“十一五”国家重点图书　　计算机科学与技术学科前沿丛书
计算机科学与技术学科研究生系列教材（中文版）

生物信息学导论
——面向高性能计算的算法与应用

王勇献　王正华　编著

清华大学出版社

生物信息学导论

“十一五”国家重点图书　　计算机科学与技术学科前沿丛书
计算机科学与技术学科研究生系列教材（中文版）

生物信息学导论
——面向高性能计算的算法与应用

王勇献　王正华　编著

清华大学出版社
北京

内 容 简 介

本书主要针对生物信息学中的典型应用，从计算方法角度介绍相关算法的原理及应用；内容分成生物学及数理基础、生物序列分析、蛋白质组学分析以及大规模生物学网络分析等四个专题，涉及生物分子序列分析、基因发现、分子进化分析、蛋白质结构预测、蛋白质肽测序、生物学网络模块划分等具体问题的求解原理及算法实现。

本书的读者对象是具有现代分子生物学及计算机科学基本知识的研究生及相关科研人员，在附加习题后也可作为生物信息学方面的入门及进阶教材，供生物医学工程、计算机应用等专业学生使用。

图书在版编目(CIP)数据

生物信息学导论：面向高性能计算的算法与应用/王勇献，王正华编著. —北京：清华大学出版社，2011.6

（计算机科学与技术学科前沿丛书　计算机科学与技术学科研究生系列教材（中文版））

ISBN 978-7-302-25022-7

Ⅰ. ①生…　Ⅱ. ①王…　②王…　Ⅲ. ①生物信息论-研究生-教材　Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字 (2011) 第 045333 号

责任编辑：焦　虹　徐跃进
责任校对：李建庄
责任印制：何　芊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62795954，jsjjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：32.5　**字　数**：810 千字
版　　次：2011 年 6 月第 1 版　　**印　　次**：2011 年 6 月第1 次印刷
印　　数：1～3000
定　　价：49.00 元

产品编号：040630-01

“十一五”国家重点图书　计算机科学与技术学科前沿丛书

计 算 机 科 学 与 技 术 学 科 研 究 生 系 列 教 材

序

未来的社会是信息化的社会，计算机科学与技术在其中占据了最重要的地位，这对高素质创新型计算机人才的培养提出了迫切的要求。计算机科学与技术已经成为一门基础技术学科，理论性和技术性都很强。与传统的数学、物理和化学等基础学科相比，该学科的教育工作者既要培养学科理论研究和基本系统的开发人才，还要培养应用系统开发人才，甚至是应用人才。从层次上来讲，则需要培养系统的设计、实现、使用与维护等各个层次的人才。这就要求我国的计算机教育按照定位的需要，从知识、能力、素质三个方面进行人才培养。

硕士研究生的教育须突出“研究”，要加强理论基础的教育和科研能力的训练，使学生能够站在一定的高度去分析研究问题、解决问题。硕士研究生要通过课程的学习，进一步提高理论水平，为今后的研究和发展打下坚实的基础；通过相应的研究及学位论文撰写工作来接受全面的科研训练，了解科学研究的艰辛和科研工作者的奉献精神，培养良好的科研作风，锻炼攻关能力，养成协作精神。

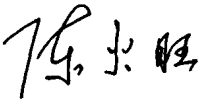
高素质创新型计算机人才应具有较强的实践能力，教学与科研相结合是培养实践能力的有效途径。高水平人才的培养是通过被培养者的高水平学术成果来反映的，而高水平的学术成果主要来源于大量高水平的科研。高水平的科研还为教学活动提供了最先进的高新技术平台和创造性的工作环境，使学生得以接触最先进的计算机理论、技术和环境。高水平的科研也为高水平人才的素质教育提供了良好的物质基础。

为提高高等院校的教学质量，教育部最近实施了精品课程建设工程。由于教材是提高教学质量的关键，必须加快教材建设的步伐。为适应学科的快速发展和培养方案的需要，要采取多种措施鼓励从事前沿研究的学者参与教材的编写和更新，在教材中反映学科前沿的研究成果与发展趋势，以高水平的科研促进教材建设。同时应适当引进国外先进的原版教材，确保所有教学环节充分反映计算机学科与产业的前沿研究水平，并与未来的发展趋势相协调。

中国计算机学会教育专业委员会在清华大学出版社的大力支持下，进行了计算机科学与技术学科硕士研究生培养的系统研究。在此基础上组织来自多所全国重点大学的计算机专家和教授们编写和出版了本系列教材。作者们以自己多年来丰富的教学和科研经验为基础，认真研究和结合我国计算机科学与技术学科硕士研究生教育的特点，力图使本系列教材对我国计算机科学与技术学科硕士研究生的教学方法和教学内容的改革起引导作用。本系列教材的系统性和理论性强，学术水平高，反映科技新发展，具有合适的深度和广度。同时本系列教材两种语种（中文、英文）并存，三种版权（本版、外版、

合作出版）形式并存，这在系列教材的出版上走出了一条新路。

相信本系列教材的出版，能够对提高我国计算机硕士研究生教材的整体水平，进而对我国大学的计算机科学与技术硕士研究生教育以及培养高素质创新型计算机人才产生积极的促进作用。



序　言

生物信息学 (又称计算生物学) 是计算机、生物、数学等多学科交叉而新兴的学科分支，以借助计算机科学、信息科学等领域的算法与工具解决生命科学中的问题为主要特征。随着生命科学研究的深入和后基因组时代的来临，生物信息学所研究的问题已经发生了巨大的变化，新的研究越来越需要借助于高性能计算环境的支持，然而综合生命科学与高性能计算两个领域的知识、有效解决现实中的应用问题并不是一件容易的事，目前也缺乏系统深入的相关图书资料。编写这本《生物信息学导论》，作者希望在向读者介绍生物信息学中与高性能计算结合最密切的一些基础性问题，讨论并总结相关的求解算法与应用技术。

尽管本书定位在生物信息学“导论”的层次，但是作者并没有打算从生物信息学领域的概念内涵、发展现状入手，也没有追求内容上的面面俱到、或者对生物信息学领域内容进行泛泛介绍，而是有所选择地介绍了生物分子序列分析、基因发现、分子进化分析、蛋白质结构预测、蛋白质肽测序、生物学网络模块划分等具体问题的求解算法及原理实现。从这个意义上讲，也可以将本书看成是生物信息学部分专题内容的汇集。在每部分专题内容中，既有对经典方法的详细讨论，也融入了作者及其合作研究者最近几年研究的创新成果，既注重理论方面的方法 (例如：利用谱分析挖掘 PPI 网络中的典型模式)，也强调具体应用方面的实现 (例如：利用 MPI 实现并行计算)。值得说明的是，结合作者知识背景与研究兴趣，本书在介绍各类生物信息学问题的求解方法时，特别关注了如何跟高性能计算技术相结合 (例如：关于序列比对的并行计算)。

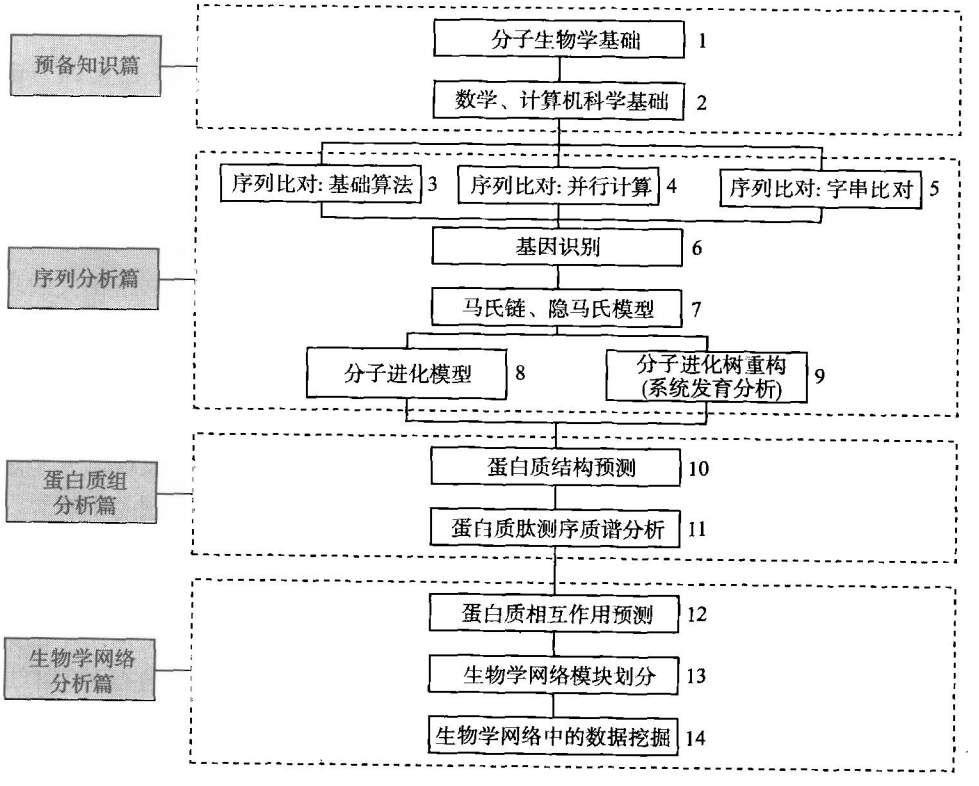
全书正文各章节结构如下图所示，共分为“预备知识篇”、“序列分析篇”、“蛋白质组分析篇”和“生物学网络分析篇”等四部分。

“预备知识篇”(包括第 1 章和第 2 章) 提供了生物信息学分析中涉及的常用分子生物学、线性代数、概率统计以及计算机算法等方面的基础知识，不同知识背景的读者可以选择自己所需的内容。读者阅读时也可以直接跳过这两章，在后续章节需要相关知识时，再回头翻阅这些内容。

“序列分析篇”是全书的重点内容之一，共包括七章内容 (第 3 章到第 9 章)，主要涵盖生物大分子的序列比对、DNA 序列上的基因识别 (或称基因发现) 和分子系统发育分析等。其中，序列比对共有三章，分别从经典基础算法、高性能并行实现以及基于字符串的模式匹配三个层面进行介绍；分子系统发育分析共有两章，分别介绍分子进化模型和进化树重构方法等内容；作为序列分析中最常用的一种模型，我们在本篇内容中专门介绍了马氏链和隐马氏模型 (第 7 章)，该模型在后续各篇中也有应用。

“蛋白质组分析篇”包括两章内容 (第 10 章和第 11 章)，分别介绍了基于序列预测蛋白质结构的计算分析，以及基于质谱数据分析的蛋白质序列测定方法等内容。

“生物学网络分析篇”是全书的最后一部分，共包含三章内容 (第 12 章到第 14 章)，重点以蛋白质相互作用为例，介绍了蛋白质相互作用预测、蛋白质相互作用网络的模块划分与功能预测等应用中的计算方法，最后一章系统总结了一般生物学网络中的数据挖掘方法与结果。



本书的读者对象是具有现代分子生物学及计算机科学基本知识的研究生及相关科研人员，在附加习题后也可作为生物信息学方面的入门及进阶教材，供分子生物学、计算机应用等专业的读者使用。为了照顾不同学科背景知识读者的需求，本书在开头还简要介绍了阅读后续章节所需要的分子生物学、数学及计算机科学基础知识；全书最后附有详细的主题索引、人名索引及以字母排序的参考文献，每条文献还特意标注了在正文中被引用处的页码，以方便读者检索。

本书是在为国防科技大学硕士研究生开设的“生物信息学导论”课程讲义以及作者从事国家自然科学基金项目研究成果的基础上整理而成。作者感谢 2005—2010 各学年选修这门课程学习的所有同学，他们参与课程讨论的许多内容构成了本书的基本素材；本书在选题与出版方面得到了国家自然科学基金 (60603054)、湖南省自然科学基金 (08JJ4021) 及国家重点基础研究发展计划课题 (2009CB723803) 的资助。

在书稿准备出版过程中，清华大学出版社的广大员工给予了大力支持与帮助，在此

一并表示感谢。

全书由王勇献主编，王正华对内容进行了统稿并提出了改进意见。由于作者水平所限，书中还有很多错误和不足之处，希望读者批评指正 (作者联系方式：湖南长沙国防科技大学计算机学院，邮编：410073，电子邮箱：yxwang@nudt.edu.cn)。

作者于长沙

2010 年 12 月

目 录