精要速览系列

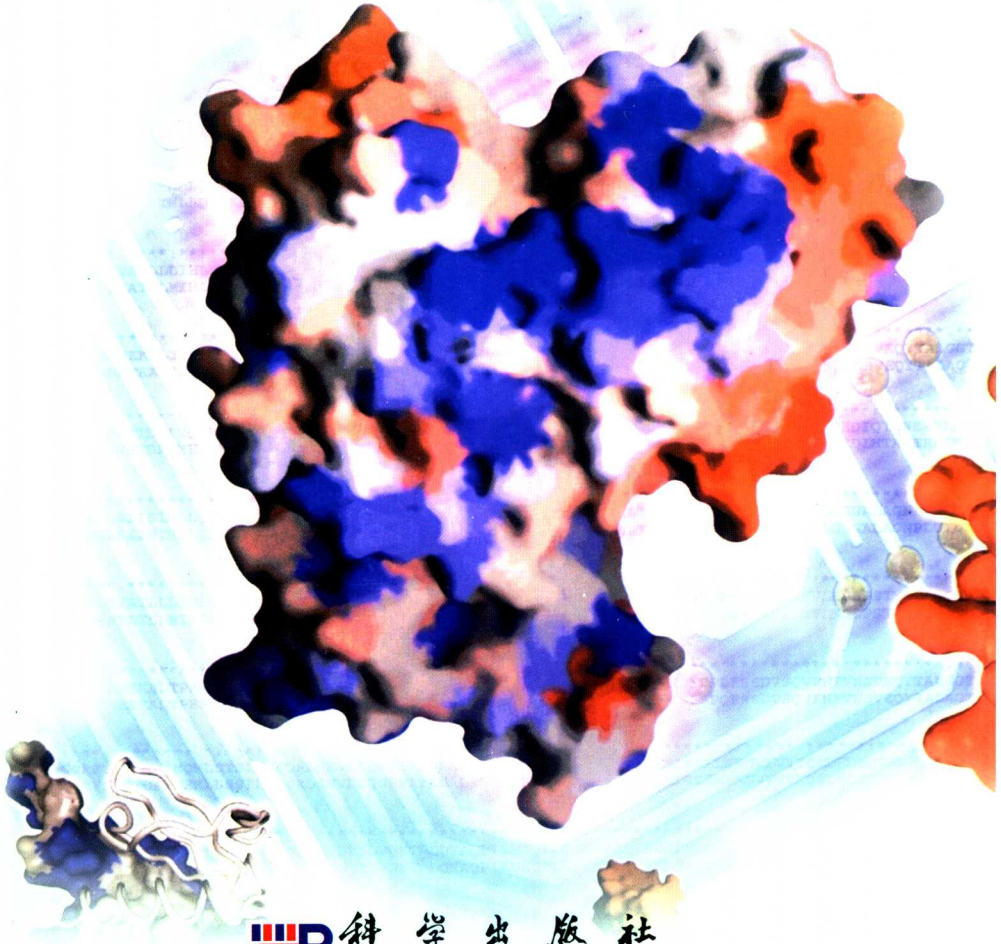
Instant Notes 先·锋·版

生物信息学（中译本）

（第二版）

Bioinformatics

D. R. 韦斯特海德
〔英〕J. H. 帕里什 著 王明怡 杨 益 吴 平 等译校
R. M. 特怀曼



科学出版社
www.sciencep.com

精要速览系列——先锋版

生物信息学

〔英〕D.R.韦斯特海德　J.H.帕里什　R.M.特怀曼　著

王明怡　杨　益　吴　平　等　译校

科学出版社

北京

图字：01-2003-6824

内 容 简 介

本书是生物学经典教材——精要速览先锋版中译本之一，包括了生物信息学的基本内容。本书内容包括：生物信息学概述、数据采集、数据库——内容、结构和注释、生物数据检索、通过序列相似性标准搜索序列数据库、多序列联配：基因和蛋白质家族、系统发育学、序列注释、结构信息学、微阵列数据分析、蛋白质组数据分析、高阶模型、生物学中的化学信息学、制药业中的生物信息学、生物信息学计算的基本原理。在书后还附有进一步阅读文献以及术语表。

本书适合大学本科生物类专业的大学生作为教材使用，也可供科研人员参考。

D. R. Westhead, J. H. Parish & R. M. Twyman

Instant Notes in Bioinformatics

Authorised translation from English language edition pubished by BIOS, a member of the Taylor & Francis Group.

图书在版编目（CIP）数据

生物信息学/(英)韦斯特海德(Westhead, D. R.)等著；王明怡等译. —北京：科学出版社，2004.9

精要速览系列：先锋版

ISBN 7-03-012894-X

Ⅰ.生… Ⅱ.①韦…②王… Ⅲ.生物信息论 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字（2004）第 008413 号

责任编辑：周 辉 彭克里 孙晓洁/责任校对：曾 茹

责任印制：安春生/封面设计：陈 敬

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2004 年 9 月第 一 版 开本：B5（720×1000）

2004 年 9 月第一次印刷 印张：18

印数：1—3 000 字数：336 000

定价：32.00 元

（如有印装质量问题，我社负责调换〈路通〉）

译者名单

（按章节顺序）

吴　平、王明怡、杨　益

张麝龙、张华荣、缪岩松

序

20世纪后期，基因组学及后基因组学的迅猛发展，无论从数量上还是从质量上都极大地丰富了生物科学的数据资源。数据资源的急剧膨胀迫使人们寻求一种强有力的工具去分析这些数据，从海量的生物学数据中寻求生物学规律，这些规律将是解释生命之谜的关键。另一方面，以数据分析、处理为本质的计算机科学技术和网络技术迅猛发展并日益渗透到生物科学的各个领域。于是，一门崭新的、拥有巨大发展潜力的新学科——生物信息学悄然兴起。

科学出版社引进的现代生物学精要速览系列书籍一直深受广大师生和科研工作者的喜爱。在生物学中的《分子生物学》等其他方向的速览书籍中文版相继出版后，《生物信息学》中文版一书便呼之欲出。《生物信息学》涉及50多个主题，基本上总括了生物信息学的基本理论、核心内容和主要技术。在每一主题中都配有“要点”栏目，有助于抓住重点，并配有简洁的插图，便于理解。

在科学出版社的主持下，浙江大学生命科学学院的几位教师和研究生翻译了该书，奉献给大家，以期帮助读者抓住重点，取得更好的学习效果。在翻译过程中，我们坚持忠实原文的原则，同时力求符合汉语的表达习惯。

参与本书翻译共有六人，其中吴平翻译了序言、目录、附录部分，王明怡翻译了A、B、C、D、H、O共六章，杨益翻译了E、F、G、I共四章，张麝龙翻译了J、K两章，缪岩松翻译了M、N两章，张华荣翻译了L章。王明怡、杨益和吴平分别对全书进行了校订，张华荣校订了M、N两章，最后由吴平、王明怡和杨益统一校订意见、规范文字和插图位置，审核并定稿。另外，也要感谢姜华武、易可可和李靖老师在本书翻译过程中给予的帮助和建议。由于时间仓促，任务繁重，翻译过程中难免会有一些错误和问题，在此敬请广大读者及时反馈指正。

吴平　教授

2003年11月3日

于浙江大学华家池

原版前言

人们从事生物序列的计算分析已经有很长时间，但是它对于生物化学家或分子生物学家的重要性仅在过去10年左右的时间内才凸显出来，因为当时的高通量测序技术带来了与许多研究计划有关的大量有价值的序列数据。于是，生物信息学作为一个非常重要的学科而问世。学科发展的核心阶段是全基因组的测序，人类基因组已在一年以前完成测序，且在不久前非生物信息学家也开始能够方便地访问这些已注释的数据。

紧随着DNA测序，迄今已有许多其他的高通量实验技术被开发出来。用微阵列可以在全基因组的规模上研究基因和蛋白质表达模式，用酵母双杂交技术也可以在同样规模上研究蛋白质的互作，甚至还有完成高通量确定结构的设想。有大规模数据产生的地方就需要数据的处理和分析，因而生物信息学的研究领域也在逐渐扩展。

目前，对从事生命科学的研究者来说，具备生物信息学知识是非常必要的，所以这一学科被设置为本科生课程。本书是为实验生物学者而准备的，可以是本科生的读物或实验生物学者在研究工作中需要该课程基本知识时的参考书。本书将对计算机工作者和其他准备转入生物学领域的人提供帮助。我们首先描述了数据的产生和数据库，然后叙述了“经典”的生物信息学问题——“对于序列我能做什么?”最后，我们速览了有关结构、表达、蛋白质组学、互作和途径等新的生物信息学问题。

本书的主旨是叙述一些分析方法的适用性和优缺点，也就是告诉读者这些方法能做什么。但是，本书不是常用软件的使用手册。那些软件的使用手册比本书提供了更详细的说明。生物信息学中的许多分析方法基于复杂的数学、统计和计算技术。因此，根本没有必要详细描述这些分析方法。但是，有时确实对它们进行了简要的描述，这时有利于对一些基本概念的理解。正如人们常说的“驾驶汽车并不需要深入理解内燃机内部的工作原理”，对生物信息学中基本方法的理解也是这个道理。所以，我们介绍了一些必要的分析方法，希望对读者有所帮助。

致谢

作者感谢Jonathan Ray、Sarah Carlson和David Hames在本书准备阶段给予的帮助、理解和支持，以及Chris Hodgson对第O章的建议和讨论。

David Westhead 谨把本书献给他的父母 Robert 和 Mavis，他的妻子 Andrea 和孩子 Elizabeth、Francis。

Howard Parish 谨把本书献给他的孙女 Scarlett。

Richard Twyman 谨把本书献给他的父母 Peter 和 Irene，他的孩子 Emily 和 Lucy，以及 Hannah、Joshua 和 Dylan。

缩略语

2D-PAGE	two-dimensional polyacrylamide gel electrophoresis 双向聚丙酰胺凝胶电泳
AC	approximate correlation 近似相关
ACeDB	A *C. elegans* Database 线虫数据库
ADIT	AutoDep Input Tool AutoDep 输入工具
AE	annotated exon 已注释的外显子
AN	actual negative 真实的阴性
AP	actual positive 真实的阳性
AQL	ACeDB query language ACeDB 查询语言
BDGP	Berkeley *Drosophila* Genome Project 伯克利果蝇基因组计划
BIND	Biomolecular Interaction Network Database 生物分子互作网络数据库
BIOS	Basic Input-Output System 基本输入-输出系统
BRET	bioluminescent resonance energy transfer 生物发光共振能量传递
CASP	Critical Assessment of Structure Prediction 结构预测的鉴定评估
CD	circular dichroism 圆二色性谱
CD	candidate drug 候选药物
CDE	common desktop environment 通用桌面环境
cDNA	complementary DNA 互补 DNA
CDS	coding sequence 编码序列
CGI	common gateway interface 通用网关接口
CIP	Cahn-Inglod-Prelog CIP 规则
CORBA	Common Object Request Brokering Architecture 公用对象请求代管体系结构
DBMS	database management system 数据库管理系统
DDBJ	DNA Databank of Japan 日本 DNA 数据库
DEC	Digital Equipment Corporation 数字设备公司
DIP	Database of Interacting Proteins 互作蛋白质数据库
DNS	Domain Name Server 域名服务器
EBI	European Bioinformatics Institute 欧洲生物信息学研究所
EMBL	European Molecular Biology Laboratory 欧洲分子生物学实验室
ENU	ethylnitrosourea 乙基亚硝基脲
EP	Expression Profiler 表达模式
ES cells	Embryonic stem cells 胚胎干细胞
ESI	electrospray ionization 电喷雾离子化

ExPASy	Expert Protein Analysis System (Switzerland) 蛋白质分析专家系统（瑞士）
FE	false exon 假外显子
FN	false negative 假阴性
FP	false positive 假阳性
FRET	fluorescent resonance energy transfer 荧光共振能量传递
FTP	file transfer protocol 文件传输协议
GASP	Gene Annotation aSsessment Project 基因注释评估计划
GEO	Gene Expression Omnibus 基因表达汇编
GFP	green fluorescent protein 绿色荧光蛋白
GGTC	German Gene Trap Consortium 德国基因捕获协会
GNOME	GNU network object model environment GNU 网络对象模型环境
GOLD	Genomes Online Database 基因组联机数据库
GOR	Garnier-Osguthorpe-Robson GOR 二级结构预测法
GRAIL	Gene Recognition and Assembly Internet Link 基因识别和汇集互联网连接
GSS	genome survey sequence 基因组调查序列
GST	glutathione S-transferase 谷胱甘肽转移酶
GUI	graphical user interface 图形用户界面
HIV	human immunodeficiency virus 人类免疫缺陷病毒
HMM	hidden Markov model 隐马尔科夫模型
HSP	high-scoring segment pair 高分值片段对
HTG	high-throughput genomic sequence 高通量基因组序列
HTML	hypertext markup language 超文本标记语言
HTS	high-throughput screening 高通量筛选
http	hypertext transfer protocol 超文本传输协议
IP	Internet Protocol 互联网协议
ISP	Internet Service Provider 互联网服务提供商
KDE	K desktop environment K 桌面环境
KEGG	Kyoto Encyclopedia of Genes and Genomes (日本) 京都基因和基因组百科全书
LCA	last common ancestor 最后共同祖先
LOG	Laplacian of Gaussian 高斯的拉普拉斯
m/e or m/z	mass/charge ratio 分子质量/电荷比率
MAD	multiwavelength anomalous diffraction 多波长不规则衍射
MAGE	microarray and gene expression 微阵列和基因表达
MAGE-ML	microarray and gene expression markup language 微阵列和基因表达标示语言
MAGE-OM	microarray gene expression object model 微阵列基因表达对象模型
MALDI	matrix-assisted laser desorption/ionization 基质辅助激光解吸附/离子化
ME	missing exon 丢失的外显子
MGED	Microarray Gene Expression Database 微阵列基因表达数据库

MIAME	minimum information about a microarray experiment 微阵列实验最低限度信息
MIME	Multiple Internet Mail Extensions 多用途互联网邮件扩充协议
MIR	multiple isomorphous replacement 多路同形置换
MMDB	Molecular Modeling Database 分子建模数据库
mRNA	messenger RNA 信使 RNA
MS	mass spectrometry 质谱
MSD	Macromolecular Structure Database 大分子结构数据库
MS-DOS	Microsoft Disk Operating System 微软磁盘操作系统
MSF	multiple sequence format 多序列格式
NBRF	National Biomedical Research Foundation 国家生物医学研究基金会
NCBI	National Center for Biotechnology Information 国家生物技术信息中心
NDB	Nucleic Acid Data Bank 核酸数据库
NJ	neighbor joining 邻近相连
NMR	nuclear magnetic resonance 核磁共振
NNSSP	Nearest Neighbour Secondary Structure Prediction 最近邻二级结构预测
NOE	nuclear Overhauser effect 原子的 Overhauser 效应
OMIM	OnLine Mendelian Inheritance in Man 孟德尔人类遗传联机数据库
ORF	open reading frame 可读框
PAGE	polyacrylamide gel electrophoresis 聚丙酰胺凝胶电泳
PAM	accepted point mutations 可接受点突变
PAUP	phylogenetic analysis using parimony 采用简约法的系统发育分析
PCNA	proliferating cell nuclear antigen 增殖细胞核抗原
PCR	polymerase chain reaction 聚合酶链反应
PDB	Protein Data Bank 蛋白质数据库
PE	predicted exon 预测的外显子
PERL	Practical Extraction and Reporting Language 实用提取和报告语言
PH	pleckstrin homology pleckstrin 同源结构域
PHYLIP	phylogenetic inference package 系统发育推断软件包
pI	isoelectric point 等电点
PIR	Protein Information Resource 蛋白质信息资源
PN	predicted negative 预测的阴性
PP	predicted positive 预测的阳性
RCSB	Research Collaboratory for Structural Bioinformatics 结构生物信息学合作研究协会
RMSD	root mean square deviation 均方差
rRNA	ribosomal RNA 核糖体 RNA
RT	reverse transcription 反转录
SAGE	serial analysis of gene expression 基因表达连续分析法
SDS	sodium dodecyl sulfate 十二烷基硫酸钠

SELDI	surface-enhance laser desorption/ionization 表面增强激光吸附/离子化
SH2, SH3	Src-homology domain Src 同源结构域
SMART	Simple Modular Architecture Research Tool 简单模块结构研究工具
SMILES	Simplified Molecular Input Line Entry Specification 简化分子输入行条目规范
SNP	single nucleotide polymorphism 单核苷酸多态性
SOM	self-organizing map 自组织映射
SPR	surface plasmon resonance 表面等离子共振技术
SQL	symbolic query language 符号查询语言
SRS	sequence retrieval system 序列检索系统
SSE	secondary structure element 二级结构元件
STS	sequence tagged site 序列标记位点
T_C	Tanimoto coefficient Tanimoto 系数
TCP	Transmission Control Protocol 传输控制协议
TE	true exon 真正的外显子
TN	true negative 真阴性
TP	true positive 真阳性
TrEMBL	Translated EMBL 翻译的 EMBL
tRNA	transfer RNA 转运 RNA
UML	Unified Modeling Language 统一建模语言
UPGMA	unweighted pair group method using arithmetic mean 不加权的算术平均组对法
UPGMC	unweighted pair group method using centroid value 采用重心值的不加权组对法
URL	uniform resource locator 统一资源定位
WE	wrong exon 错误外显子
WPGMA	weighted pair group method using arithmetic mean 加权的算术平均组对法
WPGMC	weighted pair group method using centroid value 采用重心值的加权组对法
WST	watershed transformation 分水岭转换
WWW	World Wide Web 万维网
XML	eXtensible Markup Language 扩展标记语言
Y2H	yeast two-hybrid 酵母双杂交

目　录

A 生物信息学概述

A1 生物信息学的范围

要 点

什么是生物信息学?

生物信息学是生物和信息技术的结合,是现代科学的又一个分支学科,它利用计算机对大量生物数据进行分析。生物信息学把用于存储和搜索数据的数据库开发,与用于分析和确定大分子序列、结构、表达模式和生化途径等生物数据集之间的关系的统计工具和算法的开发结合在一起。

计算机在生物信息学中的作用

计算机在生物信息学中是必需的,生物信息学需要计算机的处理速度(以及允许快速和系统地执行重复任务)和处理问题的能力。然而,生物信息学中涉及的许多问题仍需要专家的人工处理,而原始数据的完整性和质量也很关键。

本书的范围

本书的目的是给生物信息学的初学者提供足够的信息,使他们理解生物信息学应用的原理,并获得应用能力。正文部分包括介绍基本内容如互联网的作用,以及生物信息学的一些关键领域:数据库使用、序列和结构分析工具、注释工具、表达分析以及生化和分子途径分析。第O章是有关计算机操作系统和软件基本知识的附录。

生物信息学WWW网站速览

全书各处都提到存放信息资源、数据库和生物信息学工具的各种WWW网站。由于WWW的不断发展,这些网站的地址也可能会有所变动。为了方便,这些网站的链接列在由本书作者定时更新的WWW网站上(http://www.bios.co.uk/inbioinformatics)。如果书中所列的URL不能访问,通过这一网站上提供的链接寻找所需要的站点也许是最好的途径。

相关主题 生物信息学与互联网(A2)
有用的生物信息学 WWW 网站(A3)

什么是生物信息学 **生物信息学**是生物和信息技术的结合,这一学科包括了用来管理、分析和操作大量生物数据集的任何计算工具和方法。生物信息学主要由三个组成部分:建立可以存放和管理大量生物信息学数据集的**数据库**;开发确定大数据集中各成员关系的**算法**和**统计方法**;使用这些工具来分析和解释不同类型的生物数据,包括 DNA, RNA 和蛋白质序列、蛋白质结构、基因表达以及生化途径。

生物信息学这个术语从 20 世纪 90 年代开始使用,最初是 DNA、RNA 及蛋白质序列的数据管理和分析的同义词。自从 20 世纪 60 年代就有了序列分析的计算机工具,但是那时并未引起人们很大的关注,直到测序技术(参见 B1)的发展使 GenBank(参见 C2)之类的数据库中存放序列的数量快速增长。现在这一术语已扩展到其他类型生物学数据,如蛋白质结构、基因表达和蛋白质互作等,这些领域都需要有它自己的数据库、算法和统计方法,其中部分内容会在本书中讨论。

计算机在生物信息学中的作用 生物信息学尽管不是专门的计算机学科,但在很大程度上以计算机为基础。计算机在生物信息学中非常重要的原因有两个。第一,许多生物信息学问题需要重复相同的任务数百万次。例如,将一条新序列与数据库中其他每条序列作比较(参见 E3)或系统地比较一组序列来确定进化关系(参见 G2)。在这些情况下,计算机处理信息和快速测试不同解决方案的能力是必不可少的。第二,生物信息学需要计算机解决问题的能力。这类生物信息学需要解决的典型问题包括通过给定氨基酸序列得到蛋白质的折叠途径,或者通过给定搜集的 RNA 表达数据来推测生化途径。计算机可以帮助解决这些问题,但专家的输入和可靠的原始数据也是很重要的。

本书的范围 本书基于作者在本科生和研究生生物信息学教学中的经验而完成的。对于那些生物信息学的初学者来说,一个共同的起点是"对于序列我能做什么?"本书旨在提供给读者广泛的背景知识,以便理解生物信息学中所用的方法,并给出充足的事例和技术细节使读者能够解决一些真实的问题。我们描述了生物信息学中互联网的作用(参见 A)、如何产生生物信息学数据(参见 B)、数据库的重要性(参见 C)以及如何访问和搜索(参见 D)。我们讨论了序列分析(参见 E、F、

G)、序列注释(参见 H),结构分析和预测(参见 I)、基因和蛋白质表达分析(参见 J 和 K)、蛋白质互作的生物信息学(参见 L 和 M)和生物信息学在制药业中的一些应用(参见 N)。第 O 章由一系列附录组成,提供有关文件格式,计算机操作系统和软件的背景知识。

我们有意省略了有关计算生物学的论题,也省略了为精细结构而专门设计的软件、自动化的设备(包括机器人)和其他类型的数据搜集。我们说明了分子图形学的方法,但是,同样我们省略了文件显示的图形和其他帮助工具。

生物信息学 WWW 网站速览

本书引用了许多数据库和计算机软件工具,在万维网和各种信息丰富的网站中也可以找到它们。尽管书中列出了这些资源的地址,但是由于互联网的不断更新,这些网址也会定期发生变化,为了方便,正文中讨论的所有网站可以在本书的 WWW 网站(http://www.bios.co.uk/inbioinformatics)中找到。WWW 站点也包含了进一步的信息更新和本书中没有提到的链接。