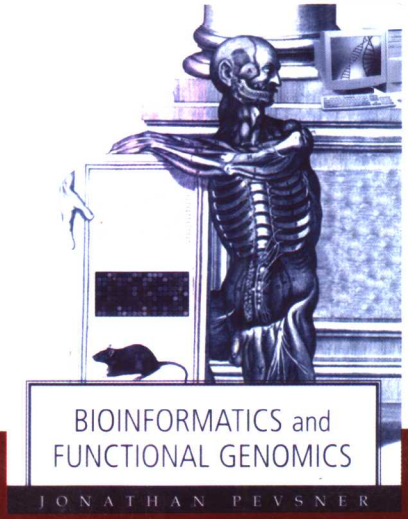
国外优秀科技著作出版专项基金资助

生物信息学与功能基因组学

BIOINFORMATICS and FUNCTIONAL GENOMICS

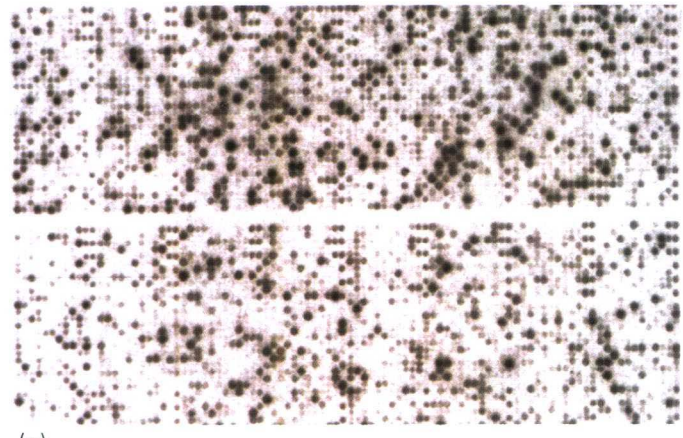
[美] 乔纳森·佩夫斯纳 著
(JONATHAN PEVSNER)

孙之荣 主译

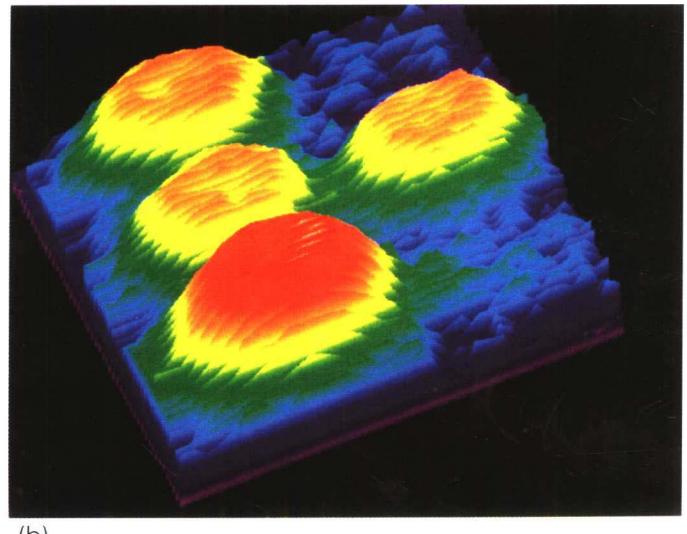


Chemical Industry Press

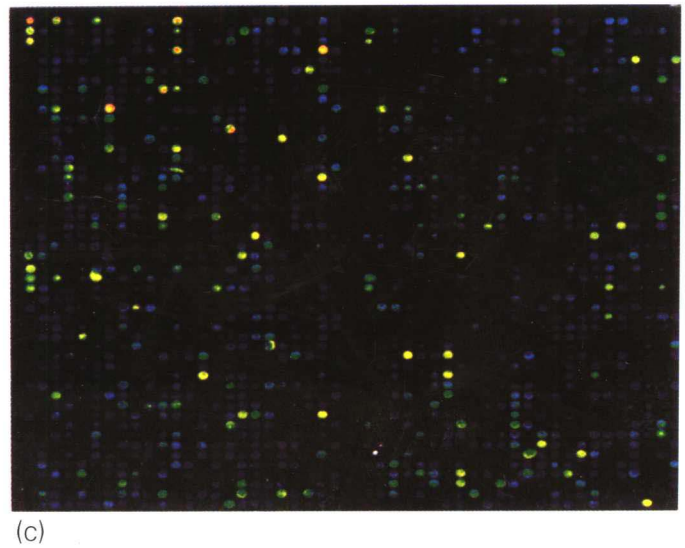
化学工业出版社
现代生物技术与医药科技出版中心



(a)



(b)



(c)

图 6.20 Research Genetics公司的GeneFilter微阵列（探针为从唐氏综合征患者死后脑部海马区得到的cDNA）

（a）微阵列上共有5000种cDNA。基因在所有阵列上的分布必须随机化；

（b）用NIH图像软件看到的6个信号点。图像分析软件必须详细说明每个信号的性质，包括一个强信号（图中左下角）“溢出”到弱信号（右下角）的可能性；

（c）NEN Perkin-Elmer公司的一种微阵列（MICROMAX，有2400个基因），用和图6.19中同一个瑞特综合征患者和相应对照组的脑组织样本做探针。这里使用的是经荧光标记和竞争性杂交的样本

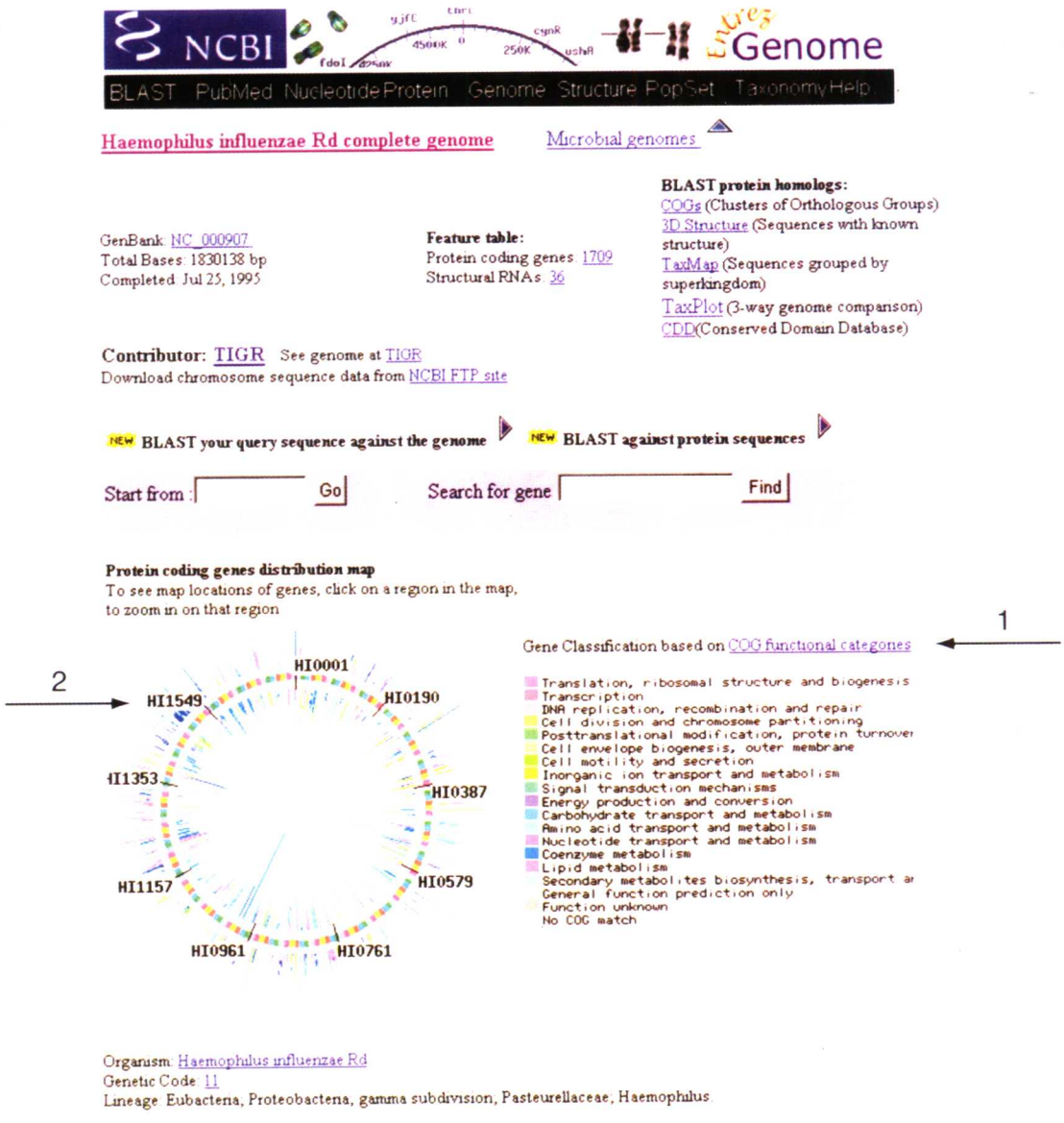


图 12.9 第一个完成基因组测序工作的独立生活的生物体 *H. influenzae* Rd的Entrez基因组记录。这个记录是从Entrez基因组资源中点击其左侧栏的“bacteria（细菌）”得到的。记录的顶端包含了基因组的编号以及大小等信息。整个核苷酸序列都可以从这里下载。右上方是关于这个基因组所编码的1709个蛋白质的一些资源。从这里可以进行 *H. influenzae* 的BLAST搜索。记录的主要部分是由一个代表这个基因组的彩色圆环和一个基于COG（见下文）的功能分类组成的。环形图表被点击以后（图12.10）将显示这个基因组的基因和蛋白质的详细信息。底部是这个细菌的基因代码以及分类学起源。这个记录还包含了参考文献（图中没有显示出来），包括基因组研究所Fleischmann等在1995年的最初的基因组序列报告

译者序

后基因组时代的生物学研究对每一个生物学工作者来说都充满了挑战。随着人类基因组测序工作的初步完成，越来越显现出生物信息学的重要作用。目前，虽然生物信息学已经在国内外得到了蓬勃的发展，但是国内相关书籍的水平还十分参差不齐，至今仍没有一本非常翔实的中文教材。与此同时，很多生物学研究者特别是生物信息学工作者都迫切需要一本能够详尽介绍生物信息学概要的中文教材类用书，而《生物信息学与功能基因组学》的出现恰可部分满足这种需求。此书内容翔实、语言通俗、图文并茂，对目前生物信息学的各个主要研究方向讲解详细而清楚，引用了国外生物信息学领域专家最新的研究成果。特别是第 3 篇对各个模式生物基因组序列所作的单独讲解很有特色，无论是对于生物信息研究者还是生物学研究者都将会是十分实用的。书后的生物信息学专业名词术语表设计，便于读者在阅读和平时学习及工作中作简单查阅。我们将此书整理翻译出版并介绍给广大的生物学研究工作者，希望能对他们的研究工作有所帮助，同时此书对从事生物信息学研究和学习的大学本科生和研究生也是一本很好的参考教材。

《生物信息学与功能基因组学》涵盖了现代生物学的许多新的研究领域，对功能基因组有比较详细的介绍，提供了一些重要的生物学资源。本书将生物信息学与功能基因组学结合在一起讲述，更清楚地描述了功能基因组时代生物信息学的新发展。全书分为 3 篇，第 1 篇着重分析了当前 web 数据库中的 DNA、RNA 和蛋白质序列资源；第 2 篇着重在基因组层次上对蛋白质和 RNA 进行了分析；第 3 篇着重对多个具体的已经测序的基因组完整序列进行了详尽的分析；书后的附录部分还附带了生物信息学专业名词术语表。

本书的翻译工作由清华大学生物科学与技术系生物信息学与系统生物学研究所的师生集体完成。参加本书翻译校对工作的有李立博士、郑家顺博士、夏学峰、沈金城、孟祥波、王世雷、周云、袁麟、李薇、黄泥、张松、陈虎、吴明、黄波、陈香、梁灵等。全书由孙之荣教授负责组织和审校。希望本书的翻译和出版将对后基因组时代我国生物学研究的发展及对生物信息学研究人才的培养能够提供良好的帮助和促进作用。

孙之荣

2006 年 1 月于清华园

原序

近10年来生物信息学的成长基于一个简单的原因：数据。逐渐积累的大量实验数据驱使我们开发数据的表示方法和算法来管理这些信息。生物学正在发生一场革命，如果计算的革命不同时伴随着生物学的革命，那么计算机能为生物学服务的内容就比较有限，如同20世纪80年代中期那样：一些科学家（如结构生物学家）利用计算机来进行模型优化、结构模拟和可视化，而另一些热心的科学家会醉心于管理专门的“精品店”式的数据库，大多数生物学家则仅会偶尔地使用计算机进行文献搜索、结果的图形化、定性模型的描述，以及少量的序列搜索和比较。

但是，生物学发生了革命。实验测定生物分子（蛋白质和核酸）序列能力的增长超出许多人的预期。在20世纪90年代早期我们已经清楚我们需要改进方法来储存、分析和传播这些生物学信息。科学团体启动人类基因组计划（和相关的测序计划）的决定使得那些懂得如何分析数据和拥有发明新方法的能力的科学家们大有用武之地。其结果是，来自于数学、统计学和计算机科学的研究者们导致了第一次突破，这些发现可以用于帮助拼接基因组序列、创建序列数据库以及开发序列分析的快速搜索和比较算法。

在早期，从事生物信息学的人员来自于各个领域。研究某些特定问题的生物学家进入了计算领域：X射线结晶学、NMR波谱学、系统发生学和群体（或统计）遗传学。同时，三维结构计算也吸引了一些计算机科学家的注意，他们的兴趣在于计算机图形学和计算几何。对字符串分析算法感兴趣的计算机科学家发现序列分析也是富于挑战和吸引人的课题。近年来，一些统计学家钟情于分析富有噪声但又有意义的微阵列数据。这些先驱研究者们创造了大量的方法。一些方法用途广泛，为常见的问题提供了解决方法（如BLAST），一些能成功用于解决特定的生物学问题，但应用于更广泛领域的能力则有限，一些方法虽然被阐述得很有意义，但不能用于解决生物学问题。

基因组测序计划从两个方面扩展了生物信息学的覆盖范围。首先，它们提供了研究的关键材料，使得该领域有了自己的使命并且能够成长起来。另一方面，生物学家们意识到基因组计划的成果能通过计算分析的方法为他们所用。因此，在这个领域发展的早期就得到了大量的来自生物学团体的关键支持。此种情形促使了一些支持生物信息学研究和培训计划的产生，并且在生物学家和生物信息学家之间建立了良好的关系，生物信息学家们及时地开发了一系列有用的工具，因此建立了学科互动和产出的可持续模式。

生物信息学的未来在有些方面是清楚的，在另一些方面则是模糊的。毫无疑问，许多生物学家赞成生物学研究中的“高通量”方法，并且正在发明各种新方法来获取数量巨大的有用信息。在发生快速测序方法的革命之后，其他大规模测量数据的方法也随之而来。迄今影响最大的是测量细胞内mRNA表达量的微阵列技术。实验室的小型化也加速了我们其他许多实验手段和能

力的发展。例如，使用二维凝胶电泳和质谱技术分析全细胞的蛋白质（和它们与翻译后修饰的蛋白质的区别）提供了海量数据。生物学家产生数据的能力和兴趣是没有限制的，因此，从这个方面来说，生物信息学是不用担心的。事实上，数据积累的速度将会加速，所以模型会变得更复杂，也会需要模拟和可视化方面更进一步的创新。

另一方面，生物信息学将以何种方式成熟起来，也是存在疑问的。一些人认为“工具开发者”和“工具使用者”之间是存在重要区别的，开发者接受技术学科的培训，并针对由使用者描述的一般需求来产生可靠的方法。他们甚至位于专门的“生物信息学”系，并和其他的工具开发者们相互联系与合作。使用者们能理解这些工具，并合作参与开发，但是他们的主要兴趣在于帮助他们解决问题。另外一些人则认为生物信息学将以一种更加整合的方式并入生物科学的框架。下一代生物学家将会拥有定量和计算领域的知识和技巧，它们将会是在生物学研究中碰到需要解决的问题而发展起来的新的表示方法和算法的主要来源。开发者和使用者之间的区别将消失，因为生物学家研究解决问题的方法将会集中于生物信息学方法和其他方法，例如实验设计。通常，答案位于中间的某处。开发工具所需要的专门的定量和计算方法来自于生物学之外，但是生物学家们会在他们的工作中使用这些技术学科的知识和方法。这本书就是为这种整合而迈出的重要一步。在《生物信息学和功能基因组学》中，Jonathan Pevsner 写下了令人感兴趣的文字。

表面上，这本书是针对那些想解决问题的生物学家的，例如，它的组织是基于基本的生物学法则，即从 DNA 到 RNA 再到蛋白质，同时也从单基因分析到多基因分析再到基因组分析；对相关方法的描述是为了帮助用户理解这些算法的关键特性，而不涉及不必要的实现上的技术细节。但是，这本书不止如此。它不仅仅描述算法，以作为解决问题时的手册之用，它也提供了近 10 年来生物学上的许多重要发现和突破。有一点不得不说：生物信息学方法是基因组革命的一部分，因此在解释生物信息学工具时，必须描述开发这个工具的生物学背景和问题。因此，这本书不仅对生物学家（他们希望解决问题）而且对计算科学家（他们希望了解生物信息学处理的生物学问题，哪些已被解决，哪些重要问题还未被解决）来说都是有用的。

要合适地开发一个工具并提供有意义的解释需要一定的生物学知识深度，这一点使我产生深刻印象。例如，对分子系统发生和进化的讨论（第 11 章）提供了分子进化整个领域的优秀教程，它和对计算工具的描述同等有用。类似地，对基因表达的讨论（第 6 章、第 7 章）非常好地概述了现在的基因表达模型，例如这些数据如何用于提出假设，以及什么情况下这些假设是不足的。最后，对基因组的综述（第 12～第 17 章）极好地概括了我们面对的各个物种的生物学挑战，阐明了模式生物的重要性，也因此清晰地说明了我们需要处理的重要信息学问题。

因此，留给我们的是激动人心的新发现，对生物学界和计算/信息科学界都是如此。对生物学家而言，我们可以为针对重要的生物学问题开发重要工具提供宝贵的指导意见。一段写作清晰的文字需要附上观点、不足、疑问、问题、测验和参考资料（在线的和离线的），其目的是为了让一个严肃的科学工作者能快速地获取在生物信息学领域中分析当前问题的领先的方法。对计算/信息科学界而言，我们需要对这些方法拟解决的关键问题做一概要的描述，如为什么这一问题比较困难，现有的方法解决到怎样的水平。因此，提供关于生物信息学的生物学动因的重要介绍，也可用来评估生物信息学的发展机会和现有进度。基于这些理由，我热心地推荐此书，同时希望您利用本书既能理解当前生物信息学工具的状态（和如何熟练地使用它们），也能了解它们背后的生物学问题（包括解决的和留下的）。

Russ Altman
斯坦福大学

前言

1. 本书的起源

这本书源于几年以前我在 Johns Hopkins School of Medicine 为介绍生物信息学和基因组学课程准备的讲义。第一次课包括大约 70 位研究生和数百名听众，他们中有博士后研究员、技术员、本科生和教职人员。参加课程的人员来自于各个领域——遗传学、神经科学、免疫学和细胞生物学的学生，对特殊疾病感兴趣的临床医生，统计学家和计算机科学家，病毒学家以及微生物学家。他们有一个共同的兴趣，就是想知道如何使用计算机来解决生物学问题。对生物信息学，我简单地定义为计算机科学和分子生物学的交叉。这个新兴的领域依赖于使用计算机算法和计算机数据库来研究蛋白质、基因和基因组。功能基因组学就是使用基因组范围的实验和计算方法来研究基因的功能。

2. 比较

本质上说，生物信息学的研究领域是关于比较研究的领域。本书的前 1/3 我们将学习如何从数据库中取得 DNA 或蛋白质序列，然后成对地相互比较它们或者搜索整个数据库。对一名对一条特定基因感兴趣的学生来说，一个自然的问题是“其他什么基因（或蛋白质）和我这个基因相关?"。

在本书的前 1/3 部分，我们沿着从 DNA 到 RNA（基因的表达）再到蛋白质的路线进行描述，并进行一系列的比较。我们会比较两个细胞株的基因表达，它们是经过药物处理和未经处理过的两个细胞株，或者是野生型小鼠心细胞和基因敲除后的小鼠心细胞，或者处于发育不同阶段的蛙细胞。这些比较会扩展到蛋白质的世界，那里我们将应用蛋白质组学工具处理多种生理条件下的复杂生物学样本。多个相关的 DNA 和蛋白质序列比对是另一种形式的比较。这种相关性可以用系统发生树来进行可视化。

本书的后 1/3 部分论述了生命之树（the tree of life），这提供了另一个层次上的比较。哪种人类免疫缺陷病毒（HIV）对我们有威胁，我们如何能通过比较各种 HIV 亚型来帮助我们开发疫苗？蚊子和果蝇是怎么相关的？脊椎动物（如鱼和人）共享哪些基因，哪些基因是各系统发生谱系特有的？

我相信是这些各种不同的比较使得新兴的生物信息学和基因组学区别于传统的生物学领域。生物学总是关注于比较。书中，我提到了 19 世纪的生物学家，如 Richard Owen、Ernst Haeckel 和 Charles Darwin，他们在生物体的层次上进行了比较研究。我们准备解决的问题实质上没有改变。我们仍在寻求对生物学的统一概念的更完整的理解，例如生命的组成部分（如蛋白质和基

因）的组织、复杂生物系统的行为和生命进化的连续性。所改变的是我们寻求这种完整理解的方法。这本书描述了富有各种基因的原始信息和基因产物的数据库，以及用来分析这些数据的工具。

3. 人类疾病的挑战

我将学生当作一名分子生物学者和神经科学者进行培训。我的实验室研究儿童脑部疾病（如 Down 综合征、孤独症和铅中毒）的分子基础，它位于 Kennedy Krieger Institute，是一家儿童发育疾病的医院（你可以从 http://www.kennedykrieger.org 获取更多的信息）。每年有超过 1 万名病人访问该机构。医院开设了各种儿童疾病（包括语言障碍、饮食失调、孤独症、智力迟钝、脊柱分裂以及脑外伤）的临床部门。其中一些是常见的疾病，如 Down 综合征（大约 700 名新生儿中有 1 名受影响）和智力迟钝；也有一些罕见的疾病，如 Rett 综合征或者肾上腺脑白质营养不良（*adrenoleukodystrophy*）。

现在，世界公共数据库中储存的碱基数量已达到近百亿，这在第 2 章中会有所描述。我们也已经得到了人类基因组的序列，而且，自 1995 年以来有数百个基因组已被测序。贯穿本书，你将会沿着科学进步的路线学会如何对 DNA 测序，研究它的 RNA 和蛋白质产物。当今，科学进步的脚步让人眼花缭乱。

但是，与此同时我们对人类疾病仍然所知甚少。有上千种疾病是由于单个基因的一个缺陷导致的病态现象。不过即使我们发现了疾病中所发生缺陷的基因，例如囊肿纤维症（cystic fibrosis）、肌营养不良、肾上腺脑白质营养不良和 Rett 综合征，但要做到有效治疗或治愈还是很困难。与单基因疾病相比，复杂疾病更常见，如孤独症、抑郁症和智力迟钝，它们可能是由多基因突变引起的。与遗传疾病相比，感染性疾病又更常见。我们对于为什么一个病毒株仅感染人类而与之紧密相关的另一个病毒株仅感染黑猩猩所知更少。我们不明白为什么一种菌株是致病的而另一菌株又是无害的。我们还不了解如何开发有效的疫苗来对付真核致病原，包括从原生动物（例如恶性疟原虫*Plasmodium falciparum* 导致疟疾）到寄生性的线虫。

近来，生物信息学工具的开发导致在这些领域取得进步的前景是鼓舞人心的。我们仅仅刚刚开始理解致病体和易感性宿主的遗传基础。我们的希望是，在生物信息学数据库中迅速积累起来的信息，能通过研究转化成对人类疾病和一般生物学机制的洞察理解。

4. 对读者的提醒

本书描述了和生物信息学与功能基因组相关的 1000 个以上的网址。所有这些地址会随时间

变化（并且一些会消失）。为了让这些网址链接与时俱进，一个相伴的网址（http://www.bioinfbook.org/）基本上维护了书中所有的网址链接，并按书中的章节排列。我们尽我们所能随时间维护这些网址。每个月，我们用一个程序扫描所有链接，并且如有必要就更新它们。

另一个网址是针对教师的，包含了对问题的详细解答（参见 http://www.wiley.com/）。

5. 致谢

这本书的写作过程是一个很好的学习经历。非常感谢许多对此做出贡献的人们，尤其是 Kennedy Krieger 研究所和 Johns Hopkins 医学院，这里的知识氛围异常浓厚。书中的各章节是由一个介绍生物信息学课程的讲义发展而来的。学习该课程的前三年中，Johns Hopkins 医学院的教员为 Jef Boeke（负责酵母功能基因组学）、Aravinda Chakravarti（负责人类疾病）、Neil Clarke（负责蛋白质结构）、Kyle Cunningham（负责酵母）、Garry Cutting（负责人类疾病）、Rachel Green（负责 RNA）、Stuart Ray（负责分子系统发生学），以及 Roger Reeves（负责人类基因组）。他们对这些领域的见解令我获益匪浅。

我由衷地感谢本书的许多审阅者，包括一群匿名的审阅者，他们提供了相当有建设性和详细的建议。审阅本书的包括 Russ Altman、Christopher Aston、David P. Leader 和 Harold Lehmann（各个章节），Conover Talbot（第 2 章），Edie Sears（第 3 章），Tom Downey（第 7 章），Jef Boeke（第 8 章和其他一些章节），Michelle Nihei 和 Daniel Yuan（第 8 章），Mario Amzel 和 Ingo Ruczinski（第 9 章），Stuart Ray（第 11 章），Marie Hardwick（第 13 章），Yukari Manabe（第 14 章），Kyle Cunningham 和 Forrest Spencer（第 15 章）和 Roger Reeves（第 16 章）。Kirby D. Smith 阅读了第 18 章并且对其他各章也提出了见解。每位同事都花了大量的时间和精力来帮助提高文中的内容，每一位都堪称导师。许多学生也阅读了书中一些章节，我要提及的有 Rong Mao、Ok-Hee Jeon 和 Vinoy Prasad。我特别感谢 Mayra Garcia 和 Larry Frelin，他们在我写作的过程中提供了宝贵的帮助。我还要感谢 John Wiley & Sons 的编辑 Luna Han 对我的鼓励。

我也要感谢 Gary W. Goldstein（Kennedy Krieger Institute 的院长）和 Solomon H. Snyder（Johns Hopkins 神经科学系主任），他们都提供了鼓励和帮助，并且让我在从事研究的同时有机会来写这本书。

我感谢我的家人，因为他们提供了关爱和支持；同时也感谢 N. Varg、Kimberly Reed 和 Charles Cohen。最后，我要感谢我的未婚妻 Barbara Reed，是因为她的耐心、忠诚和爱。

国外优秀科技著作出版专项基金资助

生物信息学与功能基因组学

BIOINFORMATICS and FUNCTIONAL GENOMICS

[美] 乔纳森·佩夫斯纳 著
(JONATHAN PEVSNER)

孙之荣 主译

化学工业出版社
现代生物技术与医药科技出版中心
·北 京·

图书在版编目（CIP）数据

生物信息学与功能基因组学/[美] 乔纳森·佩夫斯纳（Pevsner J.）著；孙之荣主译. —北京：化学工业出版社，2006.5

书名原文：Bioinformatics and Functional Genomics

ISBN 7-5025-8383-1

Ⅰ.生… Ⅱ.①佩…②孙… Ⅲ.①生物信息论②基因组-研究 Ⅳ.①Q811.4②Q343.2

中国版本图书馆 CIP 数据核字（2006）第 020646 号

Bioinformatics and Functional Genomics/by Jonathan Pevsner

ISBN 0-471-21004-8

北京市版权局著作权合同登记号：01-2005-2537

生物信息学与功能基因组学

[美] 乔纳森·佩夫斯纳 著

孙之荣 主译

责任编辑：邵桂林 郎红旗

责任校对：陶燕华

封面设计：胡艳玮

*

化学工业出版社
现代生物技术与医药科技出版中心 出版发行

（北京市朝阳区惠新里 3 号 邮政编码 100029）

购书咨询：(010)64982530

(010)64918013

购书传真：(010)64982630

http://www.cip.com.cn

*

新华书店北京发行所经销

北京永鑫印刷有限责任公司印刷

三河市万龙印装有限公司装订

开本 787mm×1092mm 1/16 印张 45¾ 字数 1130 千字

2006 年 6 月第 1 版 2006 年 6 月北京第 1 次印刷

ISBN 7-5025-8383-1

定 价：95.00 元

原书缺页

原书缺页

原书缺页

原书缺页

原书缺页