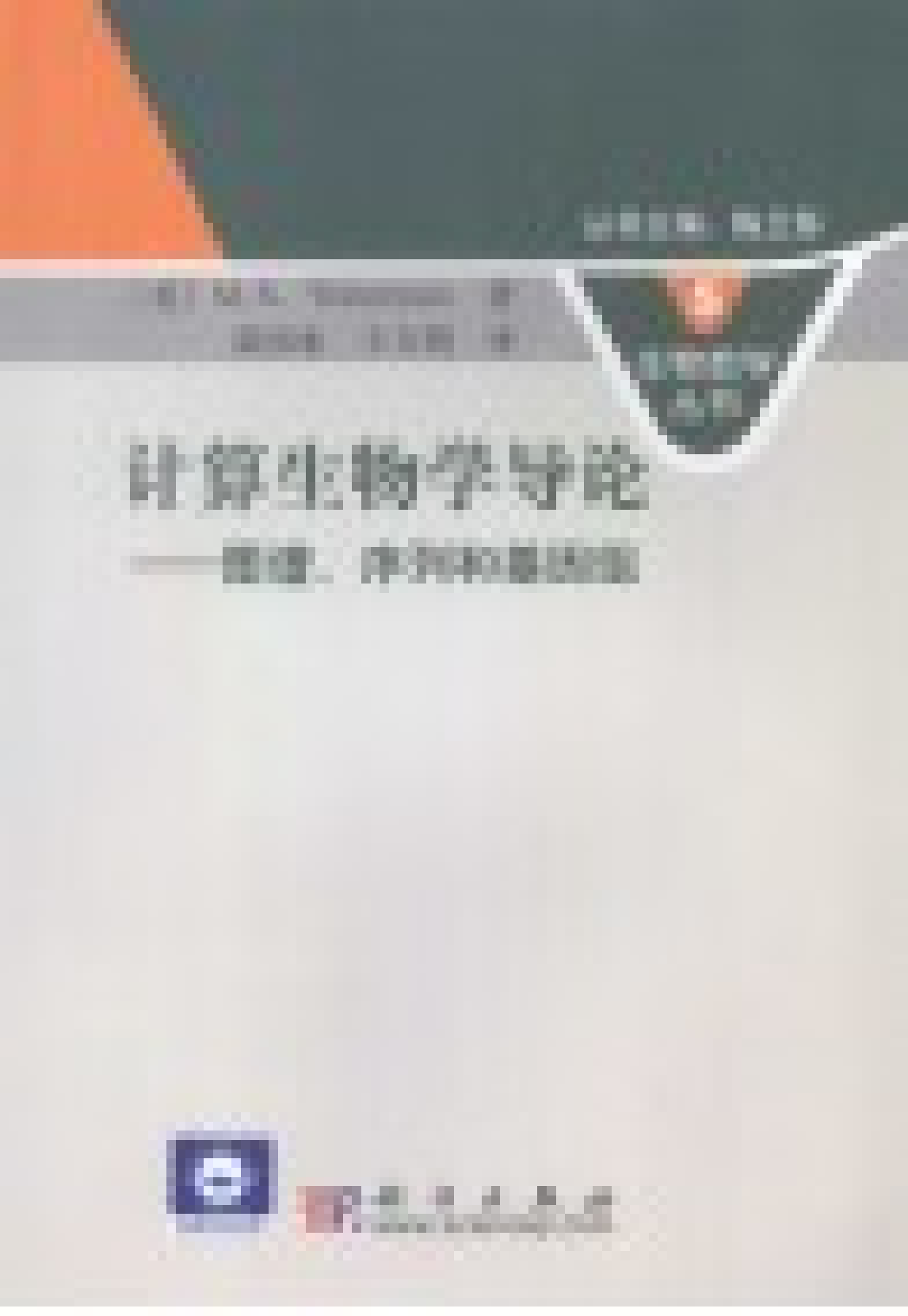
丛书主编：陈兰荪

〔美〕M.S. Waterman 著
黄国泰 王天明 译

5
生物数学丛书

计算生物学导论
——图谱、序列和基因组

科学出版社
www.sciencep.com



生物数学丛书 5

计算生物学导论

——图谱、序列和基因组

〔美〕M. S. Waterman 著
黄国泰 王天明 译

科学出版社
北京

图字：01-2009-2608

内 容 简 介

本书是 *Introduction to Computational Biology* 的中文译著，本书的意图是针对有数学技能的人介绍令人着迷的生物数据和问题，并建立更实际的生物数学的基础。

本书共分 15 章，其中第 1 章介绍分子生物学的基本常识，第 2—4 章介绍限制图谱和多重图谱，第 5、6 章研究克隆和克隆图谱，第 7 章讨论 DNA 序列相关的话题，第 8—11 章是共同模式下序列比较问题，第 12 章涉及序列中模式计数的统计问题，第 13 章叙述 RNA 二级结构的数学化论述，第 14 章给出有关序列的进化历史，最后第 15 章给出某些关键文献的原始出处. 本书结构完整，内容更新、更全面.

本书适合高等院校数学和生物专业的高年级大学生、研究生和教师阅读参考，也适合科研单位的研究人员参考.

图书在版编目(CIP)数据

计算生物学导论：图谱、序列和基因组/（美）M. S. Waterman 著，黄国泰，王天明译. —北京：科学出版社，2009
(生物数学丛书；5)
ISBN 978-7-03-025156-5
Ⅰ. 计… Ⅱ. ①M… ②黄… ③王… Ⅲ. 分子生物学-计算方法 Ⅳ. Q7
中国版本图书馆 CIP 数据核字(2009) 第 134405 号

责任编辑：陈玉琢 房 阳／责任校对：钟 洋
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

漷滋印刷有限责任公司印刷

科学出版社发行 各地新华书店经销
*
2009 年 8 月第 一 版 开本：B5(720×1000)
2009 年 8 月第一次印刷 印张：23 1/4
印数：1—3 000 字数：449 000

定价：68.00 元
(如有印装质量问题，我社负责调换〈环伟〉)

《生物数学丛书》序

传统的概念：数学、物理、化学、生物学, 人们都认定是独立的学科, 然而在 20 世纪后半叶开始, 这些学科间的相互渗透、许多边缘性学科的产生, 各学科之间的分界已渐渐变得模糊了, 学科的交叉更有利于各学科的发展, 正是在这个时候数学与计算机科学逐渐地形成生物现象建模, 模式识别, 特别是在分析人类基因组项目等这类拥有大量数据的研究中, 数学与计算机科学成为必不可少的工具. 到今天, 生命科学领域中的每一项重要进展, 几乎都离不开严密的数学方法和计算机的利用, 数学对生命科学的渗透使生物系统的刻画越来越精细, 生物系统的数学建模正在演变成生物实验中必不可少的组成部分.

生物数学是生命科学与数学之间的边缘学科, 早在 1974 年就被联合国科教文组织的学科分类目录中作为与 "生物化学"、"生物物理" 等并列的一级学科."生物数学" 是应用数学理论与计算机技术研究生命科学中数量性质、空间结构形式, 分析复杂的生物系统的内在特性, 揭示在大量生物实验数据中所隐含的生物信息. 在众多的生命科学领域, 从 "系统生态学"、"种群生物学"、"分子生物学" 到 "人类基因组与蛋白质组即系统生物学" 的研究中, 生物数学正在发挥巨大的作用, 2004 年《Science》杂志在线出了一期特辑, 刊登了题为 "科学下一个浪潮 —— 生物数学" 的特辑, 其中英国皇家学会院士 Lan Stewart 教授预测, 21 世纪最令人兴奋、最有进展的科学领域之一必将是 "生物数学".

回顾 "生物数学" 我们知道已有近百年的历史：从 1798 年 Malthus 人口增长模型, 1908 年遗传学的 Hardy-Weinberg"平衡原理"; 1925 年 Voltera 捕食模型, 1927 年 Kermack-Mckendrick 传染病模型到今天令人注目的 "生物信息论", "生物数学" 经历了百年迅速地发展, 特别是 20 世纪后半叶, 从那时期连续出版的杂志和书籍就足以反映出这个兴旺景象; 1973 年左右, 国际上许多著名的生物数学杂志相继创刊, 其中包括 Math Biosci, J. Math Biol 和 Bull Math Biol; 1974 年左右, 由 Springer-Verlag 出版社开始出版两套生物数学丛书：Lecture Notes in Biomathermatics (二十多年共出书 100 册) 和 Biomathematics (共出书 20 册); 新加坡世界科学出版社正在出版 "Book Series in Mathematical Biology and Medicine" 丛书.

"丛书" 的出版, 既反映了当时 "生物数学" 发展的兴旺, 又促进了 "生物数学" 的发展, 加强了同行间的交流, 加强了数学家与生物学家的交流, 加强了生物数学学科内部不同分支间的交流, 方便了对年轻工作者的培养.

从 20 世纪 80 年代初开始, 国内对 "生物数学" 产生兴趣的人越来越多, 他 (她)

们有来自数学、生物学、医学、农学等多方面的科研工作者和高校教师, 并且从这时开始, 关于 "生物数学" 的硕士生、博士生不断培养出来, 从事这方面研究、学习的人数之多已居世界之首. 为了加强交流, 为了提高我国生物数学的研究水平, 我们十分需要有计划、有目的地出版一套 "生物数学丛书", 其内容应该包括专著、教材、科普以及译丛, 例如: ① 生物数学、生物统计教材; ② 数学在生物学中的应用方法; ③ 生物建模; ④ 生物数学的研究生教材; ⑤ 生态学中数学模型的研究与使用等.

中国数学会生物数学学会与科学出版社经过很长时间的商讨, 促成了 "生物数学丛书" 的问世, 同时也希望得到各界的支持, 出好这套丛书, 为发展 "生物数学" 研究, 为培养人才作出贡献.

陈兰荪

2008 年 2 月

前　言

仅仅在 1953 年才确定了著名的 DNA 双螺旋结构. 自从那时起, 出现了一系列惊人的发现. 阐明遗传密码仅仅是开始. 了解基因和它们在真核生物, 如人类基因组中不连续性的细节, 已经导致能够研究和操作 Mendel 的抽象概念 —— 基因本身. 学会越来越快地阅读遗传材料使我们能够试图解读整个基因组. 像我们正在接近 21 世纪一样, 我们也正在接近生物学不可思议的新纪元.

分子生物学的革新率惊心动魄. 一代人为写博士论文必须煞费苦心掌握的实验技术, 对现代大学生来说成为例行实验. 数据的积累已经使建立国际核酸、蛋白质、单个生物体, 甚至染色体的数据库成为必要. 粗略地度量核酸数据库的大小进展过程成指数增长, 从而新的学科 (如果这样说太自大了): 生物学和信息科学结合的新的专门领域正在不断产生. 在巨大的数据库中寻找相关事实和假设, 对生物学来说变得非常重要. 这本书是关于生物学数据库, 特别是关于序列和染色体的数学结构的.

数学书名趋向于简洁、隐匿的观点, 而生物学的书名通常比较长, 包含的信息更多, 相当于数学家给出的简单摘要. 相应地, 生物学家的摘要有数学家引言的长度和细节. 为了努力填补到目前为止几乎孤立的两种文化之间的鸿沟, 我的书名反映了这些冲突的传统. "计算生物学导论" 是一个短书名, 可以用作许多不同书的名字. 书名的副标题 "图谱, 序列和基因组" 是让读者知道这本书是关于分子生物学应用的. 即使这样也太短, "计算生物学导论 ·····" 应该为 "计算, 统计和数学分子生物学 ·····".

在第 1 章详细说过, 打算读本书的读者应该学过概率和统计的基本课程, 也应该掌握微积分. 计算机科学中的算法和复杂性的概念也是有帮助的. 至于生物学, 大学入门课程也非常有用, 是每个受教育的人在任何场合都应该知道的材料. 本书打算给具有数学技能的人介绍令人着迷的生物数据和问题, 而不是给那些喜欢自己学科纯洁又封闭的人. 在如此迅速发展的学科中所做工作有立即变废的重大危险. 我已经试图在我认为不大会改变的基础上和那些会被明天更巧妙的技术淘汰的数据结构和问题之间建立一个平衡. 例如, 物理图谱 (如限制图谱) 的基本性质依旧重要. 虽然 20 年来一直关心双消化问题, 它有变成过时的可能. 序列装配也容易受到技术的影响而发生许多改变. 序列比较总是有意义的, 并且动态规划算法是一个好的简单的框架, 这些问题都可以嵌入其中, 如此等等. 我试图介绍生物学引起的数学, 但不完全, 而且省略了一些重要的课题. 构造进化树值得写一本书, 到现在还

没有写. 蛋白结构是一个巨大的课题通常与数学无关, 这里没有涉及. 我试图做的是给出与基因组研究有关的一些有趣的数学.

对恰当确定与本书有关的研究领域的课题给予了很多关注. 甚至, 书的名字还没处理好. 数学生物学看起来并不满意, 一部分是由于更早时期的不幸, 并且这种选择相对计算生物学和信息学更窄. (如果后半部分名字成功, 我希望它用法语发音.) 更重要的是这个学科由哪些部分组成? 有三种主要的见解: ① 它是生物学适当的子集和能满足其需要的数学和计算机科学; ② 它是数学科学的子集, 生物学是遥远的动机所在; ③ 有许多真正的交叉学科成分, 具有生物学的原始动机的数学问题, 而这些问题的解又给生物学实验以提示, 如此等等. 我个人的观点是, 虽然最后一种是最值得鼓励的行动, 但所有这三种不仅是值得做而且是不可避免的和适当的做法. 在建立和阐述数学知识时, 我希望本书能帮助建立更实际的生物学中交叉学科的基础.

应该感谢的人很多, 鉴于篇幅有限, 敬请包涵. 在此, 预先对一些重要的疏漏 —— 他们被疏忽了表示歉意. Los Alamos 的 Stan Ulam 和 Bill Beyer 是使我进入这个领域的关键人物. Stan 相信在新的生物学中有数学, 没有给出任何细节, 可是以他的风格影响了许多人. 开始时, 我一点都不懂生物学, Temple Smith 给我有数学和统计内容的很好的问题, 并和我一起解决它们. 当别人还不清楚这个领域的实质时, Gian-Carlo Rota 鼓励我做这项工作. 在这项工作中, 他和后来的 Charlie Smith(而后是系统发展基金) 给予我重要的支持. 没有南加利福尼亚大学同事们的帮助, 本书会短得多、乏味得多, 他们是 Richard Arratia, Norman Arnheim, Caleb Finch, David Galas, Larry Goldstein, Louis Gordon 和 Simon Tavaré. 这些年来博士后 Gary Benson, Cary Churchill, Ramana Idury, Rob Jones, Pavel Pevzner, Betty Tang, Martin Vingron, Tandy Warnow 和 Momiao Xiong 非常友好地教我他们知道的东西, 使之成为更丰富的学科. 我的学生 Daniela Martin, Ethan Port 和 Fengzhu Sun 阅读了草稿, 做了习题, 并普遍地改进和改正了本书. 三个刻苦工作的天才将我的手稿相继翻译成 LaTex, 他们是 Jana Joyce, Nicolas Rouquette 和 Kengee Lewis. 我的工作得到了系统发展基金、国家健康研究所和国家自然基金的资助. 最后我要对 Walter Fitch, Hugo Martinez 表示感谢, 特别要对这个学科的先驱 David Sankoff 表示感谢, 从这个学科一开始他就参与并一直到现在. 在本书结束时, 希望读者将错误告诉我. Donald Knuth 在他精彩的多卷著作《程序设计的艺术》中为他的读者发现的每一处错误奖励一美元, 后来奖励两美元. 为做到最大限度消灭错误, 我也想提供类似的奖励, 但我怀疑能否付得起与仍存错误数成正比的总数. 取而代之, 仅能提供我最真诚的感谢. 我将做一个软件、勘误表和其他有关本书的信息由 ftp 或 http://hto-e.usc.edu 提供给大家.

数 学 符 号

函数

$\lfloor x \rfloor$	最大整数 $\leqslant x$
$\lceil x \rceil$	最小整数 $\geqslant x$
$x \wedge y$	x 与 y 的最小值
$x \vee y$	x 与 y 的最大值
x^+	$x \vee 0$
$a_n \sim b_n$	$\lim\limits_{n\to\infty} \dfrac{a_n}{b_n} = 1$
$f(x) \approx g(x)$	$f(x)$ 约等于 $g(x)$
O	若存在一个常数 c, 当 $x \to \infty$ 时, 使 $\lvert f(x) \rvert \leqslant cx^3$, $f(x)$ 为 $O(x^3)$
o	若当 $x \to \infty$ 时, $f(x)/x^3 \to 0$, 则当 $x \to \infty$ 时, $f(x)$ 为 $o(x^3)$
$\boldsymbol{A}^{\mathrm{T}}$	矩阵 $\boldsymbol{A}$ 的转置

实数的子集

$\mathbb{N}$	自然数: $1, 2, \cdots$
$\mathbb{Z}$	整数
$\mathbb{R}$	实数

集合符号

$\varnothing$	空集
$A \cup B$	A 和 B 的并集
$A \cap B$	A 和 B 的交集
A^{c}	A 的补集
$A \sim B$	$A \cap B^{\mathrm{c}}$
$\mathrm{limsup} A_n$	$\bigcap\limits_{n\geqslant 1} \Big(\bigcup\limits_{m\geqslant n} A_m \Big)$
$\mathrm{liminf} A_n$	$\bigcup\limits_{n\geqslant 1} \Big(\bigcap\limits_{m\geqslant n} A_m \Big)$
$\lvert A \rvert$ 或 $\#A$	A 的元素数目
$\boldsymbol{I}_A, \boldsymbol{I}(A)$	A 的示性函数

概率

$\boldsymbol{P}(A)$	A 的概率
$\boldsymbol{E}(X)$	随机变量 X 的期望
$\mathrm{Var}(X)$	X 的方差

$\mathrm{cov}(X,Y)$	(X,Y) 的协方差
$\mathrm{cor}(X,Y)$	(X,Y) 的相关系数
$X^d=Y$	X 和 Y 同分布
$X_n \xrightarrow{d} Y$	X_n 依概率收敛到 Y
iid	独立同分布
$B(n,p)$	参数为 n 和 p 的二项分布
$P(\lambda)$	均值为 λ 的 Poisson 分布
$N(\mu,\sigma^2)$	均值为 μ 方差为 σ^2 的正态分布

目　录

第0章 引 言

序言中简略地提到了给人留下深刻印象的一些分子生物学取得的进展. 分子生物学是实验学科, 虽然构成生物体的材料服从熟知的化学和物理规律, 但在生物学中没有几个真正的普遍规律, 即使描述核苷酸 (DNA 字符) 三联子到氨基酸 (蛋白的字符) 映身, 即所谓普遍遗传密码, 在所有的生命系统中也并非完全一样. 我曾听到一个学数学的同事嘟囔: "他们为什么不把它叫做几乎普遍的规律?" 问题是进化已经发现不同的问题有不同的解, 或者在相关的不同物种中进化对其结构进行了不同的修正. 生物学家经常寻找普遍规律. 可是, 不管发现了什么规律, 总存在各种变形. 为了严密起见, 生物学家总是仔细描述生物体和实验条件. 用类似的方式, 数学家仔细地叙述能用来证明定理的假设. 尽管数学家和生物学家有着使工作有效的共同愿望, 但像数学家和物理学家一样, 他们并不经常交往.

20 世纪初, 由 Fisher, Haldane, Wringt 和其他人精心提出的数学模型处于生物学前沿. 今天, 分子生物学的各种发现已经使数学科学远远落后. 然而, 生物序列数据库和分析这些数据的压力, 使这些邻域之间的联系正在加强. 生物学处在新纪元的开端, 有希望出现有意义的发现, 而它们要用装配信息数据库来刻画.

在第 1 章分子生物学简单叙述之后, 第 2~4 章将研究 DNA 的限制图谱、更详细的根本序列的初略标志图谱. 然后, 在第 5, 6 章研究克隆和克隆图谱, 形成基因组生物学文库和构造基因组的 "拼接图" 或基因组图谱是非常重要的. 第 7 章给出与阅读 DNA 序列本身有关的某些问题. 第 8~11 章介绍为寻找共同模式进行序列比较的一些问题. 两个或更多序列比较是数学在生物学中最重要的应用之一, 这些生物学问题引起算法和概率统计的一些进展. 在第 12 章涉及序列中模式计数的统计, 它令人惊讶得精细, 生物学中的分子结构是一个中心问题. 蛋白质结构是一个巨大的、基本上没有解决的问题, 本书不涉及它. RNA 二级结构的更数学化论述的课题在第 13 章处理. 最后, 给定一族有关序列, 我们试图推断出它们的进化历史, 这个问题在第 14 章讨论. 经典的遗传学、遗传图谱和聚结是值得充分地阐述, 而本书没有涉及.

在研究论文中, 正确的理解历史是本质的, 在一本引论中不能介绍大量的文献, 也不能完全忽略原始材料的参考文献, 在第 15 章中给出某些关键文献的原始出处. 此外, 也给出少数几篇位于当今研究前沿的文章. 这个学科发展得非常迅速, 读者不应该认为我已经介绍了任何给定课题的最新工作, 这里仅仅提供了入门材料, 然后, 需要去查文献, 查数据库和到实验室. 如图 0.1 所示, 这本书有许多独立的模型,