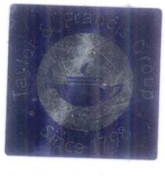
生物信息学数据分析丛书

理解生物信息学

Understanding Bioinformatics

〔英〕M. 泽瓦勒贝　J.O.鲍姆　著
李亦学　郝　沛　主译



科学出版社

理解生物信息学

Understanding Bioinformatics

生物信息学数据分析丛书

Understanding Bioinformatics

理解生物信息学

〔英〕M. 泽瓦勒贝　J. O. 鲍姆　著

李亦学　郝　沛　主译

科学出版社

北京

图字：01-2009-1133 号

内 容 简 介

本书是一本集生物信息学专业参考书和教材于一体的书，共分为 7 部分：基础知识、序列联配、进化过程、基因组特征、二级结构、蛋白质三级结构、细胞和组织，以及附录和字符表等。每部分由不同章节构成，大多数章节可以被归为应用章节或理论章节。因此在每部分开始时，都有应用章节，描述了特定研究领域较实用的方面。理论章节则紧随其后，解释了其科学、理论基础以及在已有应用中所使用的技术。本书还提供了思维导图、流程图、扩展阅读等其他书不常见的内容，以供读者能够在每一章、每一节开始时对整体内容有所把握，并能够了解更多扩展知识、发展技能的参考文献。

本书适合分子生物学、生物信息学专业及生物医学领域的师生和研究者参考使用。

图书在版编目(CIP)数据

理解生物信息学/（英）泽瓦勒贝（Zvelebil，M.）等著；李亦学，郝沛主译. —北京：科学出版社，2012

书名原文：Understanding Bioinformatics

（生物信息学数据分析丛书）

ISBN 978-7-03-032832-8

Ⅰ.①理… Ⅱ.①泽… ②李… ③郝… Ⅲ.①生物信息论-高等学校-教材 Ⅳ.①Q811.4

中国版本图书馆 CIP 数据核字（2011）第 239302 号

责任编辑：李 悦 孙 青/责任校对：钟 洋

责任印制：钱玉芬/封面设计：陈 敬

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

源海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

*

2012 年 1 月第 一 版 开本：787×1092 1/16

2012 年 1 月第一次印刷 印张：38 1/2 插页：22

字数：1 263 000

定价：168.00 元

（如有印装质量问题，我社负责调换）

译者名单

主　译　李亦学　郝　沛

译　者

第1章　李　芸（yli01@sibs. ac. cn）

第2章　李　芸（yli01@sibs. ac. cn）

第3章　叶　琳（yllilith@gmail. com）

第4章　叶　琳（yllilith@gmail. com）

第5章　余　曜（yuyao84@gmail. com）

第6章　许　涛（xutaoseu@hotmail. com）

第7章　李　虹（lihong@sibs. ac. cn）

第8章　叶　琳（yllilith@gmail. com）

第9章　王　振（zwang01@sibs. ac. cn）

第10章　董　潇（dongxiao@sibs. ac. cn）

第11章　平　捷（pingjie811@gmail. com）

第12章　叶　琳（yllilith@gmail. com）

第13章　王靖方（jfwang8113@gmail. com）

第14章　王靖方（jfwang8113@gmail. com）

第15章　俞　辉（yuhui@sibs. ac. cn）

第16章　叶志强（zqye@scbit. org）

第17章　李圆圆（yyli@scbit. org）

附录A　叶　琳（yllilith@gmail. com）

附录B　李　芸（yli01@sibs. ac. cn）

附录C　李　芸（yli01@sibs. ac. cn）

译者序

作为世界进入信息化时代的重要标志之一，科学数据的积累速度在迅速增加。人类社会最近30年所积累的科学数据总量已经超过了人类5000年发展历史所积累的数据量总和。科学数据的大量积累是重大科学规律发现的基础。海量生物学数据资源的快速查询、深度解读和综合利用已经成为现代生命科学和生物医学发展新的巨大需求。正是这种强大的需求推动了21世纪信息技术和生物技术两大学科在更大的深度和广度上进一步交叉融合，推动新兴科学——生物信息学——真正成为生命科学研究、生物医药技术发展的不可或缺的关键技术和驱动力。现在已经没有人会怀疑，生物医药技术的高速发展需要生物信息学不断提供强有力的帮助和技术支撑。

但是生物信息学的普及和应用总体来看还远远落后于生物学数据的增长，人才极为匮乏。我们急需培养大量的生物信息学专业人才，生物学家也需要对生物信息学有比较深入的理解。但是一直以来，好的生物信息学专业参考书籍和教材的阙如，使得我们在人才培养和学术普及方面缺乏标准，难以取得令人满意的进展。对此，我一直引为遗憾。几年前，科学出版社的编辑找到了我，拿出了Zvelebil博士著述的《理解生物信息学》一书，她们审慎地询问此书是否值得翻译出版。拜读此书，我欣喜地发现，这本书正是我和我的同事们作为生物信息学研究者一直在寻找的“好的”生物信息学专业参考书籍和教材。我立即表示，这本书非常值得翻译和介绍给我国的生物信息学专业领域的学生及研究人员，经过同科学出版社的同志们讨论，马上决定由我的团队主笔翻译此书，交由科学出版社出版。

Zvelebil博士是英国著名的肿瘤生物信息学家。早在1993年就在英国伦敦大学肿瘤研究所创建了生物信息学的研究小组，而那个时候“Bioinformatics”这个英文单词还没有出现在英文词典中。近20年来，她的研究范围纵贯生物信息学的几乎所有领域，并且不断与时俱进。近年来她又十分关注第二代测序技术的发展，不断在新的高通量生物技术相关的生物信息学研究方向上有重要的研究成果产出。这本著作《理解生物信息学》是Zvelebil博士对其多年来从事肿瘤生物信息学研究工作的一个详细总结。在这本书中，她结合自己的研究实践，从肿瘤生物学和计算生物学两个不同的视角出发，令人惊叹地实现了生物和计算的有机融合，给计算生物学注入了生物学的活力，阐明了生物信息学的精髓所在。本书分为7个部分，共有17个章节以及3个附录，涵盖了生物信息学的各个方面，内容丰富全面，系统地阐述了生物信息学的理论、技术与方法，内容涉及生物信息学的发生、发展和前沿与动态。各个章节的选取体现了作者对生物信息学的深刻理解，尤其可贵的是，全书论述深入浅出，使初学者非常易于理解，而这一点却往往为许多生物信息学的专业书籍所忽视。此外，本书同时又提供了拓展阅读的内容，使专业研究者的需求也能得到满足。

非常感谢我的研究团队和同事，没有他们的辛勤劳动，这本书是不可能付梓的，也非常感谢科学出版社的编辑，没有她们的介绍，这本书也可能到今天还只是被淹没在众多英文书籍中的一本专业书籍，作为“阳春白雪”而仅为“小众”所欣赏。

望本书帮助更多的读者了解生物信息学，走近生物信息学！

李亦学

2012.1.10

前　言

在过去 15 年内，互联网的影响以及不断发展成熟和准确的生物信息技术为生物医学研究所产生数据的分析带来了革命性的进步。我们希望分子生物学及生物医学领域的所有研究者都能进行序列分析相关领域的工作，同时还希望他们能进行蛋白质结构分析，甚至掌握更高级的生物信息学技术。

当我们在 20 世纪 80 年代早期开始研究时，由于数据库以及用户友好的应用程序都还没有得到开发且都只能安装在实验室的计算机上，因此所有在现今已被囊括到生物信息学范畴中的技术，其操作都仅局限于专家。到 90 年代中期，我们已能从互联网上获得很多数据库以及分析程序。那些产生数据的科学家便开始了类似于序列自身联配等的工作。然而，这些数据所需的全面的训练集的产生却稍显延迟。在 90 年代末期，我们开始将生物信息学技术扩展到大学本科生和研究生水平，并很快意识到我们需要有一本可以起到桥梁作用的教科书。这是因为简单的导论往往将重点放在结果上，几乎不考虑其背后的科学原理，而非常详细的专著则往往着重于阐述一系列技术背后的理论基础，而教科书则能填补两者间的不足。

因此，一方面，我们想要将解释程序方法的材料包括进来，因为我们相信若要进行合理的分析仅仅了解如何使用这些程序，以及它能产生什么类型的结果（和错误!）是不够的。我们还必须对这些程序所使用的技术和所基于的原理有一定的了解。另一方面，我们也希望这本书对生物信息初学者来说是容易理解的，同时我们也意识到即使是较高年级的学生，对某个应用的用途偶尔也只需要一些速览，而并不需要阅读全部基础理论。

为了解决这一明显的两难问题，我们将应用和理论归于不同的章节。在整本书中，我们所写的专门的应用章节为生物信息应用提供一定的实用知识，这部分内容浅显且易于理解。在大多数时候，一个应用章节后往往紧跟着一个理论章节。理论章节解释其程序方法以及背后的科学原理。我们发现这不可避免地造成了两者间小部分内容的重叠，但如果这能让读者自由选择他们能从事哪种水平的生物信息学课题，那对我们来说也只是一个小小的牺牲。

在此我们撰写了一本对任何生物信息学新生来讲都易于理解，且在研究生学习中也能继续使用的书。这本书假设读者对生物学背景都有一定的了解，如基因和蛋白质结构等，而重要的是要知道有更多的类型存在，而不是仅仅记得一年级课本中所提到的几个典型例子。此外，从一定的数学水平来详细描述技术也是需要的，这更适合于较高年级的学生。我们发现很多生物信息学的研究生都有一定的计算机科学和数学基础，他们将在书中发现很多熟悉的算法策略，但在这里这些算法策略却应用在了他们所不熟悉的领域。在他们阅读这本书的时候，将发现要想真正胜任生物信息学的工作，他们还需要一定的生物医学知识。

任何编书的过程似乎都会遇到怎样选择编入书本的内容抉择，生物信息学作为一门学科已经发展到了相当的程度，我们在编写时必须非常小心，不能为了将所有可能的主题都融入本书而降低了其教学价值。我们已经尽可能多地收入了广泛内容，但是仍有一些舍弃。例如，我们没有讲述从单个解码来构建一个核苷酸序列的方法，也没有讲述那些更为专业的基因组注释内容。

最后一章是对发展极快的系统生物学主题的介绍。同样地，我们不得不在包括更多内容和有限的可用篇幅之间作一权衡。但是我们希望这一章能让读者了解它所包括的主题以及它所想要回答的问题。虽然这一章不可能回答每个读者提出的关于系统生物学的问题，但是如果它能促使更多人投身到更深入的研究中，这就已经是一个进步了。

我们想要感谢很多人，他们在本书撰写工作中给予我们很多帮助。若没有他们的热情以及 Matthew Day 的支持，我们是不可能完成的。其中 Matthew Day 在整个过程中指导我们完成了第一份草稿。从第一份草稿到全书的完成，若没有 Chris Dixon、Dom Holdsworth、Jackie Harbor，以及其他来自 Garland Science 的研究者的支持也是不可能实现的，在写书期间他们给了我们很多有价值的建议和鼓励。我们还要感谢 Eleanor Lawrence 将我们的文本整理成型，并感谢 Nigel Orme 为本书制作美妙的插图。我们收到了来自很多其他地方的鼓舞和鼓励，不能一一记述，但还是要感谢我们的学生和阅读我们草稿的人。

最后，我们还要感谢朋友和家人在我们写这本书的时候所付出的艰辛。特别是，Jeremy Baum 想要感谢他的妻子 Hilary 的支持和忍耐，Marketa Zvelebil 想要特别感谢她的母亲 Martin Scurr、Nick Lee 以及与她一起工作的同事。

Marketa Zvelebil
Jeremy O. Baum
2007 年 5 月

给阅读者的短笺

本书的结构

应用章节和理论章节

在组织本书时，经过深思熟虑，我们决定通过两种方式将不同章节组织编排在一起。首先，根据主题将章节组织成 7 个部分。在每一个部分中，又有第二个水平的组织，且这一组织方式稍异于传统的组织方式，即大多数章节可以被划到应用章节或理论章节。在本书设计时，我们希望这本书既易于使想要对生物信息应用有所了解的学生阅读，也易于那些想要知道这些应用如何实现并编写自己的应用的学生阅读。因此在大多数部分开始时，都有应用章节，它描述了特定研究领域较实用的方面，并希望它能起到有用的传单式介绍（hands-on introduction）的效果。理论章节则紧随其后，它解释了其科学、理论基础以及在已有应用中所使用的技术。在获得了一定的运行程序的经验后，读者更需要对这一部分进行阅读。为了真正成为这些技术的专家，你需要阅读并理解这些技术性的方面。每一章的开始页和内容表格都很清楚地提示了这一章是应用章节还是理论章节。

第 1 部分：基础知识

基础知识为本书涉及的关键知识提供了 3 个介绍性章节。前两章包含的内容对具有生物医学背景的人来说应该是非常熟悉的。第 1 章描述了核酸结构以及它们在生命系统中扮演的一些角色，包括对基因组 DNA 如何转录成 mRNA 并随后翻译成蛋白质的简单描述等。第 2 章描述了蛋白质的结构和组织。这两章都只介绍了最基本的信息。从任何角度讲，掌握这些信息都不足以能从事任何严谨的工作。设置这两章的目的是为了提供足够的信息使本书前后连贯。第 3 章从非常初级的水平描述了数据库。很多从事生物医学研究的工作者都有需要分析的大数据集，它们需要以一个方便且可行的方式进行存储。数据库则为这一问题提供了一个完整的解决方案。

第 2 部分：序列联配

序列联配包含 3 章，包括了各种序列分析的内容，所有内容都与相似性鉴定有关。其中，第 4 章是对这方面的实践性介绍，包含不同分析方法可能存在的问题及正确结果的例子。序列分析有很多种不同的技术，第 5 章和第 6 章描述了其中的几种。第 5 章将重点放在两序列联配以及用于数据库搜索的特定方法上，期间描述了很多技术，包括动态规划、后缀树（suffix tree）、哈希（hashing）算法和链式方法（chaining）。第 6 章描述多序列分析涉及的方法，定义常见的模式、定义相关蛋白质组成家族的谱式（profile）以及构建多重联配，本章展示的一个关键技术是隐马尔可夫模型（HMM）。

第 3 部分：进化过程

进化过程展示了用于从一个序列集合获得系统发育树的方法。这些树是序列进化历史的重构，其前提假设是认为它们共享同一个祖先。第 7 章解释了涉及的基本概念，随后展示了不同的方法是如何应用于不同的科学问题的。第 8 章详述涉及的技术以及它们如何与研究进

化过程的假设联系起来。

第 4 部分：基因组特征

基因组特征描述解析原始基因组序列数据所需要的分析。虽然当一个基因组序列在研究杂志上发布时，一些初级的分析已经被执行过了，但是通常在这之前也有些可用的但却未分析过的序列。本部分描述了一些用于尝试在序列上定位基因的技术。第 9 章描述了一些可用的程序，展示其结果的复杂性并阐述一些可能的缺陷。第 10 章展示了一个对所用技术的调研，特别是不同的马尔可夫（Markov）模型以及单独每个组分，如核糖体-结合位点的模型是如何能构建出整个基因的模型的。

第 5 部分：二级结构

二级结构由两章构成，着重介绍基于序列（或一级结构）预测二级结构的方法。第 11 章介绍二级结构预测的方法，并讨论了解释预测结果的各种方法和途径。接着，讲述如何处理特殊的二级结构预测问题，如蛋白质的跨膜区、由两个或两个以上的 α 螺旋组成的超螺旋结构、亮氨酸拉链结构和 RNA 二级结构。第 12 章讲述二级结构预测方法的基本原理和细节，从基本概念到深入理解预测技术，如神经网络、马尔可夫模型在此类问题中的应用等。

第 6 部分：蛋白质三级结构

三级结构是第 5 部分内容的延伸，它为蛋白质三级结构和四级结构的预测与建模提供保证。第 13 章为读者介绍了能量函数、最小化以及从头开始预测（*ab initio* prediction）的概念。这一章更详细地介绍了穿线法，并将重点放在蛋白质结构的同源建模，以渐进的方式使读者掌握这些知识，该章的结尾用一个具体例子阐述了这一技术。第 14 章包含了进一步分析结构信息的方法和技术，并描述了结构和功能关系的重要性。这一章告诉我们折叠的预测是怎样帮助我们鉴定功能的，并对配体对接和药物设计进行了介绍。

第 7 部分：细胞和组织

细胞和组织由 3 章组成，较详细地描述了表达分析，并简单介绍了系统生物学。第 15 章介绍了分子蛋白质和基因表达数据的可用技术，向读者展示了从这些实验技术中所能获得的信息，以及这些信息是如何用于进一步分析的。第 16 章展示了第 15 章涉及的一些聚类技术和统计学，这些技术也常用于基因和蛋白质表达分析。第 17 章是一个单独的章节，它描述了系统建模的过程，向读者介绍了系统生物学的基本概念，并展示了这些前景广阔且快速发展的领域在将来可能获得的成果。

附录

本书提供了 3 个附录，将文内主要部分涉及的一些概念进行了扩展介绍，这对好学的以及高级阅读者非常有用。附录 A 介绍概率和贝叶斯分析（Bayesian analysis），附录 B 主要与第 6 部分关联，介绍分子能量函数，而附录 C 则描述了函数优化的技术。

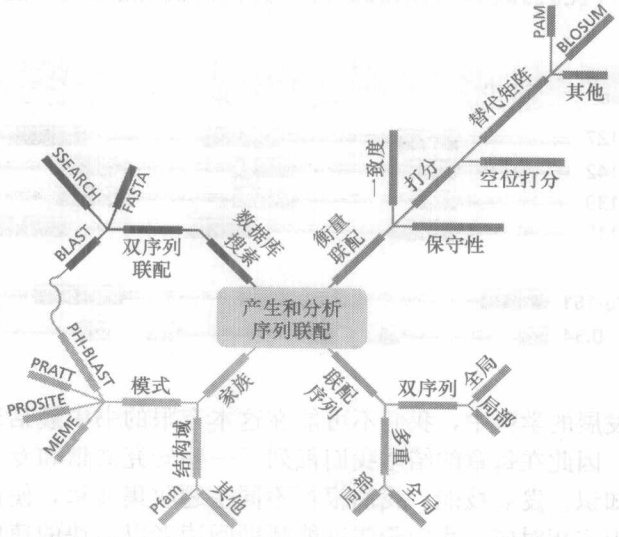
各章的结构

学习效果

每章开篇都有一个学习效果列表，它总结了该章所涉及的主题，可作为一个反馈清单（revision checklist）。

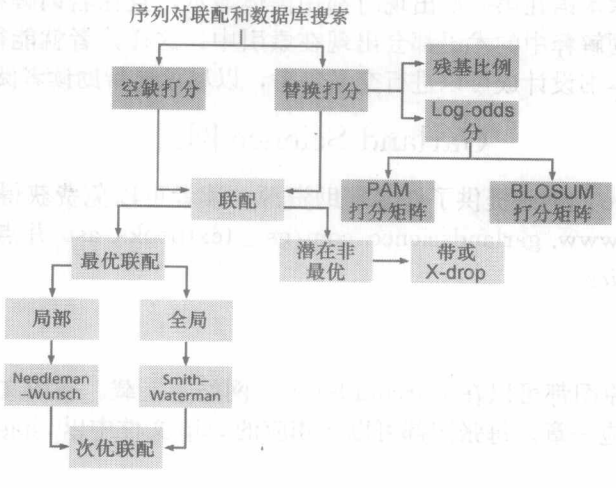
思维导图

每一章都含有一个思维导图，这是本书一个特别的教学特征，它确保每个学生都能看到并记住一些特定应用中所必需的步骤。例如，第 4 章的思维导图，包含了“产生和分析序列联配”这一主题的 4 个主要方面（area）：衡量联配、数据库搜索、联配序列和家族。为清晰起见，每一方面都用不同深浅的黑色标注，并包含所涉及的关键概念。从视觉上帮助读者对这一领域所讨论的材料有一个大致的了解。偶尔地，思维导图的两个独立方面也可能有着重要的关联。例如，BLAST 和 PHI-BLAST 之间就存在一定联系，后者是直接从前者发展而来的，但却具有非常不同的功能，因此，这两者将出现在思维导图两个不同的方面。



流程图

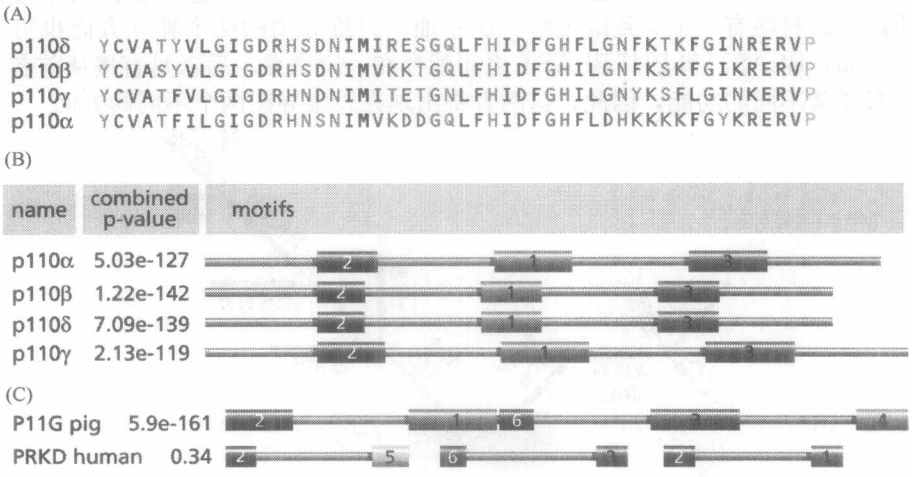
每一章的每个小节都有一个流程图以帮助读者记忆该小节所涵盖的主题。作为示例，下面给出了第 5 章的一个流程图，其中在本节将要解释的概念用深灰色框标注，且相互间用箭头连接起来。例如，两种主要类型的最优联配：局部和全局将在本章的这一节描述。那些已在之前小节描述过的概念用浅灰色框标注，这样我们就很容易了解本节涉及的主题和已介绍过的主题间的联系。例如，构建联配需要为空缺（gap）打分的方法和为替换（substitution）打分的方法，两者都已经在这一章描述过了。通过这种方式，整章涉及的主要概念以及相互间的关系就能渐渐地被构架出来。



插图

每一章都配有插图。插图的配置是经过充分考虑的，以保证既简单易懂又与本书其他章节保持连贯一致。

图 4.16 便是一个示例。



扩展阅读

在这么一个快速发展的学科中，我们不可能在这本有限的书中囊括现有的所有知识，更不用说将来的发展了。因此在每章的结尾我们都列了一些研究文献和专业著作的参考文献以帮助读者进一步扩展知识、发展技能。我们根据不同主题收集文章，使得扩展阅读中每节都与这一章相应小节的内容相对应。我们希望这能帮助阅读者以最快的速度找到他们感兴趣的扩展材料。

字符表

生物信息学需要使用很多符号，对还不了解生物信息的人来说，许多符号都是不熟悉的。为了帮助读者了解本书适用的符号，我们在本书后面给出了引用的每个符号、它的定义以及它在本书最常出现的位置的列表。

名词解释

在文中，所有技术术语在第一次出现时都用黑体显示，且在名词解释中列出其相应的解释。此外，每个在名词解释中的术语都会出现在索引中，这样读者就能很快获得详细介绍这一术语的相应页码。本书设计成可以进行交叉参考，以尽可能帮助读者阅读。

Garland Science 网址

在 Garland Science 的网站提供了许多辅助资源。读者可以免费获得而不需要任何密码。更详细的信息可登录 www. garlandscience. com/gs _ textbooks. asp 并点击链接进入 *Understanding Bioinformatics*。

图版

本书所有的英文原图都可以在 Garland Science 网站上下载。插图文件以 . zip 格式保存，其中每个 . zip 文件对应一章。每张图都可以从相应的 . zip 文件中以 . jpg 的格式解压出来。

更多材料

Garland Science 的网站还包括一些与本书主题相关的额外的材料。7 个部分中任何一部分都对应一个 .pdf 文件，它通过一系列与这些章节内容相关的有用的网址链接，能链接到一些的有用的数据库、文件格式定义、免费的程序以及允许数据在线分析的服务器上。此外，在阐述分析方法时所用到的数据也会被提供。这就允许读者对同一数据重新进行分析，重现本书所显示的结果，并尝试其他技术。

致谢名单

Stephen Altschul	National Center for Biotechnology Information, Bethesda, Maryland, USA
Petri Auvinen	Institute of Biotechnology, University of Helsinki, Finland
Joel Bader	Johns Hopkins University, Baltimore, USA
Tim Bailey	University of Queensland, Brisbane, Australia
Alex Bateman	Wellcome Trust Sanger Institute, Cambridge, UK
Meredith Betterton	University of Colorado at Boulder, USA
Andy Brass	University of Manchester, UK
Chris Bystroff	Rensselaer Polytechnic University, Troy, USA
Charlotte Deane	University of Oxford, UK
John Hancock	MRC Mammalian Genetics Unit, Harwell, Oxfordshire, UK
Steve Harris	University of Oxford, UK
Steve Henikoff	Fred Hutchinson Cancer Research Center, Seattle, USA
Jaap Heringa	Free University, Amsterdam, Netherlands
Sudha Iyengar	Case Western Reserve University, Cleveland, USA
Sun Kim	Indiana University Bloomington, USA
Patrice Koehl	University of California Davis, USA
Frank Lebeda	US Army Medical Research Institute of Infectious Diseases, Fort Detrick, Maryland, USA
David Liberles	University of Bergen, Norway
Peter Lockhart	Massey University, Palmerston North, New Zealand
James McInerney	National University of Ireland, Maynooth, Ireland
Nicholas Morris	University of Newcastle, UK
William Pearson	University of Virginia, Charlottesville, USA
Marialuisa Pellegrini-Calace	European Bioinformatics Institute, Cambridge, UK
Mihaela Pertea	University of Maryland, College Park, Maryland, USA
David Robertson	University of Manchester, UK
Rob Russell	EMBL, Heidelberg, Germany
Ravinder Singh	University of Colorado, USA
Deanne Taylor	Brandeis University, Waltham, Massachusetts, USA
Jen Taylor	University of Oxford, UK
Iosif Vaisman	University of North Carolina at Chapel Hill, USA

目　录

第2部分 序列联配