数学与现代科学技术丛书 2

生物计算

——生物序列的分析方法与应用

杨晶 胡刚 王奎 沈世镒 编著

科学出版社
www.sciencep.com

数学与现代科学技术丛书 2

生物计算

——生物序列的分析方法与应用

杨 晶 胡 刚 王 奎 沈世镒 编著

科学出版社

北京

内 容 简 介

本书介绍生物计算中的几种主要方法，如序列比对、系统发育分析、蛋白质序列的语义分析与结构预测、基因识别与生物芯片的数据分析等，给出它们的基本问题与有关的方法及应用. 全书由三部分组成. 第一部分介绍这些问题的由来与主要内容，给出它们的基本原理、计算与分析方法及应用意义，同时介绍一些国际上较为通用的软件包. 第二部分是生物学备忘录，介绍有关生物学的基础知识. 第三部分是数学备忘录，介绍与这些生物计算有关的数学理论与方法.

本书可作为数学、生物、医学、化学等专业的本科生或研究生教材，其中第一部分内容可作为各专业的公共部分，而第二、三部分内容可供各专业适当选用.

图书在版编目(CIP)数据

生物计算：生物序列的分析方法与应用/杨晶等编著. —北京：科学出版社，2010

(数学与现代科学技术丛书；2)

ISBN 978-7-03-026393-3

Ⅰ.生… Ⅱ.杨… Ⅲ.生物信息论-研究 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字 (2010) 第 007551 号

责任编辑：赵彦超／责任校对：鲁 素

责任印制：钱玉芬／封面设计：王 浩

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

*

2010 年 3 月第 一 版 开本：B5(720×1000)

2010 年 3 月第一次印刷 印张：23 3/4

印数：1—3 000 字数：456 000

定价：69.00 元

(如有印装质量问题，我社负责调换〈双青〉)

《数学与现代科学技术丛书》序

当代数学在向纵深发展的同时，被空前广泛地应用于几乎一切领域. 一方面，它与其他学科交汇，形成了许许多多交叉学科（例如，信息科学、计算机科学、系统科学、数学物理、数学化学、生物数学、数学语言学、数量经济学、金融数学、复杂性科学、科学计算等）；另一方面，它又被应用于高新技术（如信息安全、信息传输、图像处理、语音识别、网络、海量数据处理、网页搜索、遥测遥感、交通管理、医疗诊断、手术方案、药物检验、商业广告等方面）的开发，成为一些高新技术的核心. 应用数学的这种发展趋势急剧地扩展了数学的疆界，也深刻地改变了数学的面貌.

中国的经济正在迅猛发展，其中的科技含量也与日俱增. 为了提高自主创新能力，我国已经有不少数学工作者投身于这类应用数学的研究中，还有更多的数学工作者则正在密切关注这方面的进展，看好它的前景. 愈来愈多的人希望了解这类应用数学的现状，寻找入门之径.

《数学与现代科学技术丛书》是力图反映这个发展趋势的一套应用数学丛书，它将较全面地向我国读者介绍当今数学在现代科学技术各个领域中应用的状况，通过必要的准备知识，逐步把读者引向相关的研究前沿.

从事交叉学科研究和高新技术开发的应用数学家，除了要精通所需的数学知识外，还必须深入了解其所研究问题的来龙去脉. “建模”是应用数学研究实际问题的关键. 这也是一门数学艺术：从复杂的实际问题中抽象出关键的“量的关系”，它既能反映出问题的基本特征，又能用现阶段的数学工具对其加以处理. 有鉴于此，这套丛书的一个特点就是不但要介绍有关的数学理论和方法，还必须介绍问题的来源与背景、数学建模以及如何运用数学工具来解决实际问题.

本丛书适用于数学及相关专业的大学生和研究生，以及与数学有关的各专业科技工作者.

张恭庆

2009 年 11 月

前　言

生物计算中的理论、方法与应用越来越被生物、医学及其他医务工作者所需要与关注, 特别是在人类基因组计划实施以来, 该学科的发展与研究更凸显出重要的作用. 基因、基因组、蛋白质、蛋白质组等生物学信息的数据采集、储存与分析及其生物学意义, 是生物计算乃至生物、医学与医药的重点研究内容之一. 因此在国内外的许多医科院校均被作为重要课程, 与生物信息学和生物计算相关内容的课程不仅是研究生的必修课程或选修课程, 也是多个专业本科生的专业必修课程或选修课程. 我们先后用了近三年的时间, 在开展教学和研究工作的同时编写了本书, 目的是为生物学和医学相关专业的本科生与研究生提供一本既通俗易懂, 同时又可深入了解相关内容的教材, 为该学科的建设与发展服务.

自 2004 年以来, 本人有幸多次参加南开大学数学科学学院沈世镒教授主持的"生物信息学" 讨论班. 在讨论与学习过程中, 不仅掌握了一些解决生物序列分析与计算的具体算法, 更重要的是学到了解决生物序列分析的一些新方法和新思想. 如生物序列的多种比对算法、数据结构中的语义分析及其在蛋白质结构分析中的应用等. 这些方法从不同角度对生物计算中的有关问题进行研究与探讨, 并在许多方面得到了很好的应用. 在学习过程中, 与南开大学数学科学学院胡刚、王奎博士等合作, 对生物计算中的算法以及相关软件包的使用等问题有了更深入与确切的理解, 使本书得以顺利完成. 我们希望能将该领域中的主要内容与方法介绍给读者.

"生物计算" 与 "生物信息学" 在本质上无大的区别, 国内外的许多院校均把它们看作同一领域的学科. 在本书中, 我们把 "生物计算" 看作较偏重于原理与方法, 同时注重它们的实现与应用, 在介绍国外先进与常用算法的同时, 增加了相应软件包的使用与分析等内容.

本书共分为三个部分, 第一部分是本书的重点和核心, 主要介绍生物信息学中的常用算法与软件, 包括它们的基本原理、方法与应用, 并介绍国际上常用的计算机软件包, 对其性能、使用方法与具体实施加以介绍, 以方便读者更快和更有效地实现与应用这些算法. 该部分内容尽量简练并具有实用性与可操作性. 为便于更多的读者阅读此书, 本书第二与第三部分分别介绍了有关的生物学和数学的基础知识. 第二部分介绍了与生物计算相关的生物学基础知识, 是为具有数学与计算机等专业背景的读者准备的, 已经具有生物学与医学等专业背景的读者可以将这部分内容略去不读. 第三部分介绍与生物计算相关的数学理论与方法, 因为它涉及多个数学学科领域, 内容较为广泛, 对理解与掌握生物计算的基本方法是有益的, 也可供

需要作更深入研究 (如自行设计或编写算法程序与软件等) 的读者参考.

本书的编写是在沈世镒教授的建议、组织与策划下由多人编撰完成, 第 5~10 章由杨晶编撰, 第 2~4, 13 章由胡刚编撰, 第 1, 11, 12, 14 章由王奎编撰, 沈世镒审校全书. 书中介绍的一些新算法, 如序列比对的 SPA 算法与蛋白质一级结构的语义分析等, 是南开大学生物信息学研究组取得的新成果, 书中未详细介绍的部分结果可在参考文献 [174, 228] 中找到. 另外, 特别感谢天津医科大学生物医学工程系田心教授在科研和教学工作中给予的精心指导和帮助. 同时感谢天津医科大学内分泌研究所郭刚教授、张镜宇教授和基础医学院解用虹教授等在分子生物学方面的帮助和指导. 感谢天津理工大学计算机学院宋金杰教授及其学生周新博的帮助.

本书的出版获国家自然科学基金项目 (20836005; 30870791)、天津市自然科学基金项目 (07JCDJC08100) 以及天津医科大学和南开大学教务处的支持, 特此表示感谢.

本书可作为医学、生物学、计算机和数学等专业本科生或研究生“生物信息学”或“生物计算学”课程的教材或参考书. 不同专业对本书内容可有不同的侧重与选择. 生物信息学的研究还处在起步阶段, 一些理论与方法都在不断发展中, 因此有关内容还会不断更新. 另外, 本书内容涉及医学、生物学、数学与计算机等专业, 具有跨学科特征. 由于我们的专业知识与工作背景的限制, 书中错误或不妥之处在所难免, 真诚希望读者批评指正.

杨　晶

2009 年 3 月于天津

目　录

第二部分　生物学备忘录

第三部分　数学备忘录

第一部分

基 本 方 法

第 1 章　生物序列突变与比对分析

1.1　生物序列突变与比对问题

1.1.1　生物序列的类型与结构

1. 生物序列的定义与记号

生物序列一般指 DNA 序列、RNA 序列或蛋白质序列, 它们是具有活性的生物大分子, 在一定的环境条件下具有新陈代谢与特定的生物学功能, 都是由大量生物小分子按一定的顺序排列而成. 其中 DNA 与 RNA 序列由不同的核苷酸排列组成, 而蛋白质是由不同的氨基酸排列而成. 如果把核苷酸或氨基酸看作一个基本单元, 那么生物序列就是由这些基本的分子单元的排列组成.

生物序列的结构有多种表示方式, 最常见的是它们的一级结构与空间结构表达. 所谓一级结构就是由这些基本分子单元按一定的顺序排列而成的序列, 因此可用以下序列记号表示:

$$A=(a_1,a_2,\cdots,a_n),\quad B=(b_1,b_2,\cdots,b_m), \tag{1.1.1}$$

其中 A,B 等英文大写字母表示序列, a_i,b_i 表示在第 i 位置上的基本分子单元, 它们取自某一个特定的分子集合 $V_q=\{0,1,\cdots,q-1\}$, 当 A,B 是 DNA 或 RNA 序列时, 取 $q=4$, 这时 $V_4=\{a,c,g,t\}$ (或 $\{a,c,g,u\}$) 表示四种不同的核苷酸. 当 A,B 是蛋白质序列时, V_q 就是氨基酸的集合, 常用的氨基酸有 20 种, 这时 $q=20$, V_{20} 就表示这 20 种常用的氨基酸.

在 (1.1.1) 式中, 下标 n,m 分别是 A,B 的序列长度. 对于一般的多重序列 (序列组), 记之为

$$\mathcal{A}=\{A_1,A_2,\cdots,A_s,\cdots,A_m\}, \tag{1.1.2}$$

其中每个 A_s 是 V_q 上的序列, 记之为

$$A_s=(a_{s,1},a_{s,2},\cdots,a_{s,n_s}),\quad s=1,2,\cdots,m, \tag{1.1.3}$$

而 n_s 是序列 A_s 的长度, 称 m 为该多重序列的重数.

2. 生物序列的类型

生物序列有多种类型, 从分子结构类型来分, 有核苷酸 (DNA 与 RNA) 序列与氨基酸序列. 从数据内容来分, 有一级结构与空间结构等, 一级结构一般指生物大