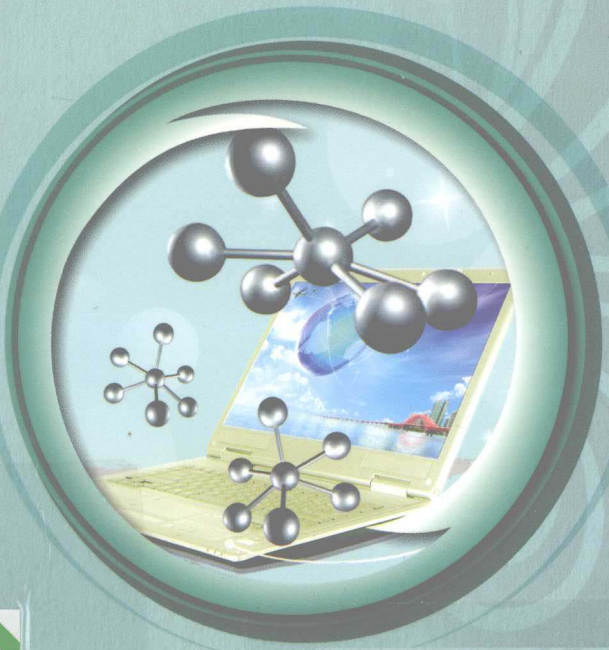
生物信息学中计算机技术应用

陈绮 ◎ 编著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

生物信息学中计算机技术应用

生物信息学中
计算机技术应用

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书结合计算机图形图像处理、模式识别、数据挖掘等技术，利用当前关于蛋白质结构研究的最新成果，讨论蛋白质三维结构比较中兼顾全局比较与部分比较的关键问题，有效解决了蛋白质三维结构的多层次比较问题。内容包括蛋白质三维结构的可视化、蛋白质三维结构统一坐标系的建立、统一坐标系下蛋白质结构频谱建立、基于灰色关联的蛋白质三维结构相似性算法研究，以及蛋白质三维结构空间复杂性的分形研究。

本书可作为大学高年级本科生及硕士生的参考书，也可作为科研人员的参考资料。

图书在版编目（CIP）数据

生物信息学中计算机技术应用/陈绮编著.—北京：电子工业出版社，2010.5

ISBN 978-7-121-10731-3

Ⅰ. ①生… Ⅱ. ①陈… Ⅲ. ①计算机应用－生物信息论 Ⅳ. ①Q811.4-39

中国版本图书馆 CIP 数据核字（2010）第 070996 号

策划编辑：董亚峰
责任编辑：李 蕊
印 刷：北京天宇星印刷厂
装 订：涿州市桃园装订有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：850×1168 1/32 印张：7.125 字数：166 千字
印 次：2010 年 5 月第 1 次印刷
印 数：2000 册 定价：25.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

2001 年 2 月，人类基因组序列图谱公开发表，这意味着后基因时代的到来。当基因序列在细胞中的角色都定位清楚后，人们才能真正理解其价值。人类后基因组计划由序列（结构）基因组学向功能基因组学转移，生命科学的研究重心从基因组学（Genomics）转变为蛋白质组学（Proteomics），其中心任务是阐明基因组所表现的真正执行生命活动的全部蛋白质的表达规律和生物功能，其目的是对基因组生物学功能进行研究和应用。

目前，有关蛋白质的各种研究如火如荼，但是人类对蛋白质的认识还只是冰山一角。对基因所表达的各种蛋白质都通过实验进行研究是不切实际的，新测得的蛋白质可通过与已知特性的蛋白质进行结构或序列比较后推知其生物功能。在生物的演化过程中，蛋白质的结构比其序列更保守，序列变化不一定会改变蛋白质的结构，但相似结构的蛋白质却可能具有不同的序列，而具有相似结构的蛋白质往往具有相似的功能，因此结构比较更受重视。在活性细胞中，蛋白质通过与其他分子的适当结合，执行了几乎全部的主要功能，蛋白质的结构对其功能具有重要的意义。结构基因组学（Structure Genomics）就是在此条件下蓬勃发展的，通过不断提高对蛋白质结构的认识，可以改进和完善由结构比较推测新蛋白质的方法。

蛋白质结构的相关研究可帮助生物学家鉴定新的蛋白质、对蛋白质进行分类、预测蛋白质的功能、进行同源分析及辅助药物设计等，在医药、农业、畜牧业、微生物应用及人类健康和社会

经济等领域具有重大意义。

人类基因组计划的目的之一在于阐明人类约 10 万种蛋白质的结构、功能、相互作用及与各种人类疾病之间的关系。蛋白质的三维结构与功能有着密切的关系，对蛋白质结构的研究是蛋白质组学中最核心、最基本的问题，而蛋白质三维结构的比较研究却属于三维物体形状比较的一个分支。本文针对蛋白质三维结构比较中兼顾全局比较与局部比较的关键技术和主要算法进行了深入的研究，将数据挖掘技术与图形图像处理技术有机地结合起来，研究了蛋白质三维结构数据的规范化处理、蛋白质三维结构特征提取、蛋白质空间结构多层次比较等关键技术。

本书可作为大学高年级本科生及硕士生的参考书，也可作为科研人员的参考资料。

本书的编著得到了海南省自然科学基金项目（609003）、海南大学科研项目（hd09xm84），以及海南大学科研基金资助，在此表示深深的谢意！由于本人的研究水平和阅历有限，书中错漏之处在所难免，恳请读者不吝赐教，在此表示衷心的感谢！如发现问题，请与作者联系，联系地址为海南大学信息科学与技术学院（邮编 570228）。

陈　绮

于海南大学

目　录

第1章 绪 论

本章主要介绍生物信息学中计算机技术的应用、研究的背景、意义、研究内容及全书的组织结构，并介绍生物系统的背景知识。

1.1 引言

生物信息学（Bioinformatics）这一名词最早出现于 1991 年的电子出版物中，是近年来新兴并正蓬勃发展的一门新学科。关于生物信息学的定义，学者众说纷纭，但其要旨是用数理和信息科学的观点、理论和方法去研究生命现象、组织和分析呈现指数增长的生物学数据的一门学科。它研究遗传物质的载体 DNA 及其编码的大分子蛋白质，以计算机为其主要工具，发展各种软件，对逐渐增长的 DNA 和蛋白质的序列和结构进行收集、整理、存储、发布、提取、加工、分析和研究，目的在于通过这样的分析逐步认识生命的起源、进化、遗传和发育的本质，破译隐藏在 DNA 序列中的遗传语言，揭示人体生理和病理过程的分子基础，为人类疾病的预测、诊断、预防和治疗提供最合理和有效的途径[Wu02]。

1990 年 10 月 1 日，美国国会正式批准“人类基因组计划”（Human Genome Project，HGP），标志人类历史上规模最大的科

研工程的正式启动。至今，世界上许多国家都已相继成立了一大批具有影响力的生物信息学中心。2001 年 2 月，人类基因组序列图谱公开发表，这意味着后基因时代的到来。人类后基因组计划是由序列（结构）基因组学向功能基因组学的转移，其目的是要对基因组生物学功能进行研究和应用，阐明人类约 10 万种蛋白质的结构、功能、相互作用，以及与各种人类疾病之间的关系，而这一切都离不开计算机技术的支持。中国科学院院士张春霆在 2000 年发表了“生物信息学的现状与展望”一文，下面予以引用，作为对生物信息学这门新兴学科的介绍。

1.2 生物信息学

1.2.1 生物信息学产生的背景

有人说，基于序列的生物学时代已经到来。尽管对“序列生物学”这一提法可能有所争议，但是今日像潮水般涌现的序列信息却是无可争辩的事实。自从 1990 年美国启动人类基因组计划以来，人与模式生物基因组的测序工作进展极为迅速。迄今，已完成了约 40 多种生物的全基因组测序工作，人类基因组约 3×10^9 碱基对的测序工作也接近完成。至 2000 年 6 月 26 日，被誉为生命“阿波罗计划”的人类基因组计划，经过美、英、日、法、德和中国科学家的艰苦努力，终于完成了工作草图，这是人类科学史上又一个里程碑式的事件。它预示着完成人类基因组计划已经指日可待。截至目前，仅登录在美国 GenBank 数据库中的 DNA 序列总量已超过 70 亿碱基对。在人类基因组计划进行过程中所积累起来的技术和经验，使得其他生物基因组的测序工作可以完

成得更快捷。可以预计，今后 DNA 序列数据的增长将更为惊人。生物学数据的积累并不仅仅表现在 DNA 序列方面，与其同步的还有蛋白质的一级结构，即氨基酸序列的增长。此外，迄今为止，已有 1 万多种蛋白质的空间结构以不同的分辨率被测定。基于 cDNA 序列测序所建立起来的 EST 数据库，其记录已达数百万条。在这些数据基础上派生、整理出来的数据库已达 500 余个，这一切构成了一个生物学数据的海洋。可以打一个比方来说明这些数据的规模，有人估计人类（包括已经去世的和仍然在世的）所说过的话的信息总量约为 5 艾字节（1 艾字节等于 10^{18} 字节），而如今生物学数据信息总量已接近甚至超过此数量级。这种科学数据的急速和海量积累，在人类科学研究历史中是空前的。

数据并不等于信息和知识，但却是信息和知识的源泉，关键在于如何从中挖掘它们。与正在以指数方式增长的生物学数据相比，人类相关知识的增长（粗略地用每年发表的生物、医学论文数来代表）却十分缓慢。一方面是巨量的数据，另一方面是人们在医学、药物、农业和环保等方面对新知识的渴求，这些新知识将帮助人们改善其生存环境和提高生活质量。这就构成了一个极大的矛盾，这个矛盾催生了一门新兴的交叉学科，这就是生物信息学。在美国人类基因组计划实施 5 年后的总结报告中，对生物信息学做了以下定义：生物信息学是一门交叉学科，它包含了生物信息的获取、处理、存储、分发、分析和解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具，阐明和理解大量数据所包含的生物学意义。生物信息学这一名词的出现仅仅是几年前的事情，但是计算生物学这一名词的出现要早得多。鉴于这两门学科之间并没有或难以界定严格的分界线，在这里统称为生物信息学。

1.2.2 生物学数据库

《Nucleic Acids Research》杂志连续7年在其每年的第1期中详细介绍最新版本的各种数据库。在2000年1月1日出版的第28卷第1期中详细地介绍了115种通用和专用数据库，包括详尽描述和访问网址。迄今为止，生物学数据库总数已达500个以上。在DNA序列方面有GenBank、EMBL和DDBJ等；在蛋白质一级结构方面有SWISS-PROT、PIR和MIPS等；在蛋白质和其他生物大分子的结构方面有PDB等；在蛋白质结构分类方面有SCOP和CATH等。应该指出，几乎所有这些数据库对学术研究部门或人员来说都是免费的，可以免费下载或提供免费服务。但是，鉴于相当多的数据库经营者们面临着资金紧缺的境地，这种免费的局面还能维持多久就不得而知了。有的数据库，如SWISS-PROT，已开始向商业用户每年收取数千至数万美元不等的使用费。其他数据库暂时还是免费的，但不知是否永远免费。如果一些重要的数据库对学术研究部门开始收费，这对于我国生物信息学的发展是非常不利的。中国是一个基因信息资源大国，应当抓紧建设我国自有的数据库，在世界上作出贡献，在平等的基础上与国外共享生物信息资源。

1.2.3 生物信息学的主要研究内容

生物信息学主要包括以下几个研究领域，但是限于篇幅，这里仅列出其名称并做简单介绍。

1. 序列比对（Alignment）

基本问题是比较两个或两个以上符号序列的相似性或不相似性。序列比对是生物信息学的基础，非常重要。两个序列的比对有较成熟的动态规划算法，以及在此基础上编写的比对软

件包——BALST 和 FASTA，可以免费下载使用。这些软件在数据库查询和搜索中有重要的应用。有时，两个序列总体并不很相似，但某些局部片断相似性很高。Smith-Waterman 算法是解决局部比对的好算法，缺点是速度较慢。两个以上序列的多重序列比对目前还缺乏快速而又十分有效的算法。

2. 结构比对

基本问题是比较两个或两个以上蛋白质分子空间结构的相似性或不相似性。目前已有一些算法。

3. 蛋白质结构预测

包括二级和三级结构预测，是最重要的课题之一。从方法上来看有演绎法和归纳法两种途径。前者主要是从一些基本原理或假设出发来预测和研究蛋白质的结构和折叠过程，分子力学和分子动力学属这一范畴。后者主要是从观察和总结已知结构的蛋白质结构规律出发来预测未知蛋白质的结构，同源模建和指认（Threading）方法属于这一范畴。虽然经过 30 余年的努力，但是蛋白质结构预测研究的现状却远远不能满足实际需要。

4. 计算机辅助基因识别（仅指蛋白质编码基因）

基本问题是给定基因组序列后，正确识别基因的范围和其在基因组序列中的精确位置。这是最重要的课题之一，而且越来越重要。经过 20 余年的努力，提出了数十种算法，有 10 种左右的重要算法和相应软件在网上提供免费服务。原核生物计算机辅助基因识别相对容易些，结果好一些。从具有较多内含子的真核生物基因组序列中正确识别出起始密码子、剪切位点和终止密码子，是个相当困难的问题，研究现状不能令人满意，仍有大量的工作要做。

5. 非编码区分析和 DNA 语言研究

这是最重要的课题之一。在人类基因组中，编码部分仅占总序列的 3%～5%，其他通常称为“垃圾”DNA，但其实只是人们暂时还不知道其功能。分析非编码区 DNA 序列需要大胆的想象与崭新的研究思路和方法。DNA 序列作为一种遗传语言，不仅体现在编码序列之中，而且还隐含在非编码序列之中。

6. 分子进化和比较基因组学

这也是最重要的课题之一。早期的工作主要是利用不同物种中同一种基因序列的异同来研究生物的进化，构建进化树。既可以用 DNA 序列也可以用其编码的氨基酸序列来做，甚至可通过相关蛋白质的结构比对来研究分子进化。以上研究已经积累了大量的工作。近年来，较多模式生物基因组测序任务的完成，为从整个基因组的角度研究分子进化提供了条件。可以设想，比较两个或多个完整基因组这一工作需要新的思路和方法，当然也渴望得到更丰硕的成果。这方面可做的工作是很多的。

7. 序列重叠群（Contigs）装配

一般来说，根据现行的测序技术，每次反应只能测出 500 或更多碱基对的序列，把大量的较短的序列全体构成了重叠群（Contigs），再逐步把它们拼接起来形成序列更长的重叠群，直至得到完整序列的过程称为重叠群装配。拼接 EST 数据以发现全长新基因也有一个类似的过程。已经证明，这是一个 NP-完备性算法问题。

8. 遗传密码的起源

遗传密码为什么是现在这样的？这一直是一个谜。一种最简单的理论认为，密码子与氨基酸之间的关系是由生物进化历史上一次偶然的事件而造成的，并被固定在现代生物的共同祖先里，

一直延续至今。不同于这种“冻结”理论，有人曾分别提出过选择优化、化学和历史等三种学说来解释遗传密码。随着各种生物基因组测序任务的完成，为研究遗传密码的起源和检验上述理论的真伪提供了新的素材。

9. 基于结构的药物设计

人类基因组计划的目的之一在于阐明人类约 10 万种蛋白质的结构、功能、相互作用及与各种人类疾病之间的关系，寻求各种治疗和预防方法，包括药物治疗。基于生物大分子结构的药物设计是生物信息学中极为重要的研究领域。为了抑制某些酶或蛋白质的活性，在已知其三级结构的基础上，可以利用分子对接算法，在计算机上设计抑制剂分子，作为候选药物。这种发现新药物的方法有强大的生命力，也有巨大的经济效益。

10. 其他

如基因表达谱分析、代谢网络分析、基因芯片设计和蛋白质组学数据分析等，逐渐成为生物信息学中新兴的重要研究领域，这里不再赘述。

1.2.4 与生物信息学关系密切的数学领域

限于篇幅，仅列出它们的名称。统计学，包括多元统计学，是生物信息学的数学基础之一；概率论与随机过程理论，如近年来兴起的隐马尔科夫链模型（HMM），在生物信息学中有重要应用；运筹学，如动态规划法是序列比对的基本工具，最优化理论与算法在蛋白质空间结构预测和分子对接研究中有重要应用，拓扑学，这里指几何拓扑，是 DNA 超螺旋研究中的重要工具，在多肽链折叠研究中也有应用；函数论，如傅里叶变换和小波变换等都是生物信息学中的常规工具；信息论，在分子进化、蛋白质