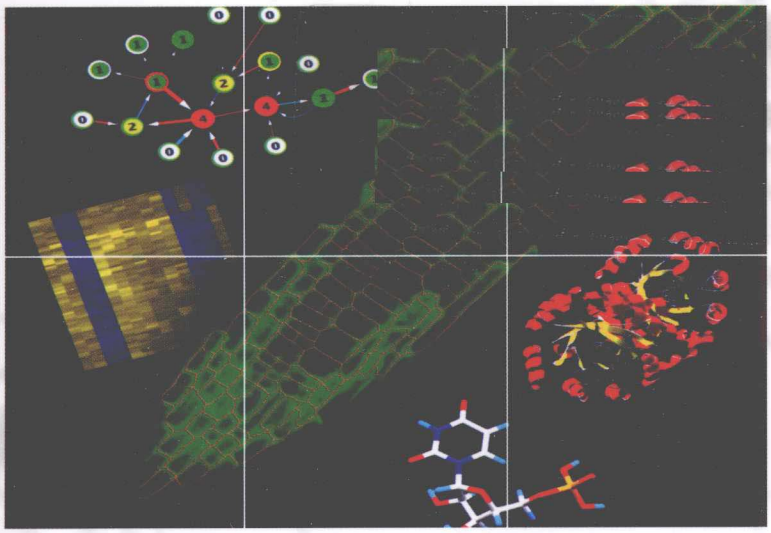
精要速览系列

Instant Notes

BIOINFOMATICS (SECOND EDITION)

生物信息学（第二版）



·导读版·

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

科学出版社
www.sciencep.com

精要速览系列

Instant Notes in

Bioinformatics

Second Edition

生物信息学

（第二版，导读版）

科学出版社

北京

图字:01-2010-1587 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:学科概况(A-B)、基础部分(C-I)、应用领域(J-R),合计18章:A 生物学研究方式的转变、B 生物信息学的定义、C 物理学要素、D 数据及数据库、E 数据类型、F 计算、G 概率与统计、H 模拟与数学技术、I 人工智能和机器学习、J 基因组及其他序列、K 转录物组学、L 蛋白质与蛋白质组学技术、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以及索引。

本书适合普通高校生命科学、医药科技相关专业,以及数学、物理、化学、计算机等理工科专业教学使用,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

ISBN978-0-415-39494-9

图书在版编目(CIP)数据

生物信息学=Bioinformatics:导读版:英文/(英)霍奇曼(Hodgman, T. C.)等编著. —2 版. —北京:科学出版社,2010
(精要速览系列)
ISBN 978-7-03-028873-8

Ⅰ.①生… Ⅱ.①霍… Ⅲ.①生物信息学-双语教学-高等学校-教材-英文 Ⅳ.①Q811.4

中国版本图书馆 CIP 数据核字(2010)第 173903 号

责任编辑:单冉东 / 责任校对:张凤琴
责任印制:张克忠 / 封面设计:耕者设计工作室

科学出版社 出版
北京东黄城根北街 16 号
邮政编码:100717
http://www.sciencep.com
双青印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2004 年 9 月第 一 版 开本:787×1092 1/16
2010 年 9 月第 二 版 印张:24 3/4
2010 年 9 月第一次印刷 字数:600 000
印数:1—3 000
定价:52.00 元
(如有印装质量问题,我社负责调换)

导读编译者

高雪峰（吉林大学生命科学学院）

赵　熹（吉林大学理论化学计算国家重点实验室）

胡建平（乐山师范学院化学与生命科学学院）

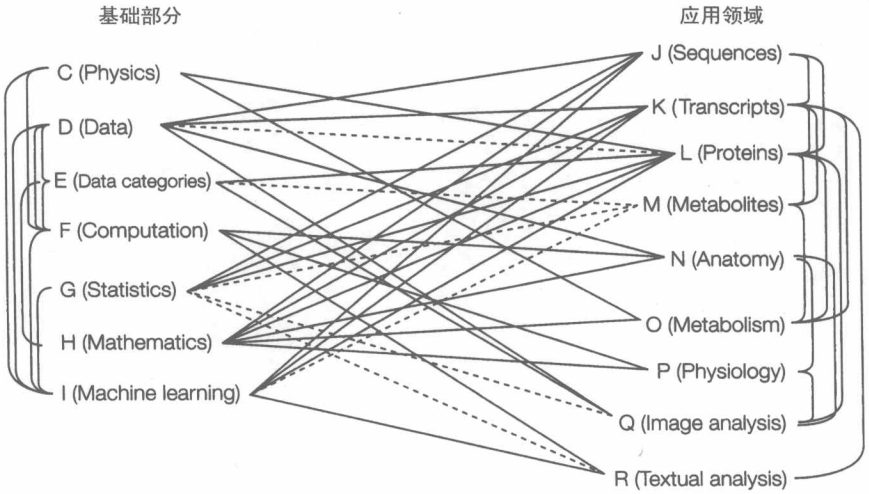
前　言

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。

本书章节主要分成三部分，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，经过的令人振奋（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的 21 世纪的简要历史（通过对一系列的生物信息学这一术语的定义）。

第二部分是信息学的基础部分（从 C 章到 I 章）：物理学，数学和计算机科学。但缺少一项重要内容——计算机编程是生物信息学的基本技能，受图书篇幅限制，无法进行有用语言的充足训练。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。

第三部分包含生物学的应用领域（从 J 章到 R 章）。它包括三个部分：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。第二和第三部分的关联如下图所示，这强调基础部分的基础重要性。从两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。



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，因此本书的三位主要作者一致认为某些特定章节由他人编写更为合适。在此对做出贡献的相关人员致以谢意：J 章的尼古拉·金，K 章的亚历克斯·马歇尔，L 章的尼古拉·戈尔德和汤姆·加拉格尔和 L 章的罗布·林福思。许多人还检查各章节的准确度和清晰度，故此非常感谢阿拉斯泰尔·米德尔顿，利亚·邦德，汤姆·加拉赫，金肯·诺比，特别是简·霍奇曼（他校对了许多章节）。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。

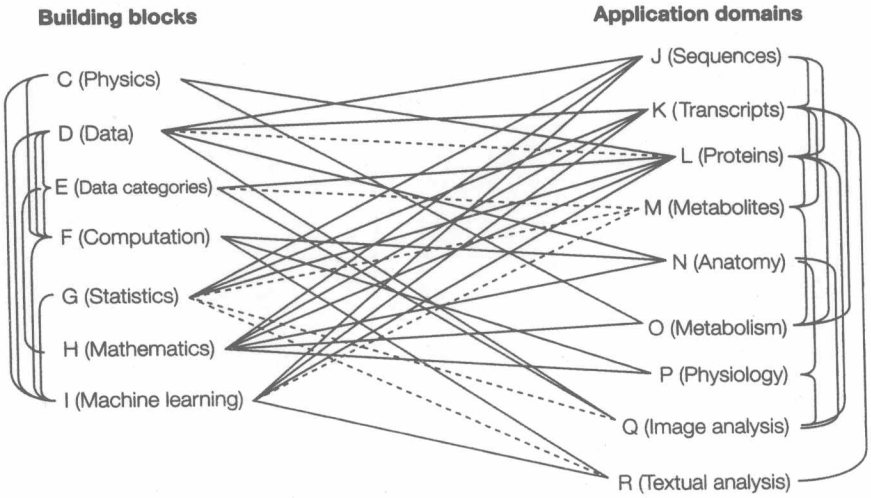
PREFACE

Since the first edition of *Instant Notes – Bioinformatics*, the field has progressed substantially and is well on the way to becoming an established discipline in its own right. We are grateful, therefore, to the publishers for giving us the opportunity to produce a second edition. This has enabled us to restructure the book into a form that has two aims. The first is to provide material for informatics students irrespective of their chosen field – biology, chemistry, medicine, neuroscience, etc. The second is to show how these generic informatics skills are being applied to most aspects of the life sciences, and not simply molecular biology where bioinformatics first flourished.

The Sections have been grouped into three parts. The first (Sections A and B) provides an introduction to the subject. Section A outlines the factors leading to bioinformatics becoming an essential activity. Section B is a brief history of the subject (through a series of definitions of the term *bioinformatics*) from its origins in the 1960s through the exhilarating (if not intoxicating) 1990s to the twenty-first century, where bioinformatics is being applied to all types of biological information.

The second part comprises the building blocks of informatics (Sections C–I): physics, mathematics, and computer science. It contains one major omission, however. Computer programming is an essential skill in bioinformatics, but the space constraints here do not allow us to include adequate training in any useful language. Since this is a particularly practical activity, it is better to leave this to the many other books that are available. However, we have tried to outline the rudiments of good data management and programming practice.

The third part contains the biological application domains (Sections J–R). It comprises three subgroups: molecular biology; metabolism, anatomy, physiology; and complex information sources (specifically image datasets and natural-language texts). The latter remain the hardest places from which to extract accurate and quantitative data. The second and third parts interlock as shown in the figure below. This emphasizes the foundational importance of the building blocks. The consequent reliance on them from all of the application areas is clear from the tight mesh of interconnections linking the two together.



Interconnections between sections. The sections of the book for the building blocks and application areas are shown in two columns. The connections between them are shown by solid lines where they are linked by related topics, and dashed lines where the connections are implicit in the text.

Bioinformatics is now sufficiently broad that even the three main authors of this work felt it appropriate to draw in others to write particular Sections. Thus, we wish to thank several people for contributing: Sections J (Nicola Gold), K (Alex Marshall), L (Nicola Gold and Tom Gallagher), and M (Rob Linforth). Various people also checked the accuracy and clarity of various Sections, so we gratefully acknowledge the help of Alastair Middleton, Leah Band, Tom Gallagher, Kim Kenobi, and particularly Jane Hodgman (who also proofread many of the sections). We are grateful to members of the UK Centre for Plant Integrative Biology for providing microscope images ahead of publication. Readers are liable to find some duplication, but this is retained for the purposes of clarity. Finally, we hope that students and teachers alike will appreciate the breadth of the subject and enjoy reading this work.

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
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
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	资料、资讯、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	基因组调查序列
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	隐藏马尔可夫模型
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	网络协议
ISO	International Organization for Standardization	国际标准化组织
IT	information technology	信息技术
J RE	Java runtime environment Java	运行环境
KEGG	Kyoto Encyclopedia of Genes and Genomes	京都(日本)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	马尔可夫链蒙特卡尔理论
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	分子模拟数据库
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	孟德尔人类遗传在线数据库
OODB	object-orientated database	面向对象数据库
OOP	object-oriented programming	面向对象程序设计
ORF	open reading frame	开放阅读框
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 parsimony	采用简约法的系统发育分析
PCs	personal computers	个人电脑
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	定量结构-活性相关
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	基因表达连续分析法
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	简化分子输入条目规范的扩展
SMILES	Simplified Molecular Input Line Entry System	简化分子线性输入规范
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	结构式查询语言
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	通用描述、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

A　生物学研究方式的转变

要　点

介　绍

生物学的研究逐渐地多样化，并且专业分化与日俱增，这主要是由于20世纪70年代中期四个研究方式转变的驱动力的推动作用，生物学的研究日渐多样化和专门化。本节将依次讨论。

分子与万物

分子生物学与遗传学相结合是识别生物过程中组分的非常有效的研究方式，最初是针对基础生物化学过程，但最终几乎涉及每一个生物转化过程。

小型化与自动化

生物技术专家开发并且继续寻找从愈来愈少的生物样品中获取尽量多的信息的方法。机器人技术的出现同样带来了连续处理大量样品的能力。这些就是通常说的高通量技术。

图像分析

为了能够对从上述技术得到的大量数据（兆字节/样品）进行可靠的处理，现在原始的输出通常是由可以被计算机程序解释的图像组成。

计算和统计模型

通过对高通量技术所产生的大量数据都进行的一系列统计分析，确定生物学对象（基因和蛋白质等）的个体与群体的性质。生物过程源于这些对象之间的相互作用，计算方法的应用范围非常广泛，目前已经用来描述这些过程。这些生物过程的最简单形式包括生化、调控或遗传的相互作用网络。然而，更多的生物学现象（或体系）的动态和定量的行为要由更复杂的数学模型来描述。这些模型可用于进行“what-if”计算模拟（*in silico*）实验。模型的质量关系到它们的模拟性能，特别是对系统状态的预测。这些数学模型只有在现代高速的计算机上才有可能实现应用。这意味着现在的生物学正逐渐变得越来越像物理学，因为理论生物学家正赶上并有可能很快超越纯粹的实验生物学家。

这些研究方式转变的结果

这些变化的结果是使生物学家日益将更多时间花在数据分析上，而在实验本身上时间较少。这些改变同样导致了对于被称为生物情报员的人的需求迅速增长，他们能够及时以生物学有意义的方式熟练的管理和解释这些海量数据。章节B通过系列的定义讲述了这个学科的简短历史。

相关章节

生物信息学的定义（B）
概率和统计（G）
模建与数学技术（H）
转录物组学（K）
蛋白质与蛋白质组学技术（L）
图像分析（Q）

C 物理学要素

要 点

质量守恒	这条定律阐明系统质量是守恒的。意味着所有生物反应的反应物质量必须等于产物的质量。
热力学	热力学第一定律阐明，一个封闭系统的能量是恒定的。这样的描述类似于上一个定律，然而它是指一个化学反应释放的能量会变成热，反之亦然。第二定律阐明，一个不平衡系统的熵将倾向于增加，直至达到平衡。该定律在生物学中有非常重要的含义，当大分子将趋于降解成更小片段（熵增加）时，能量用在补偿生成小片段时造成的熵损失。第三定律阐明，熵将随着温度的降低而减少（在绝对零度时趋近极小值）。因此，大分子通常在低温环境里稳定。这项原则也被应用于计算机优化技术，最常见的就是模拟退火。
物理原理在计算中的应用	物理原理通过两个途径应用到计算中。首先，当计算机的零件温度上升的时候，物理学原理通过影响这些零件限制了计算机的计算能力；其次，物理学原理可以用于建立算法提高计算效率。例如，模拟退火搜索，它模拟了冷却的物理过程，有助于防止搜索陷入局部最小。
物理学的研究方法	几个世纪以来，物理学的研究方法有时被称为奥坎姆的剃刀。当较简单的描述无法胜任的时候，这种依据最低数量限度的组成（或概念）和它们之间关系进行的最简单描述的方法必定变得更加复杂。

相关章节 计算（F）

F 计　　算

要　点

中央处理器

中央处理器(CPU)主要由两个部分组成,控制单元和逻辑运算单元(ALU)。前者提取并执行包含基本操作的指令,例如加、减法。后者在数字上实行运算和逻辑测试。

寄存器

CPU 的寄存器组织的像一个阵列,其中的单元被称作寄存器,并且允许处理器储存少量用于计算的数据。这些寄存器与主存储器相比能更迅速地被访问。常常有好几种寄存器用于不同的用途,例如,用于存储数据的数据寄存器,存储存储器地址的寄存器、保持堆栈数据结构的栈寄存器,等等。

机器码

CPU 的控制单元执行机器代码指令。机器代码本身基本上是一组由汇编语言写的更易读的二进制编码,而汇编语言用助记码代替了原始的二进制。机器代码往往是特异针对特定类型的处理器,因此被称为本机(代)码。

高级语言

这种语言在风格上脱离机器代码,更像是自然语言。实际上从机器代码分离出来后,它们往往是更易于移植,因为它们可以在一个特定的计算机上被转换成本地机器代码。本章将举例介绍 Java、C++、Perl、Python 和 PHP。

面向对象的程序设计

Java 和 C++是面向对象的语言,这意味着,数据和处理指令是封装在"对象"之内,并且它们可以彼此联系。对象是依照种类来定义的,包含关于可以存储数据的属性和可以完成的处理信息。

指　针

指针指出数据的位置。传递指针到一个子程序或对象,而不是实际的数据,可以节省大量的处理时间。

正则表达式

正则表达式是描述寻找匹配的文本字符串搜索模式的一种方法。他们允许用户组合使用文本、通配符和特殊字符进行精确地指定他们搜索的一组字符串。

软件版本

作为开发中的软件,它有助于记住出现在限定步骤中的发展。有记载这些步骤的方法,使用主要和次要的版本号,"alpha"、"Beta"和"候选版本"等标记。在本节 CVS(concurrent versions system 并发版本系统)将作为一个处理软件版本的解决方案来讨论。

计算机网络

计算机之间相互通讯除了使用传输控制协议(TCP)以外,还使用其他协议。

多层计算

多层计算通常包括三个主要层面:顶层是用户界面层,中层逻辑层执行大部分逻辑处理,底层是数据层。这种方法的优点是可以在不同的机器上执行不同的层面,因此可以分散处理负载。

网格计算和 web 服务

网格计算依靠网络计算机的网格(grid)提供处理资源。任务提交给网格控制软件,在那里通过网格分配作业的处理要求,并汇编网格中的单个节点完成的结果。web 提供了一种软件与其他可通过 WWW 提供处理服务的软件进行联系的途径。

生物信息学软件的设计和管理的好习惯

这涉及软件开发中的设计,执行和维护方式。它也介绍了有关编程设计格式的一些想法,以及一些在生物信息学中良好的程序设计习惯。

相关章节

数据及数据库(D)
数据类型(E)
人工智能和机器学习的计算方法(I3)
文本分析(R)

G 概率与统计

G1 概率和概率分布

要 点

概 率

这是一个在生物信息学中关键的概念，对于解释生物分子数据非常重要。

事件的并发概率

这是一个乘法定则，$P(A \text{ and } B)=P(A)P(B)$，它在针对独立事件的众多概率论应用中有重要的地位。其他重要公式有：对于相互排斥的事件 $P(A \text{ or } B)=P(A)+P(B)$，以及 $P(\text{not } A)=1-P(A)$。

概率分布和密度函数

概率分布是一个数学函数给出的关于一个变量能得到某个特定的值或属于某个值域的几率。或许最简单的概率分布就是二项式分布，它给出在 N 次试验中的任何成功次数(R)的概率，每次实验只有两个结果，成功(概率$=P$)和失败(概率$=1-P$)。它适用于在连续投硬币的过程中产生正面朝上的次数。

泊松分布

和二项式分布相似，它是一个给出事件发生概率的函数，但是用一个参数 E 来表示期望的事件发生的次数。在测试的数量很大的时候它近似于二项式分布，并且任何一个测试的成功率非常低。

正态(高斯)分布

这是一个数学函数，它给出一个变量能获得连续(实数)值的概率。它通常表示为关于两个参数(均值和标准偏差)的钟形(bell-shaped)概率密度函数。概率是计算出来的密度函数曲线下面的面积。它在大量统计理论中是非常重要的。

相关章节

条件概率和贝叶斯法则(G2)
基本的统计学检验(G3)
人工智能和机器学习的统计方法(I2)
数据库和数据源(J1)
转录谱(K1)
相互作用蛋白质组学(L2)

G2 条件概率和贝叶斯法则

要 点

条件概率

这表达了事件可能不是独立的思想。一个已经发生的事件(A)可能会影响另一个事件(B)的发生概率，这可以用符号表示，$P(B|A)$所指的是假定 A 已经发生时 B 的重现概率。它产生了改进的概率乘法定则：$P(A \text{ and } B)=P(A)P(B|A)$。

贝叶斯法则

这是重新排列上述的乘法定则得到的：$P(A|B)=P(A)P(B|A)/P(B)$。

贝叶斯统计论

它使用了一个主观的概率解释。它广泛应用贝叶斯法则以这样的形式，认为事件 A 被视为一种假设 H，B 是被视为一些数据或观测 D，给出这样的法则 $P(H|D)=P(H)P(D|H)/P(D)$[①]。这使 H 为真的预先概率($P(H)$)与 H 为真的事后概率($P(H|D)$相关联，并受到观测数据的影响。

马尔可夫链模型

这是一个模拟序列状态的实用统计工具，例如，与 DNA 序列相关联的字母序列。序列中字母的同一性只依赖于紧邻它前面的字母的同一性，从而模拟了一些生物分子序列中的单相关性。

相关章节	概率与概率分布(G1)	序列的家族、对比和系统发育(J4)
	基本的统计学检验(G3)	多元技术和网络推理(K4)
	多元技术和序列分析(J3)	

注①:原书 57 页公式错误,应为 $P(H|D)=P(H)P(D|H)/P(D)$。

G3 基本的统计学检验

要 点

统计学和可变性

从根本上,统计学处理可变性并将它归结于不同原因。它对于许多生物学来源的可变性的量化是非常重要的,它包括实验的不精确,相同种类的不同生物体间的差异,以及环境中的差异。

测试的统计学意义

统计显著性检定的目的是辨别由那些具有真正生物学意义的随机变化所导致的影响。

t-检验和可能性

t-检验的目的是检验正态分布数据平均值之间的差异显著性。当数据不符合正态分布时,可以选择 Wilcoxon 检验和 Mann-Whitney 检验。

方差分析

当在实验中有超过两组数据的时候(t-检验只能处理两组数据),方差分析的目的在于量化组平均的统计差异显著性。

卡方检验和费歇尔精确检验

这些检验采用离散"计数"型数据,经常表现为关联表。他们旨在如果预期虚假设是真的,评估计数之间的统计差异显著性,以及每个类别中的实际观察计数。

基于再取样的检验

为了计算 P 值,假定虚假设是真的,那么这种检验就可以代替大多数标准的检验,并且通常包含了数据的随机取样。

多重检验

当执行一次以上的统计检验的时候,多重检验对于多重取样的校正是非常重要的,能给出小的错误 P 值。常见的校正模式是 Bonferroni 和 Benjamini-Hochberg 校正。

相关章节	概率与概率分布(G1)	转录分析的统计问题(K2)
	条件概率和贝叶斯法则(G2)	基因表达的差异分析(K3)
	序列分析(J3)	多元技术和网络推理(K4)

H 模型与数学技术

H1 系统特征

要 点

生物系统 一个生物系统是一个对象的集合，在大小上包括从分子到生物种群，它们在显示群体功能或作用的途径上互相影响。这种群体行为符合生理学、传染病学以及生物体的群体生物学。

模 型 模型是过程或对象的简化表示，模型可以在一组指定的环境下描述它们的行为。

系统特征 一个系统的属性是由一系列事物的各个属性组成的。使用的模拟技术类型取决于应用于系统研究中的事物(在标题下列出)。

抽 象 抽象是将某些生物学过程映射到以数学关系表示的一系列概念的过程。

相关章节

图论及其应用(H2)
常微分方程和代数学(H3)
高级模拟技术(H4)
形状、变形和成长(H5)

H2 图论及其应用

要 点

图 形 数学家将图形描述为一个由一系列的节点沿边线连接构成的假设结构。包含回路或交替路径的图形逐渐被称为网络。节点和边缘可以有一系列被定义为颜色的属性，它们可能有定量的值，称为权重。当描述基因调控通路的代谢信号转换时，生物学家通常是依照图形来思考。

图形的计算机表示 在计算机内部的图形总是被描绘成各种各样的矩阵，包括邻接矩阵、邻接表和化学计量矩阵。在计算机内有很多方法可以用程序描述图形，因此图形的结构和显著的特征更容易了解。

网络拓扑及其属性 整个网络的拓扑可以是随意无规则的或有特别属性的，使其对于干扰具有较强的坚固性和稳定性。生物学网络像因特网一样，有一个自由尺度的拓扑，用非常短的网络半径给出所包含的节点数。

通用换算表示法 节点没有必要在相同的尺度内描述生物学对象，因此可以用边线将一个节点(一个分子)链接到另一个表示细胞或组织的节点上。这条边线就象征这个分子对细胞或组织施加的影响。

皮特里网和P-系统 这些是为一系列特殊软件开发的特定类型的图形。它们有望被计算机科学家用来模拟广泛的生物学过程。

相关章节

基因表达的差异分析(K3)
相互作用数据库和网络(L3)
新陈代谢网络的研究(O1)
生理学(P)

H3 常微分方程和代数学

要 点

连续变化 本节是有关于平稳的变化，而不是跃变。它涉及的值可能是实数，而不仅仅是整数。