国外经典教材·计算机科学与技术

PEARSON
Addison Wesley

Fundamental Concepts of Bioinformatics

生物信息学概论

(美) Dan E. Krane & Michael L. Raymer 著
孙啸 陆祖宏 谢建明 等 译

清华大学出版社

China 需要全本请访问 www.ertongbook.com

Fundamental Concepts of Bioinformatics

对于生物信息学初学者，需要有一本很好的入门教材。本书就是为不同学科背景的学生学习生物信息学而精心设计的。作者编著本书的目的是为了使学生掌握生物信息学的基本概念和基础知识，能够像生物学家或计算机科学家一样地思考问题和解决问题。本书介绍了生物学和计算机应用的基础知识，提出了许多生物信息学的基本问题，阐述了生物信息学的基本分析方法。此外，本书还给出了许多实用算法和例程。本书是一本非常好的生物信息学概论或导论教材，可以作为生物信息学方向高年级本科生或研究生的入门教材。

ISBN 7-302-09430-6

9 787302 094302 >

定价:36.00元

PEARSON

Addison Wesley

封面设计：久久度文化
读者信箱：thjd@thjd.com.cn
信息网站：http://www.ePress.cn
http://www.34.cn
http://www.pearsoned.com

China

国外经典教材·计算机科学与技术

生物信息学概论

Fundamental Concepts of Bioinformatics

（美）Dan E. Krane & Michael L. Raymer 著

孙 啸 陆祖宏 谢建明 等译

清华大学出版社

北 京

China

内 容 简 介

本书是为不同学科背景的学生学习生物信息学而精心设计的一本入门教材，目的是使学生掌握生物信息学的基本概念和基础知识，学习生物学家或计算机科学家考虑问题和解决问题的方法。

本书介绍了生物学和计算机应用的基础知识，提出了许多生物信息学的基本问题，阐述了生物信息学的基本分析方法。此外，本书还给出了许多实用算法和例程。

本书是一本优秀的生物信息学概论或导论教材，可作为生物信息学方向高年级本科生或研究生的入门教材。

Simplified Chinese edition copyright © 2004 by PEARSON EDUCATION ASIA LIMITED and TSINGHUA UNIVERSITY PRESS.

Original English language title from Proprietor's edition of the Work.

Original English language title: Fundamental Concepts of Bioinformatics, 1e by Dan E.Krane & Michael L. Raymer Copyright © 2003

EISBN: 0-8053-4633-3

All Rights Reserved.

Published by arrangement with the original publisher, Pearson Education, Inc., publishing as Pearson Education

This edition is authorized for sale only in the People's Republic of China (excluding the Special Administrative Region of Hong Kong and Macao).

本书中文简体翻译版由 Pearson Education 授权给清华大学出版社在中国境内（不包括中国香港、澳门特别行政区）出版发行。

北京市版权局著作权合同登记号 图字：01-2003-1778

本书封面贴有 Pearson Education（培生教育出版集团）激光防伪标签，无标签者不得销售。

版权所有，翻印必究。举报电话：010 62782989 13901104297 13801310933

本书封面贴有清华大学出版社激光防伪标签，无标签者不得销售。

图书在版编目（CIP）数据

生物信息学概论/（美）奎恩（Krane, D.E.），（美）雷默（Raymer, M.L.）著；孙啸等译. 北京：清华大学出版社，2004.10

（国外经典教材·计算机科学与技术）

书名原文：Fundamental Concepts of Bioinformatics

ISBN 7-302-09430-6

Ⅰ.生… Ⅱ.①奎…②雷…③孙… Ⅲ.生物信息论-高等学校-教材 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字（2004）第 0908 号

出 版 者：清华大学出版社 **地 址**：北京清华大学学研大厦
http://www.tup.com.cn **邮 编**：100084
社 总 机：010-62770175 **客户服务**：010 62776969
组稿编辑：刘利民
文稿编辑：鲁秀敏
封面设计：久久度文化
版式设计：张红英
印 刷 者：北京密云胶印厂
装 订 者：三河市李旗庄少明装订厂
发 行 者：新华书店总店北京发行所
开 本：185×260 **印张**：20 **字数**：433 千字
版 次：2004 年 10 月第 1 版 2004 年 10 月第 1 次印刷
书 号：ISBN 7-302-09430-6·Q·39
印 数：1～4000
定 价：36.00 元

本书如存在文字不清、漏印以及缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：(010)62770175 3103 或(010)62795704

China

国外经典教材·计算机科学与技术
编审委员会

主任委员

孙家广　　清华大学教授

副主任委员

周立柱　　清华大学教授

委员（按姓氏笔画排序）

王成山　　天津大学教授

王　珊　　中国人民大学教授

冯少荣　　厦门大学教授

冯全源　　西南交通大学教授

刘乐善　　华中科技大学教授

刘腾红　　中南财经政法大学教授

吉根林　　南京师范大学教授

孙吉贵　　吉林大学教授

阮秋琦　　北京交通大学教授

何　晨　　上海交通大学教授

吴百锋　　复旦大学教授

李　彤　　云南大学教授

杨宗源　　华东师范大学教授

沈钧毅　　西安交通大学教授

邵志清　　华东理工大学教授

陈　纯　　浙江大学教授

陈　钟　　北京大学教授

陈道蓄　　南京大学教授

周伯生　　北京航空航天大学教授

孟祥旭　　山东大学教授

姚淑珍　　北京航空航天大学教授

徐佩霞　　中国科学技术大学教授

徐晓飞　　哈尔滨工业大学教授

秦小麟　　南京航空航天大学教授

钱培德　　苏州大学教授

曹元大　　北京理工大学教授

龚声蓉　　苏州大学教授

谢希仁　　中国人民解放军理工大学教授

China 需要全本请访问 www.ertongbook.com

出 版 说 明

近年来，我国高等学校的计算机学科教育进行了较大的改革，急需一批门类齐全、具有国际水平的计算机经典教材，以适应当前的教学需要。引进国外经典教材，可以了解并吸收国际先进的教学思想和教学方法，使我国的计算机学科教育能够与国际接轨，从而培育更多具有国际水准的计算机专业人才，增强我国信息产业的核心竞争力。Pearson、Thomson、McGraw-Hill、Springer、John Wiley 等出版集团都是全球最有影响的图书出版机构，它们在高等教育领域也都有着不凡的表现，为全世界的高等学校计算机教学提供了大量的优秀教材。为了满足我国高等学校计算机学科的教学需要，我社计划从这些知名的国外出版集团引进计算机学科经典教材。

为了保证引进版教材的质量，我们在全国范围内组织并成立了“清华大学计算机外版教材编审委员会”（以下简称“编委会”），旨在对引进教材进行审定、对教材翻译质量进行评审。“编委会”成员皆为全国各类重点院校教学与科研第一线的知名教授，其中许多教授为各校相关院、系的院长或系主任。“编委会”一致认为，引进版教材要能够满足国内各高校计算机教学与国际接轨的需要，要有特色风格，有创新性、先进性、示范性和一定的前瞻性，是真正的经典教材。为了保证外版教材的翻译质量，我们聘请了高校计算机相关专业教学与科研第一线的教师及相关领域的专家担当译者，其中许多译者为海外留学回国人员。为了尽可能地保留与发扬教材原著的精华，在经过翻译和编辑加工之后，由“编委会”成员对文稿进行审定，以最大程度地弥补和修正在前面一系列加工过程中对教材造成的误差和瑕疵。

由于时间紧迫和能力所限，本套外版教材在出版过程中还可能存在一些不足和遗憾，欢迎广大师生批评指正。同时，也欢迎读者朋友积极向我们推荐各类优秀的国外计算机教材，共同为我国高等学校的计算机教育事业贡献力量。

清华大学出版社

China

译者序

随着人类基因组计划的顺利实施和各种后基因组计划的开始，涌现出海量的生物分子数据。目前，生物分子数据量每 15 个月就要翻一番。这些生物分子数据具有丰富的内涵，其背后隐藏着人类目前尚不知道的生物学知识。充分利用这些数据，通过数据分析，揭示这些数据的内涵，得到对人类有用的信息，是科学家们所面临的一个严峻的挑战。生物信息学就是为迎接这种挑战而发展起来的一门新型学科，它是由生物学、数学、计算机科学相互交叉所形成的学科，是当今生命科学和自然科学的重大前沿领域之一。

在生物信息学这样一个新型交叉学科研究领域中，由生物学家、数学家和计算机科学家共同研究生物分子信息的获取、管理、分析和利用。生物信息学以计算机、网络为工具，利用数学和信息科学的理论、方法和技术去研究生物大分子，研究生物分子信息组织和表达的规律。生物分子是信息的载体，生物分子至少携带着 3 种重要的信息，即遗传信息、与功能相关的结构信息以及进化信息。通过处理、分析和挖掘生物分子数据，认识生物分子的作用，探索人类生老病死的规律，了解生命的起源和进化，进而加深对生物世界的认识。在生物学、医学领域中利用生物分子数据及其分析结果，可以大大提高研究和应用的科学性及效率。目前，生物信息学研究的重点主要是 DNA 分子和蛋白质分子两个方面，包括研究它们的序列、结构和功能，并以系统的观念去分析它们之间的相互作用。

人类基因组计划产生的生物分子数据是生物信息学的源泉，而人类基因组计划所需要解决的问题则是生物信息学发展的动力。生物信息学已经在基因组学研究中发挥了突出的作用。然而，这仅仅是开始。在后基因组研究中，生物信息学将迎接更大的挑战，也必将发挥更大的作用。

人才培养是生物信息学目前的一个重要任务。国际上对生物信息学专业人才的需求量比较大，而在国内，随着基因组及后基因组研究工作的不断开展和推进，也迫切需要相关的人才。从事生物信息学工作的人员需要理解生物学和计算机科学，需要将两者在实际工作中紧密地结合起来。然而，目前从事生物信息学研究的人员一般来自于生物学专业，或来自于信息学专业，很少同时学习这两个科学。对于来自不同专业的研究人员需要补充另外一个专业的基础知识，需要将两个专业的知识融合起来。

对于生物信息学初学者，需要有一本很好的入门教材。本书就是为不同学科背景的学生学习生物信息学而精心设计的。作者编著本书的目的是为了使学生掌握生物信息学的基本概念和基础知识，学习生物学家或计算机科学家思考问题和解决问题的方法。本书介绍了生物学和计算机应用的基础知识，提出了许多生物信息学的基本问题，阐述了生物信息学的基本分析方法。此外，本书还给出了许多实用算法和例程。本书是一本非常好的生物

China

信息学概论或导论教材，可以作为生物信息学方向高年级本科生或研究生的入门教材。

本书的翻译工作由东南大学吴健雄实验室完成。谢建明、李敏俐、傅静、周士新、谢雪英、周童、顾方君、徐伟、焦典、刘志华、董献军等同志参加了本书的翻译工作，陶怡、汤丽华、韦芬霞等同志参加了本书文字和图表校对工作。全书由孙啸统一核实、修改和整理，由陆祖宏审定所有内容。

在此，对所有参加本书翻译工作的人员表示衷心的感谢。

孙　啸

2003 年 12 月于东南大学

China

前　言

编著本书的初衷是为那些对生物信息学感兴趣的大学生提供一本教科书。我们相信，生物信息学领域中最好的工作来自于两个学科高级研究人员的通力合作，即生物学和计算机科学。这两个学科很少有共同语言，研究方法也不一样，即使在大学或学院中所处的物理位置也截然分开。

生物学家希望从互联网的各个网站上找到那些能够用于分析他们实验数据的生物信息学软件，然而关于这些网站的分类（或介绍）的工具书非常少。同样，为计算机研究人员所写的生物信息学书也很少，计算机研究人员希望从这些专业书中了解设计生物数据分析算法的策略。一个最基本的问题是缺少为学生所准备的教科书。有一些书着重指导生命科学学生如何使用网络程序，但是不教授学生如何理解程序中各个变量的含义，以及如何在特定的应用中综合使用这些程序。一些为计算机研究人员编写的书介绍利用现有算法编制程序，然而却忽略了阐述这些算法的生物学意义。从事生物信息学工作需要理解生物学和计算机科学，而学生很少同时学习这两门科学。

从事生物信息学研究的人员需要有很好的基础和丰富的知识，生物信息学教育（包括高等通识教育）已经呈现出跨越单科教育的趋势。许多生物信息学研究生教育计划面向人才市场的需求及学生的要求，为生物学或计算机科学的大学毕业生提供另一门学科（计算机科学或生物学）方面广泛的培训。这些生物信息学专业研究生教育计划的时限比传统研究生教育计划的时限至少多一年，通常多两年。编著本书的目的是使学生能够像生物学家或计算机科学家一样去思考问题和解决问题。本书围绕生物信息学的核心算法进行论述和介绍，目标是使生物学和计算机研究人员具有共同的语言，具有相互理解的基础。本书在简要介绍分子生物学和化学之后，详细分析了各种算法，从不同角度展示计算机为分析海量并且非常复杂的数据提供难以估价的帮助。本书介绍了现有数据库，鼓励学生利用实际数据评测自身的技能。本书中虽然有许多数学和统计学计算，但是都比较简单，容易理解，不需要微积分基础。本书还提供各种例程，包括目前各种基于环球网的软件实例。

虽然许多针对分子生物学问题的新算法和新方法不断涌现，但是，计算机学家和生物学家一系列基本的概念和技术也可以解决实际问题。深入理解和熟练掌握这些基础内容，是具有不同学科背景的学生在计算分子生物学和生物信息学研究方面取得成功的基本条件。本书为不同学科背景的学生学习生物信息学进行了精心的设计，具有以下一些特点。

- 采用一种独立于计算机语言的解决问题的方法，使得学生在学习生物信息学基本算法时不必了解详细的程序语法，不必了解特定的计算机语言或操作系统。
- 使用一种特别的方法展现各种素材以及各章最后的提问，这些提问与正文中的实

China

例问题相对应。本书提供对编号为奇数问题的答案，而对于其他问题的答案，可以在教师网站上找到。

- 每一章分别阐述生物数据搜集和实验中的有关问题，同样也说明算法中的有关问题，这样，具有不同专业背景的学生能够正确评价生物信息学两个方面的可变性和复杂性。
- 附录 A 着重介绍了 Perl 编程，作为非专业程序员的入门向导。还强调了结构化程序设计语言的一般概念。掌握这些知识以后，可以很容易地学习新的程序设计语言及编程方法。
- 附录 C 提供了一些用 Perl 语言实现的示例算法，以便阐明本书各章的关键概念。这些例程避免使用 Perl 语言的特异命令，而采用一般性的问题解决策略和算法技巧，这些一般性的策略和技巧可以被用于解决各种生物信息学问题。

本书的编著工作比我们预先想象的要困难得多。我们没有打算让本书成为生物信息学领域中各种文献汇集在一起的一本百科全书，但是，我们仍然引用了许多优秀的文章，也得益于各方面大力的支持和帮助。David Paoletti 出色地设计出附录 C 中的算法；在 Wright 州立大学将本书作为本科生生物信息学课程教材时，许多学生对本书最初的文字、图表及实例提出了反馈意见，在此向他们表示衷心的感谢。同样感谢 Michele Sordi，在他的鼓励和指引之下，本书的编著由设想成为现实。感谢 Peggy Williams 创造性的设想、积极的支持和出色的才智。本书的成功来自于作品的编辑 Larry Lazopoulos，来自于 Matrix 公司的高级职员，如 Merrill Peterson 和 Michele Ostovar。最后，感谢我们的妻子，Carissa M.Krane 和 Delia F.N.Raymer，感谢她们的耐心和理解，在那段繁忙的日子里，写书成了我们家庭中最优先的事情。

China

目　录

第 1 章　分子生物学和生物化学 1
1.1　遗传物质 2
1.1.1　核苷酸 2
1.1.2　取向 2
1.1.3　碱基配对 3
1.1.4　分子生物学的中心法则 5
1.2　基因结构和遗传信息 6
1.2.1　启动子序列 6
1.2.2　遗传密码 7
1.2.3　开放阅读框 9
1.2.4　内含子和外显子 9
1.3　蛋白质的结构与功能 11
1.3.1　一级结构 12
1.3.2　二级、三级和四级结构 12
1.4　化学键的本质 14
1.4.1　原子结构 15
1.4.2　化合价 15
1.4.3　电负性 16
1.4.4　亲水性与疏水性 16
1.5　分子生物学工具 17
1.5.1　限制性酶消化 17
1.5.2　凝胶电泳 18
1.5.3　印迹和杂交 19
1.5.4　克隆 19
1.5.5　聚合酶链式反应 21
1.5.6　DNA 测序 22
1.6　基因组信息 23
1.6.1　C 值悖论 23
1.6.2　复性动力学 24
本章总结 25

China 需要全本请访问 www.ertongbook.com

阅读资料.....25
问题.....26

第 2 章 数据库搜索与两两比对.....28
2.1 点阵图.....29
2.2 简单比对.....30
2.3 空位.....31
2.3.1 简单空位罚分.....31
2.3.2 起始罚分与长度罚分.....32
2.4 打分矩阵.....32
2.5 动态规划：Needleman 和 Wunsch 算法.....35
2.6 全局比对与局部比对.....39
2.6.1 准全局比对.....39
2.6.2 Smith-Waterman 算法.....40
2.7 数据库搜索.....42
2.7.1 BLAST 及其家族.....42
2.7.2 FASTA 及其相关算法.....43
2.7.3 数据库搜索的比对得分与统计显著性.....44
2.8 多重序列比对.....45
本章总结.....46
阅读资料.....46
问题.....47

第 3 章 替换模式.....49
3.1 基因内的替换模式.....50
3.1.1 突变率.....50
3.1.2 功能约束.....51
3.1.3 同义和异义替换.....52
3.1.4 插入删除和伪基因.....53
3.1.5 替换和突变.....54
3.1.6 固定.....54
3.2 估算替换数目.....56
3.2.1 Jukes-Cantor 模型.....56
3.2.2 转换和颠换.....57
3.2.3 Kimura 的双参数模型.....57
3.2.4 多参数模型.....58
3.2.5 蛋白质序列间的替换.....59

China

3.3　基因间进化率的变化 60
3.4　分子时钟 61
3.4.1　相对速率检测 62
3.4.2　家系中变异率变化的原因 63
3.5　细胞器的进化 63
本章总结 63
阅读资料 64
问题 65

第 4 章　基于距离的系统发生分析 66
4.1　分子系统发生学的历史 67
4.2　分子系统发生分析的优点 68
4.3　系统发生树 69
4.3.1　重建系统发生树的术语 69
4.3.2　有根树和无根树 70
4.3.3　基因树和物种树 71
4.3.4　特征和距离数据 72
4.4　距离矩阵法 73
4.4.1　非加权组平均法（UPGMA） 73
4.4.2　分支长度的估计 75
4.4.3　距离变换法 76
4.4.4　近邻关系法 78
4.4.5　邻近归并法 79
4.5　最大似然法 79
4.6　多重序列比对 80
本章总结 80
阅读资料 81
问题 81

第 5 章　基于特征的系统发生分析 83
5.1　简约法 84
5.1.1　信息位点和非信息位点 84
5.1.2　无权简约法 87
5.1.3　加权简约法 88
5.2　推断出的祖先序列 89
5.3　快速搜索策略 89
5.3.1　分支约束法 89

China

5.3.2 启发式搜索 91
5.4 一致树 92
5.5 树的置信度 93
5.5.1 自举检验 93
5.5.2 参数检验 94
5.6 各种系统发生分析方法的比较 95
5.7 分子系统发生分析 95
5.7.1 生命之树 96
5.7.2 人类起源 97
本章总结 97
阅读资料 97
问题 98

第 6 章 基因组学和基因识别 100
6.1 原核基因组 101
6.2 原核基因结构 103
6.2.1 启动子元件 103
6.2.2 开放阅读框 106
6.2.3 假想翻译 106
6.2.4 终止序列 107
6.3 原核基因组中的 GC 含量 108
6.4 原核基因密度 108
6.5 真核生物基因组 109
6.6 真核生物基因结构 110
6.6.1 启动子元件 111
6.6.2 调控蛋白结合位点 113
6.7 开放阅读框 114
6.7.1 内含子和外显子 114
6.7.2 可变剪接 116
6.8 真核基因组中的 GC 含量 117
6.8.1 CpG 岛 118
6.8.2 等值区 120
6.8.3 密码子使用偏性 121
6.9 基因表达 122
6.9.1 cDNAs 和 ESTs 122
6.9.2 基因表达的串行分析 124
6.9.3 微阵列 124

China

6.10　转座 126
6.11　重复元件 126
6.12　真核基因密度 128
本章总结 128
阅读资料 129
问题 130

第 7 章　蛋白质和 RNA 结构预测 132
7.1　氨基酸 133
7.2　多肽的组成 136
7.3　二级结构 137
7.3.1　骨架的柔性，Φ 和 Ψ 137
7.3.2　预测的准确度 138
7.3.3　Chou-Fasman 和 GOR 方法 139
7.4　三级结构和四级结构 141
7.4.1　疏水性 142
7.4.2　二硫键 142
7.4.3　活性结构与最稳定的结构 143
7.5　蛋白折叠建模算法 143
7.5.1　网格模型 144
7.5.2　去网格模型 146
7.5.3　能量函数和优化 147
7.6　结构预测 149
7.6.1　比较建模 149
7.6.2　线索法：反向蛋白折叠 150
7.7　预测 RNA 二级结构 151
本章总结 152
阅读资料 153
问题 154

第 8 章　蛋白质组学 155
8.1　从基因组到蛋白质组 156
8.2　蛋白质分类 157
8.2.1　酶的命名 157
8.2.2　家族和超家族 158
8.2.3　折叠 159
8.3　实验技术 159

China

8.3.1 二维凝胶电泳 159
8.3.2 质谱分析 160
8.3.3 蛋白质微阵列 161
8.4 抑制剂和药物设计 162
8.5 配体筛选 163
8.5.1 配体对接 163
8.5.2 数据库筛选 164
8.6 X 射线晶体结构 166
8.7 核磁共振结构 169
8.8 经验方法和预测技术 170
8.9 翻译后修饰的预测 171
8.9.1 蛋白质分类 171
8.9.2 蛋白酶剪切 173
8.9.3 糖基化 173
8.9.4 磷酸化 174
本章总结 175
阅读资料 175
问题 176

附录 A 计算机编程和数据结构的简介 178
A.1 建立和执行计算机程序 179
A.2 变量和值 180
A.2.1 数据类型 180
A.2.2 基本运算 181
A.3 程序控制 182
A.3.1 语句和程序块 182
A.3.2 条件执行 183
A.3.3 循环 186
A.4 可读性 188
A.4.1 结构化编程 188
A.4.2 注释 189
A.4.3 描述性变量名 190
A.5 数据结构 190
A.5.1 数组 190
A.5.2 哈希表 192
A.5.3 串的操作 192
A.6 子程序和函数 194

China

A.7　输入和输出 199
A.8　正则表达式 201
A.9　下一步怎么办 202
阅读资料 203
问题 204

附录 B　酶动力学 205
B.1　酶是生物催化剂 206
B.2　Henri-Michaelis-Menten 方程 207
B.2.1　V_{max} 和 K_m 208
B.2.2　直接作图法 209
B.2.3　Lineweaver-Burk 双倒数作图法 211
B.2.4　Eadie-Hofstee 作图法 211
B.3　简单抑制系统 212
B.3.1　竞争性抑制作用 212
B.3.2　非竞争性抑制作用 213
B.3.3　不可逆性抑制作用 214
B.3.4　pH 和温度对酶促反应速度的影响 215
阅读资料 215
问题 215

附录 C　PERL 程序举例 216
例 1：假想翻译 217
例 2：点阵图 219
例 3：相对速率测试 223
例 4：UPGMA 224
例 5：共同的祖先 227
例 6：剪接点识别 232
例 7：疏水性计算——2D-HP 模型 235
例 8：原子的空间密度计算 240
例 9：酶动力学——线性回归 243

词汇表 246

部分问题的答案 287

China 需要全本请访问 www.ertongbook.com