

第二版前言

第二版的生物信息学(序列与基因组分析)比第一版的读者更广泛,它不仅面向想学计算和统计方法的生物学家,而且也适用于想学生物学的、尤其是遗传学和基因组学的计算生物学家。章节指南介绍了每一章所需的基本计算学(统计学)和生物学背景,接着是本章要学习内容的提纲,后面提供网络资源(表格和正文中仍会显示相关 URL),习题部分用于强化本章概念和相关技术。最后,所有网络资料都均用文字形式表述出来,放在一处。所有原来的章节都经过了相应的更新、修改和重写。

第二版里增加了三章新的内容。原来第三章里的序列比对的概率和统计分析现在单列为第四章,并增添了以序列分析为基础的假设检验和预测准确性检验。第 12 章和第 13 章涵盖了原来没有的 Perl 语言编程和芯片分析。这些比较高级的内容需要读者有一定的专业背景,但可增加各生物信息学最相关内容。增添的内容代表了国际前沿进展,是本书第二版宝贵的补充。在此,我非常感谢 Arizona 大学的同事们贡献了这些章节的内容。一遍遍修改润色非常耗时、艰苦,可他们总是非常合作,不吝时间。

第 12 章由 Nirav Merchant 和 Susan Miller 提供,他们是经验丰富的计算机系统和软件专家。本章具体叙述了使用和编写 Perl 脚本和模块满足不同任务,也包含了数据格式和建立关系数据库等内容。这章里列举的许多 Perl 脚本例子都可以从此书网站上直接下载作为项目模板使用。我们希望第 12 章可以作为很多实用 Perl 程序的起点以支持大规模基因组项目。

第 13 章由统计学家 David Henderson 博士提供。他擅长 QTL 分析、试验设计和统计分析,并在芯片试验方面具有丰富的经验。本章旨在帮助生物学家从芯片的基因表达数据中提取重要的信息。通常生物学家不习惯于设计涉及大量数据的试验以及从这些试验中提取信息。芯片试验麻烦的原因有两种:一是表达数据有各种来源的背景噪音,在复杂噪音中寻找哪个基因表达有差异是很困难的事;二是芯片试验的结果往往是一长串混乱的难以分类的基因。我们为解决这两种问题提供思路:一是为去除噪音达到特定科研目标来设计试验提供指导;二是叙述了寻找显著性表达差异基因的方法;三是介绍了相关分析方法,包括基于不同标准寻找、验证不同分类标志(生物标志)的聚类方法。最后提供了实现这些目标的程序资源。基于这些背景知识,第 13 章旨在指导试验设计使其产出尽可能多的有用信息。

大家对第一版的建设性评论也有益于第二版的改进。第一个是日本的 Yasushi Okazaki 先生,他在将此书译成日文的同时提了很多建议和修改意见。在不同场合提供帮助的还有 John Clark, Gabriel Dorado, Dan Flath, Toni Kusalik 和 Etsuko Moriyama。

此书离不开我生物信息学的同事 Ritu Pandey 和 Rob Klein 的支持以及 Arizona 大学的经济支持,尤其是 Vicki Chandler, Gene Gerner Rich Hoff。谢谢 Walt Klimecki 在图 11.2 提供的帮助和 Roger Miesfeld 在图 9.13A 的协助。我也很感激 Beck Nickerson 给许多章节提出的批评和建议。Pick Weil Lau 帮助我们校对了图片库, Eric Shen 从本书网站上收集整理了意见。

最后,要感谢冷泉港实验室出版社的员工的支持。没有他们的努力此项工作不可能如此成功。Judy Cuddihy 改进了章节的格式,对全文做了极其有益的建议,并给予了大量的鼓励、支持使此书能按时完成。Mary Cozza 指导我整理了参考书目, Patricia Barker, Kathleen Bubbeo, Daniel Debruin 和 Susan Schaefer 在时间紧迫的情况下高效工作, Jan Argentine 和 Denise Weiss 见证了他们为这本书付出的努力。

最后,我还要感谢我的妻子 Jennifer Hall, 感谢她在本书写作过程中所给予我家庭方面的支持。

D. W. M.

2004 年 5 月

David W. M. 图森

第一版前言

此书主要献给那些想理解序列和结构分析方法的科学家。我深信一个使用计算机程序的人应该理解这些程序的原理,所以我的主要目的之一就是让生物学家在使用程序时领会程序背后的算法、假设、局限性以及使用策略。我尽量避免使用复杂的数学公式和注解,而是尽可能地给以简单的数字化例子。希望此书会对那些志在多学一点生物信息学中生物学问题的计算生物学家仍然有用。此书可以用作实验室参考书,或者生物信息学教科书,而不仅仅是一套特定序列分析程序的使用指南。

大部分章节都包含了一个流程图来建议如何有序使用本章中所讨论的方法。这种图表的例子比较稀少,因为需要的假设和过分简化不一定合理。希望这些图表会对那些本领域的新手有用,而对那些较有经验的人,我期望他们能以其他更好的方式达到相同的目的。

有很多参考文献也提供了获取程序或应用方法的网站和 FTP 站点。有时候我对一些常用并重要的 blast 和 clustalW 程序提供大量如何使用的说明以及如何分析结果。然而还有许多其他重要的关于生物序列、基因组分析的工具和方法,虽然时间和篇幅限制,我仍尽量包括进来。我对比较简单的序列分析给予了特别关注,比如寻找限制性酶切位点、翻译序列和组成分析。做这些分析的商业和免费的软件包很多,也有用于基因组分析的商用软件包。

在撰写此书(也是我的第一本书)的过程中,我发现已发表的文献中的信息远比我想像的多。我尽量详尽地把序列和基因组分析中的重要问题包括进来,仍有很多优秀的文章限于时间和篇幅没有引用到,在此我向那些作出重要贡献却未提及到的同事表示歉意。由于篇幅限制及生物信息学本身日新月异,没包括在此文中的资料等和引用的站点链接、例子以及问题都将收录到本书网站(<http://www.bioinformaticsonline.org>)

此学科令我感受深刻的一个特点就是大多数研究者,尤其是本领域的前沿学者都愿意与同事分享他们的研究成果。我有幸结识了几位这样的先驱,尤其是 David Lipman, Hugo Martinez (我与其共度了学术休假年) 和 Temple Smith。这些学者取得的巨大成就因为愿意和同行免费共享他们努力的成果而更加备受称赞。正是因为这样,他们为学术界和商业界在序列分析领域的成功起到了非常重要的作用。

这个大项目需要许多的支持和帮助。本书的一部分来自于 Arizona (亚利桑那) 大学 1999 至 2000 学年的“生物信息和基因组分析”课程的课堂笔记。许多学生提出了非常有意义的建议,并且对错误的查找很有帮助;我想要特别感谢 Bryan Zeitler, 他做了很多校正工作。书中其他错误将会在本书网站上修正。我要衷心感谢 Bill Pearson 提供了 FASTA 软件包的信息,感谢 Julie Thompson 和 John Kececioglu 对第 4 章提出的意见,感谢 Steve Henikoff 帮助我阅读了第 3 章内容,感谢 Michael Zuker 对第 5 章提出的意见。Bill Montfort 为第 9 章提供了 PDB 文件的信息,另外,在第 8 章中, Roger Miesfeld 提供了复杂基因调控的例子。Jun Zhu 帮助解答了第 3 章中关于贝叶斯块联配校对器的问题。在过去的三年中,为了完成或修改其他的章节,我不得不放弃一些会议和讨论会,我所在的系给了我最大的理解和支持。这段时间里, Rob Han 和 Juwon Kim 每次都用非常短的时间为我定期地提供了大量的文献和相关书本的章节,使我有更多的时间消化这些信息。冷泉港实验室出版社的编辑 Judy Cuddihy 指导了整个写作过程,耐心地督促我遵循合理的写作进度,并给予我鼓励,帮助我完成这个项目。Elisabeth Cuddihy 确认了大部分的网站,仔细检查了书中的方程和数字示例,同时帮助撰写了部分术语表。我还要感谢出版社开发部门的 Joan Ebert 和 Jan Argentine 和生产部门的 Pat Barker 和 Denise Weiss。

最后,我还要感谢我的妻子 Jennifer Hall, 感谢她在本书写作过程中所给予我家庭方面的支持、耐心和理解。

David W. Mount