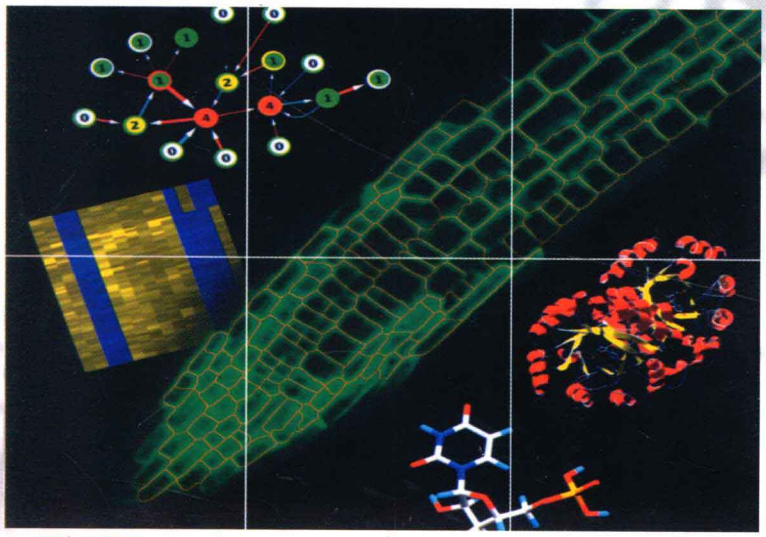
精要速览系列

Instant Notes

BIOINFORMATICS (SECOND EDITION)

生物信息学（第二版）



·中译版·

T.Charlie Hodgman,
Andrew French, David R.Westhead 编著

陈 铭 包家立 黄炳顶 译

科学出版社

(Q-3212.01)

Instant Notes

快速、准确掌握专业知识和专业外语的最佳套书！一种对教材概念的新的诠释！

- 精炼学科核心内容，以相对独立又互相关联的专题形式介绍各学科基础知识
- 版式设计独特，方便学生快速、便捷地领会学科要点，便于复习与记忆。
- 编写风格统一，提供“结构化”学习方法。
- 世界范围内的主流教材——欧洲、北美等地众多高校广泛参考和使用，国内数百家高校双语教学课程选用。

精要速览系列图书1999年面世至今受到广大读者的关注，科学出版社2009–2010年推出12个分册导读版的新版图书，2010–2011年推出10个分册的中译版。其编写风格、取材角度仍继承前版特色，在内容上根据各学科发展进行修订和扩充。

Biochemistry (3rd Edition) 生物化学 (第三版)
Molecular Biology (3rd Edition) 分子生物学 (第三版)
Genetics (3rd Edition) 遗传学 (第三版)
Immunology (2nd Edition) 免疫学 (第二版)
Microbiology (3rd Edition) 微生物学 (第三版)
Plant Biology (2nd Edition) 植物生物学 (第二版)
Animal Biology (2nd Edition) 动物生物学 (第二版)
Mathematics and Statistics for Life Scientists 生命科学中的数学与统计学
Bioinformatics (2nd Edition) 生物信息学 (第二版)
Chemistry for Biologists (2nd Edition) 生物学中的化学 (第二版)
Inorganic Chemistry (2nd Edition) 无机化学 (第二版)
Organic Chemistry (2nd Edition) 有机化学 (第二版)

高等教育出版中心
电　话：010-64030233/64019815
e-mail: bio@mail.sciencep.com

销售分类建议：生物/生物信息学/双语教学

www.sciencep.com

ISBN 978-7-03-038812-4

9 787030 388124 >

定 价：59.00 元

精要速览系列

Instant Notes in

BIOINFORMATICS

Second Edition

生物信息学

（第二版，中译版）

T. Charlie Hodgman　Andrew French

David R. Westhead

编著

陈　铭　包家立　黄炳顶　译

科学出版社

北　京

图字:01-2010-1589 号

内 容 简 介

“精要速览系列(Instant Notes Series)”丛书是国外教材“Best Seller”榜的上榜教材。该系列教材结构新颖,视角独特,重点明确,脉络分明,图表简明清晰,英文自然易懂,被许多高等院校双语教学选用。

本书在前一版基础上修订,涵盖了生物信息学的基本内容及拓展知识。全书共分三大部分,包括学科概况(A~B 章)、基础部分(C~I 章)、应用领域(J~R 章),合计 18 章:A 生物学研究新面貌、B 生物信息学的定义、C 物理学要素、D 数据与数据库、E 数据类别、F 计算、G 概率与统计、H 模型与数学技术、I 人工智能与机器学习、J 基因组与其他序列、K 转录组学、L 蛋白质与蛋白质组学、M 代谢物组学、N 超分子结构、O 生化动力学、P 生理学、Q 图像分析、R 文本分析。书前附有缩略词表,书后附有进一步阅读的文献。

本书适合普通高校生命科学、医药科技,以及生物信息学相关专业教学使用,也可供科研人员参考阅读。

T. Charlie Hodgman, Andrew French, David R. Westhead
Instant Notes in Bioinformatics, 2nd edition

ISBN 978-0-415-39494-9

图书在版编目(CIP)数据

生物信息学:第 2 版/(英)霍奇曼(Hodgman, C.)等编著;陈铭,包家立,黄炳顶译. —北京:科学出版社,2013. 10
(精要速览系列)
ISBN 978-7-03-038812-4

Ⅰ. ①生… Ⅱ. ①霍… ②陈… ③包… ④黄… Ⅲ. ①生物信息学-教材 Ⅳ. ①Q811. 4

中国版本图书馆 CIP 数据核字(2013)第 238354 号

责任编辑:刘 畅 / 责任校对:郭瑞芝
责任印制:阎 磊 / 封面设计:迷底书装

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
文林印务有限公司 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:787×1092 1/16
2013 年 10 月第 二 版 印张:15 3/4
2013 年 10 月第一次印刷 字数:414 000
定价:59.00 元
(如有印装质量问题,我社负责调换)

译者名单

陈　铭(浙江大学)

包家立(浙江大学)

黄炳顶(浙江大学)

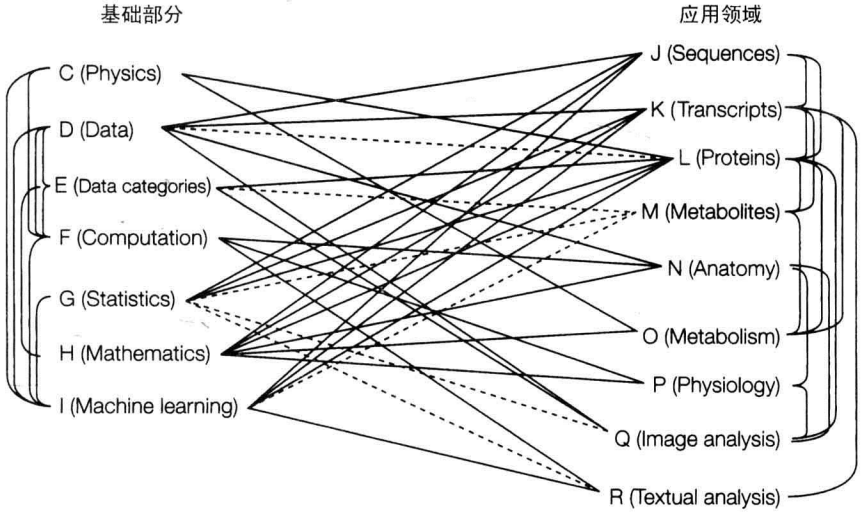
前　言

自从“精要速览系列”的生物信息学第一版出版以后，生物信息学领域已经有了实质性的发展，而且正在变成一门具有自身特点的学科。我们非常感谢出版商给我们机会出版生物信息学第二版，这使我们能够根据两个目标来重新构思这本书。首先，为化学、生物学、医学和神经学等研究领域的信息学研究者提供资料；其次，展示这些通用的信息学技术如何应用在生命科学的大多数领域，而不仅仅是在生物信息学最初活跃的分子生物学领域。

本书章节主要分成3部分，第一部分（A章和B章）主要对这个学科进行介绍。A章概述了使生物信息学成为一个必需领域的因素。B章主要介绍该学科从20世纪60年代兴起，经过令人振奋（或是令人兴奋）的20世纪90年代，直到生物信息学正应用于所有类型的生物学信息处理的21世纪的简要历史（通过一系列对生物信息学这一术语定义的演变过程）。

第二部分是信息学的基础部分（C～I章）：物理学、数学和计算机科学。但缺少一项重要内容——计算机编程，其是生物信息学的基本技能，受图书篇幅限制，无法详细阐述。由于这是一个特别实用的领域，最好是将这个问题留给大量的其他可以利用的书籍。不过，我们尽力概述有效的数据管理和程序设计习惯的基础知识。

第三部分是生物学的应用领域（J～R章）。它包括3个部分：分子生物学，新陈代谢、解剖学、生理学，复杂的信息来源（特别是图像的数据集和自然语言文本）。后者仍然是提取准确的量化数据最困难的地方。第二部分和第三部分的关联如下图所示，它强调基础部分的基础性、重要性。从二者紧密联系的网络来看，它们二者的应用领域都存在明显的相互依赖关系。



图例说明：章节间的联系。这本书的基础部分章节和应用领域章节分两列显示。如果它们之间的关系是实线，那么它们是借助相关主题相关联的，而虚线是指它们的关系隐含在正文中。

现代生物信息学涵盖的内容相当广泛，因此本书的3位主要作者一致认为某些特定章节由他人编写更为合适。在此对作出贡献的相关人员致以谢意：J章的Nicola Gold，K章的Alex Marshall，L章的Nicola Gold和Tom Gallagher，M章的Rob Linforth。许多人还检查各章节的准确度和清晰度，他们是：Alastair Middleton，Leah Band，Tom Gallagher，Kim Kenobi，特别是Jane Hodgman（他校对了许多章节），在此深表感谢。我们感谢英国植物综合生物学中心的成员在出版前提供的显微图像。读者会很容易发现一些重复，但是这是为清晰起见特意保留。最后，我们希望学生和教师都能体会到这门学科的广度，并享受阅读的快乐。

译者前言

生物信息学是生物学、计算机科学与技术及应用数学等多学科相互交叉而形成的新学科。它通过对生物学实验数据的获取、存储、处理与分析，进而达到揭示这些数据所蕴含的生物学意义的目的。随着越来越先进的生物技术的开发与运用，越来越多的组学数据的高通量测定并保存，已将生命科学推进到大数据时代。生物信息学的作用也越发显得重要。

科学出版社引进的 Charlie Hodgman 教授等编著的 *Instant Notes in Bioinformatics* 内容广泛，涉及生物信息学的诸多领域和研究基础及前沿内容，每个主题都进行了深入浅出的讲解，是一本不可多得的可供教学科研使用的书籍。为推广生物信息学在国内的普及，在科学出版社的主持下，由浙江大学几位教师和研究生对该书进行了翻译。

陈铭研究小组翻译了 A 至 F、J、K、O 至 R 共 12 章；包家立、王顺翻译了 G、H、I 共 3 章；黄炳顶、孙泽宇、马莹莹、刘贯峰翻译了 L、M、N 共 3 章。陈铭研究小组对全书进行了审核、校订并定稿。感谢以下人员参与翻译、校对并给予的帮助：金丹凤、李麒麟、原春晖、刘丽丽、王晶晶、代晓转、姜丽、王月。

翻译过程中对原文中出现的错误进行了修正，限于时间和译者水平，难免会有一些疏漏和不当之处，在此敬请广大读者反馈指正。

陈　铭

2013 年 8 月

缩 略 词

AC	approximate correlation	近似相关
ADME	absorption, disposition metabolism and excretion	吸收、分布、代谢和排泄
ADR	adverse drug reaction	药物不良反应
AE	annotated exon	已注释的外显子
AI	artificial intelligence	人工智能
ALU	arithmetic logic unit	运算器算术逻辑单元
AN	actual negative	真实的阴性
ANOVA	analysis of variance	方差分析
ANSI	American National Standards Institute	美国国家标准学会
AP	actual positive	真实的阳性
APM	accepted point mutations	可接受点突变
ArMeT	architecture for metabolomics	代谢体系结构
ASCII	American Standard Code for Information Interchange	美国信息交换标准代码
ATP	adenosine triphosphate	三磷酸腺苷
BE	boundary-element	边界元(素)
BioPAX	Biological Pathways Exchange	生物学途径交换
BLAST	Basic Local Alignment Search Tool	基本局部联配搜索工具
BLOB	binary large object	二进制大型对象
CASP	Critical Assessment of Structure Prediction	结构预测评估
CCD	charge-coupled device	电荷耦合器件
cDNA	complementary DNA	互补脱氧核糖核酸
ChEBI	chemical entities of biological interest	生物学相关的化学条目
COSY	correlated spectroscopy	关联光谱学
CPU	central processing unit	中央处理器
CT	computed tomography	计算断层照相法
CVS	concurrent versions system	并发版本系统
DARPA	Defense Advanced Research Projects Agency	美国国防部高级研究规划局
DAS	distributed annotation system	分布注释系统
DDBJ	DNA Databank of Japan	日本 DNA 数据库
DIGE	differential gel electrophoresis	差异凝胶电泳
DIKW	data, information, knowledge and wisdom	数据、信息、知识和智慧
DNA	deoxyribonucleic acid	脱氧核糖核酸
EBI	European Bioinformatics Institute	欧洲生物信息学研究所
EC	Enzyme Commission	酶学委员会
ECG	electrocardiogram	心电图
EM	expectation-maximization	期望-最大化
EMBL	European Molecular Biology Laboratory	欧洲分子生物学实验室
ESI	electrospray ionization	电喷雾离子化

EST	expressed sequence tag	表达序列标签
FAD	flavin adenine dinucleotide	黄素腺嘌呤二核苷酸
FE	false exon or finite-element	假外显子或有限元
FN	false negative	假阴性
FP	false positive	假阳性
GA	genetic algorithm	遗传学算法
GASP	Gene Annotation Assessment Project	基因注释评估计划
GATE	General Architecture for Text Engineering	文本工程通用框架
GIGO	garbage in, garbage out	无用输入,无用输出
GIS	geographic information system	地理信息系统
GUI	graphical user-interface	图形用户界面
GOLD	Genomes Online Database	基因组在线数据库
GRAIL	Gene Recognition and Assembly Internet Link	基因识别和汇集互联网链接
GSS	genome survey sequence	基因组查询序列
GXD	gene-expression database	基因表达数据库
HCA	hierarchical clustering analysis	系统聚类分析
HMM	hidden Markov model	隐马尔可夫模型
HSP	high-scoring segment pair	高分值片段对
HT	high-throughput	高通量
HTG	high-throughput genomic	高通量基因组
HTML	hypertext mark-up language	超文本标记语言
HTTP	hypertext transfer protocol	超文本传输协议
IDE	Integrated Development Environment	集成开发环境
IE	information extraction	信息提取
IEEE	Institute of Electrical and Electronic Engineers	电子和电气工程师协会
ILP	inductive logic programming	归纳逻辑编程
IP	Internet Protocol	互联网协议
ISO	International Organization for Standardization	国际标准化组织
IT	information technology	信息技术
JRE	Java runtime environment	Java 运行环境
KEGG	Kyoto Encyclopedia of Genes and Genomes	京都基因和基因组百科全书
KGML	KEGG Mark-up Language	KEGG 标记语言
LOG	Laplacian of Gaussian spot Detection	高斯光点检测的拉普拉斯算子
LOPIT	localization of organelle proteins by isotope tagging	同位素标签定位细胞器蛋白质技术
MAGE	microarray and gene expression	微阵列和基因表达
MALDI	matrix assisted laser desorption/ionization	基质辅助激光解吸附/离子化
MC	Monte Carlo	蒙特卡洛
MCMC	Markov chain Monte Carlo	马尔可夫链蒙特卡洛理论
MeSH	medical subject headings	医学主题词表
MIAME	minimum information about a microarray experiment	微阵列实验最低限度信息
MIAMET	minimal information on a metabolomics experiment	代谢实验最低限度信息
MIAPE	minimal information about a proteomics experiment	蛋白质组学实验最低限度信息
mmCIF	macromolecular crystallographic information file	大分子结晶学信息文件
MMDB	Molecular Modeling Database	分子模型数据库

MRC	Medical Research Council	医学研究委员会
MRI	magnetic resonance imaging	磁共振成像
mRNA	messenger RNA	信使核糖核酸
MS	mass spectrometry	质谱分析
NAD	nicotinamide adenine dinucleotide	烟酰胺腺嘌呤二核苷酸
NASA	National Aeronautics and Space Administration	国家航空航天局
NCBI	National Center for Biotechnology Information	国家生物技术信息中心
NJ	neighbor-joining	邻近连接
NLP	natural language processing	自然语言处理技术
NMR	nuclear magnetic resonance	核磁共振
NNSSP	Nearest Neighbor Secondary Structure Prediction	最近邻二级结构预测
NOESY	nuclear overhauser effect spectroscopy	核极化效应光谱学
ODE	ordinary differential equation	常微分方程
OMIM	Online Mendelian Inheritance in Man	人类孟德尔遗传在线数据库
OODB	object-orientated database	面向对象型数据库
OOP	object-oriented programming	面向对象型程序设计
ORF	open reading frame	可读框
PAGE	polyacrylamide gel electrophoresis	聚丙烯酰胺凝胶电泳
PAUP	phylogenetic analysis using parsimony	采用简约法的系统发育分析
PCs	personal computers	个人计算机
PCA	principal components analysis	主成分分析
PDB	protein data bank	蛋白质数据库
PDE	partial differential equation	偏微分方程
PE	predicted exon	预测的外显子
PES	potential energy surface	势能面
PHP	personal home page	个人主页
PHYLIP	Phylogenetic Inference Package	系统发育推理软件
PN	predicted negative	预测阴性
PO	plant ontology	植物本体
PP	predicted positive	预测阳性
PRPS	phosphoribosyl pyrophosphate synthetase	磷酸核糖焦磷酸合成酶
PSSM	position specific scoring matrix	位置特异打分矩阵
QSAR	quantitative structure-activity relationship	定量构效关系
RMSD	root mean square deviation	均方差
RNA	ribonucleic acid	核糖核酸
RT-PCR	reverse transcriptase polymerase chain reaction	逆转录聚合酶链反应
SAGE	serial analysis of gene expression	基因表达连续分析
SBML	systems biology mark-up language	系统生物学标记语言
SDEs	stochastic differential equations	随机微分方程
SMART	simple modular architecture research tool	简单模块结构研究工具
SMARTS	an extension of SMILES	简化分子输入条目系统的扩展
SMILES	Simplified Molecular Input Line Entry System	简化分子线性输入系统
SMRS	standard metabolic reporting structure	标准代谢报告结构
SNOMED	Systematized Nomenclature of Medicine	医学系统术语

SNP	single nucleotide polymorphism	单核苷酸多态性
SOM	self-organizing map	自组织映射
SQL	structured query language	结构化查询语言
SRS	sequence retrieval system	序列检索系统
TAP	tandem affinity purification	串联亲和纯化
TCA	tricarboxylic acid	三羧酸
TCP	transmission control protocol	传输控制协议
TE	true exon	真正的外显子
TIC	total ion chromatogram	总离子色谱图
TIFF	tagged image file format	标签图像文件格式
TN	true negative	真阴性
TP	true positive	真阳性
tRNA	transfer RNA	转移核糖核酸
UDDI	universal description, discovery and integration	统一描述、发现和集成
UML	Unified Modeling Language	统一建模语言
UMLS	Unified Medical Language System	统一医学语言系统
UniProt	Universal Protein Resource	通用蛋白质资源
USB	universal serial bus	通用串行总线
UTF	unicode transformation format	统一码变换格式
UV	ultraviolet	紫外线
WE	wrong exon	错误外显子
WSDL	Web Services Description Language	网络服务描述语言
WST	watershed transformation	分水岭变换
WWW	worldwide web	万维网
XML	extensible mark-up language	可扩展标记语言

目　录

A　生物学研究新面貌

要　点

引　言

随着各种专业与日俱增的分化，生物学的研究也逐渐多样化，从20世纪70年代中期开始，以4种新的方法与技术为核心的驱动力改变了生物学的研究方式。本节将依次介绍。

万物皆分子

从最初的基础生物化学到最终的几乎所有生物学门类，分子生物学与遗传学的结合已成为鉴定生物过程中组成成分的一大利器。

小型化与自动化

生物技术专家已经发展并将继续寻求从更小的生物样本中挖掘更多信息的技术。自动化机器的出现使得高效并可重复性地处理大量样本成为可能。这些技术被称为高通量技术。

图像分析

为了使高通量技术产生的大容量数据（每个样本MB级别）易于处理，原始输出通常由计算机程序编译过的图像呈现。

计算与统计建模

高通量技术产生的大量数据通过一系列的统计分析，确定生物对象（基因、蛋白质等）个体与群体的性质。生物过程源于这些生物对象之间的相互作用，已经有大量的计算方法用来模拟这些过程。这些生物过程最简单的形式包含生化反应或基因调控的相互作用网络。然而，生物现象（或生物系统）的动力学和量化行为也可以通过更细致的数学模型来表现。这些模型可以通过计算机模拟的方式进行“what-if”实验，在这种情形中，模型的质量关系到它的模拟能力，尤其是在预测系统的行为时。这些数学模型只有配备高性能计算机时才能得以实现，意味着生物学变得越来越像物理学，因为理论生物学家正赶上并很有可能会超越纯粹的实验生物学家。

研究方式转变的结果

这些变化的结果是生物学家将花费越来越多的时间在数据分析上，而花费在实验本身的时间较少。这些改变也导致了对能够及时以有生物学意义的方式管理和分析海量数据的人员的需求迅速增长。这些人就是生物信息学家。章节B将通过一些定义描述生物信息学的简短历史。

相关章节

生物信息学的定义（B）
概率与统计（G）
模型与数学技术（H）
转录组学（K）
蛋白质与蛋白质组学（L）
图像分析（Q）

引　言

直到20世纪的最后25年，生物学的研究仍是观察和描述“生物体”的形态，以及它们在不同环境和遗传背景下的行为和功能。最初，这些“生物体”是可见的单个植物或动物（生物学），或者是生物群体及它们相互之间和与周围环境的作用关系（种群生物学，生态学）。外科医生和植物学家开始研究动物和植物体内的器官；它们的物理分布/连通性及它们在生物体中的作用（解剖学，生理学）。显微技术开拓了研究细小生物的征途（细胞生物学，微生物学）。从19世纪末开始，化学家逐渐对生物体内的分子产生了兴趣（生物化学），他们的工作因物理学家的加入得以推进，尤其是第二次世界大战之后，物理学家们开发出了纯化和研究大分子结构与功能，并通过放射性标记研究其代谢途径的技术。然而，以4种新的方法与技术为核心的驱动力使得生物医学研究从描述型转变为假设型。这些驱动力将在下文逐一阐述。

万物皆分子

基因克隆和测序技术在20世纪70年代得以发展，导致了在80年代人们研究生物学的方法发生了改变。人们不再试图对某些生物机制中的蛋白质进行鉴定和描述，而是通过分子遗传技术描绘决定蛋白质序列的基因或转录本序列。这些生物机制从微生物学延伸到高等动植物的生理学和病理学。在果蝇中，对其胚胎发育的研究取得了重

大突破。另外,有很多人转向研究同类物种不同个体的相同基因,80 年代后期聚合酶链反应(PCR)技术的发展使其变得更加容易。因此,种群遗传学和种群动态学开始研究特定序列在种群内部和种群之间的变化。对于人类,这产生了"线粒体夏娃"的说法,指的是早期迁出非洲的少数女性的线粒体在人群中的分布。更多类似的迁移研究已经完成,从而将分子生物学带入了人类学的领域,且有些研究已经试图将生物数据和语言的发展联系到一起。某些基因组序列演化的相对较快,使得一些序列只局限于关系紧密的家庭成员。除了应用于亲子鉴定外,它们已经很大程度地改变了重大刑事案件的调查方式。如今在法医学和案发现场中收集 DNA 样本已经司空见惯。类似的 DNA 检验也已经在考古学中被用来判断出土骨骼的来源。

小型化与自动化

大多数人都已经习惯使用越来越小的容器,试管被 1.5ml 离心管代替,进而又被多孔板所取代。为了能够操作更小容量的液体,玻璃移液管被带有一次性吸头的微型移液器(分配大约 0.5ml 的液体)、微量注射器(纳米级)和玻璃毛细血管(皮米级)取代。这些小规模器械使得昂贵的试剂(如纯化的酶)能使用更长的时间;一瓶试剂可以做更多的实验,从而使得服务型企业和机构(如医院实验室)可以同时执行多次实验。然而,大量重复的执行同样的实验操作对实验者来说是非常枯燥的。开发能够执行重复试验的仪器,一方面解放了许多的实验者,另一方面,使得更多数量的实验可以完成。因此,我们有了一批被称为高通量技术的实验技术(通常缩写为"HT")。其他的实验技术也在向高通量转变。

聚丙烯酰胺凝胶电泳自 20 世纪 60 年代开始就被用于分离生物大分子。这种技术的处理量已经从几个方面得以增加。使用更薄的凝胶和更窄的泳道可以处理更少量的样本。从 80 年代后期开始,一些 DNA 测序仪通过使用"鲨鱼齿梳子"减少了泳道之间的间隙。一些凝胶电泳设备可以允许多块凝胶同时运行。

之后的一个重要技术是将大分子固定到滤膜上,然后利用已标记的试剂进行杂交检测。Ed Southern 将这一方法用于鉴定同源 DNA 片段,即发明了"Southern 杂交"(Southern blots)。几个月之后,能固定和检测 RNA(核糖核酸)和蛋白质的方法也被发明(分别被称为 Northern blots 和 Western blots)。然而 Eastern blots 还没出现。滤膜最初是由硝化纤维制成,其质地坚硬,与多核苷酸发生化学上不可逆结合时易于操作。然而,这种滤膜在干燥时易碎易燃,因此塑料材质的滤膜毫无疑问地得到了发展。这些滤膜的小型化使得照相平板印刷能够在玻璃板上将印迹聚集得非常紧密。因此我们创造了"芯片",在 $1cm^2$ 上有大约 50 000 个印迹(详细内容请见章节 K)。

光谱技术在过去 60 年内一直被改进,尤其是其敏感性。质谱技术最初只能用于分离代谢产物。然而,程序上和技术上的发展使其可以检测精确度小于 1Da 的寡肽的质量。由连续的氨基酸组成的寡肽混合物质量的差异可以被检测到,并且从大量的差异点可推断出一些容易混淆的多肽序列(如亮氨酸和异亮氨酸的质量相同)。从光谱中得到的值也可以反映转录后修饰的数据,以及实验步骤中所使用的同位素的分布情况。关于这方面的详细阐述可以见章节 L。组成光谱分析原材料的蛋白质通常来自单向或双向的聚丙烯酰胺凝胶电泳。然而最近的技术发展偏向于离子交换和反相色谱相结合,可以避免在抽提凝胶时丢失低水平蛋白质。这种方法的缺点是一个单一的实验也需要分析大量的标准样数据。

磁铁磁性的不断增强使得用核磁共振(NMR)方法来进行生物医学研究成为可能。一方面,它可以梳理并定量分析出代谢产物和蛋白质间的反应。虽然混合物复杂,代谢混合物有时还是可以分离出来,我们仍然也可以得到关于化合物种类的信息。另一方面,在另一种完全不同的领域,核磁共振现已普遍用于医院和医学研究单位中,借以无创伤获得组织内部情况的图像(http://www.medicinenet.com/mri_scan/article.htm)。在这个规模上还有正电子发射断层扫描技术和计算机断层扫描技术。

最后,随着以上实验技术的自动化和不断发展,大规模生产和结晶蛋白所需要的

时间已大大缩短。对于可溶性蛋白,之前需要数年制得的现在只需要几个月。如今通过 X 射线晶体和 NMR 可以更快地得到新蛋白质的结构。当一个蛋白质序列与另一个结构已知的蛋白质相似时,相比于用 X 射线照射含重金属衍生物的蛋白质晶体得到电子密度图谱的方法来确定新蛋白质的结构,用同晶置换的方法可以节约大量的时间。

图像分析

上面所说的一切都导致了一个结果,就是某些生物过程产生的数据量呈爆炸式增加。通过图像形式处理数据来体现计算过程的方法日趋增长,有人认为这种现象会得以缓解,当然也有人认为这种情况会恶化。第一张图像是手绘的,但是摄影技术已成为几十年来发表研究结果的支柱。计算机捕获和图像分析的出现使得自动化技术产生、解释和输出生物数据的方式有了大幅度改变。有关图像分析的技术将在章节 Q 中具体介绍,但在此阶段读者应该知道图像的产生和分析有多普遍,包括凝胶电泳(含 DNA 测序)和杂交膜探测、阵列分析(多孔板实验、免疫分析和芯片)、显微技术、影像和在前一节提及的不同的扫描方式。一个单独的图像文件的大小可能是几 kB,也可以超过一个 GB。如果一个数据集包括一系列的图像,如时间序列数据集或三维(3D)成像图,那么这个数据集可以很容易地扩展到几千 GB。那需要大量的数据储存、检索和分析。

计算与统计建模

模型是事物的简化表示法。这里的"事物"可以是生物过程或行为,也可以是和生物体有关的数据。对一个或一群物体的实验性观察可能是不准确的。当不精确度很大时,得到的实验数据就被认为有噪声。因此,需要统计学(见章节 G)从这些数据中抽取有意义的结果,将群体样本作为一个整体阐述其性质,并定量说明其中的任意个体与样本中其他部分的相似(或不同)程度(如两个蛋白质序列或叶子形状的相似性)。

一个生物系统是指在分子、细胞、生物体或不同的物种群落中发生的过程的集合。系统行为是指系统随着时间发生的改变或对某些外界变化的反应,也能由统计学技术来建模。然而,这些模型倾向用于给出一个系统行为的整体观,而对于系统内部的具体反应则提及甚少。这种方法增强了预测外界因子对系统的影响的可靠性。相应的例子包括化合物的致癌性,以及配体与蛋白质之间的亲和力(通常被称为定量构效关系,QSAR)。

那些描述系统内部机制的模型则需要其他类型的数学技术。这些数学技术总结在章节 H 中。这种模型具有巨大的潜力,并且随着计算机能够以实际时间尺度来模拟生物系统行为,它们在近几十年已经表现出自身特有的价值。一个模型仅仅当它能够模拟系统行为的时候才算好,甚至还要求能够预测系统对外界扰乱的反应。然而,机械模型也能配合系统行为的实际数据来确定不能直接测量的系统参数,这就是参数拟合。典型例子就是用 Michaelis-Menten 方程描述酶活性。这个模型长期以来被许多生化学家认为是真实存在的,而不仅仅是酶实际作用的一种表述。大多数生化专业本科生都能通过测量绘出不同的底物浓度中酶的初始反应速率,从而决定该酶固有的动力学参数:V_{max}(最大速率)和 K_m(米氏常数)。

研究方式转变的结果

这些自动化或高通量技术,结合通过高性能计算机系统处理的大容量数据,已经改变了生物学家的研究方式。如今越来越多的情况是,实验或许可以在实验室里几天内完成,但是却需要几个月的时间在办公室里处理数据。如果能够找到软件高效率地处理数据,那么整个研究过程将会大大加快,且生物学家将能够自由找出处理已有数据的新方法,或者开发出新技术来通过新方式调查他们的研究领域。反过来,IT 专家和计算机专家将走出他们的办公室,到实验室来接触可以用来计算的新实验仪器。结果是,他们将发现有关数据质量的一些技术细节,如何才能最好地处理和分析这些数据,以及怎样才能和其余的生物信息学领域最好地整合起来。

那些介于实验室和办公室之间的人的工作将会起到一个至关重要的作用。他们就是生物信息学家。他们的工作将会在章节 B 中阐述。

B　生物信息学的定义

要　点

引　言　生物信息学是近年来形成的一门学科，其涵盖的范围和重要性都在不断扩大和提高。对生物信息学的一系列定义也反映出这些变化和人们不断地关注。

起　源　生物信息学一词最早在1968年提出，而正式的生物信息学课程则是在1978年才完全形成。其后由于技术的进步而形成了今天我们看到的大量生物学数据。

王室定义　20世纪90年代中期，人们对此学科产生了浓厚的兴趣，并产生了支持和反对两种激烈的反响。但在此领域投入的资金比预想的要更多。

标准定义　生物信息学是一门融合生物学、信息科学和数学的交叉学科。这一定义拓展了学科的范围，从分析生物分子序列和结构到其他类型生物学数据（生物分子、相互作用、种群和细胞生物学）。生物信息学的多学科性质和生物信息学家作为翻译员的作用已得到认可。

功能定义　这个定义明确说明了人们如何尝试使用生物信息学，并展示了怎样的学科组成适合这门学科的体系。

公共定义　这个冗长的定义试图阐明这门学科的各个方面，包括与其他信息学科之间的关系。

相关章节

物理学要素（C）
数据与数据库（D）
数据类别（E）
计算（F）
概率与统计（G）
模型与数学技术（H）
人工智能与机器学习（I）
基因组与其他序列（J）
转录组学（K）
蛋白质与蛋白质组学（L）
代谢物组学（M）
超分子结构（N）
生化动力学（O）
生理学（P）
图像分析（Q）
文本分析（R）

引　言　如同所有新兴学科一样，生物信息学有的是更多的定义和相对较少的实践者，且这些定义本身已随着人们认识的增长而不断演变。与其规范的讲解，我们不如选择性地给出一些可以突出其演变过程的定义。

起　源　大部分生物信息学教科书都是从生物学数据（通常是大分子序列）、计算机、生物信息学的诞生，以及如何管理和分析爆炸式的生物数据的角度来定义此学科的。但几十年来这种几乎不提及学科起源的讲解却造成了一些误区。生物信息学一词（法语为bio-informatique）在1968年首次出现在Rybak的教科书中。当时最长的核苷酸序列是76bp的tRNA，遗传密码也才破译不久。当时阿波罗宇宙飞船上的计算机只有相当缓慢的CPU和只是现代手机百万倍分之一的存储空间，整个NASA的计算机运算能力都比现在的一台家用电脑弱。

书中有3个章节，Rybak概括了生物分子、细胞、组织和生物体编码信息的方式，及这些信息是如何根据热力学定律进行传递的。这种编码包括以下6种。

1. 手性氨基酸、糖类和其他代谢产物。
2. 分别存在于核酸、蛋白质和多糖中的碱基序列、氨基酸序列和糖类。
3. 不同复合物、细胞成分、细胞类型和生物体中小分子与大分子的光谱，不同组织与器官的解剖学。
4. 生物种群的密集度与分布，大分子转录、翻译和复制过程中的信息传递。
5. 生物体内通过循环系统及生物体间通过扩散代谢物的激素信号（即现在所指的信