前言

本书是为了作为高年级本科生和研究生的生物信息学教材而编写的。书中大多数章节取材于两位作者分别在加拿大滑铁卢大学和新加坡国立大学开设的生物信息学课程的讲义。该书也可以给希望了解生物信息学和计算生物学基础和工具的科研与教学人员作为参考书。

对生物信息学有兴趣的人员往往来自于不同的领域包括数学、计算机、生命科学、甚至生物科技工业。因此，本书的阅读不需要读者具有很多分子生物学方面的知识。在每章节所需要的地方，我们根据情况插入了一些必要的生物学背景知识。我们在写作的时候也不假定读者具有很强的计算机算法和概率方面的背景，尽管对这两方面的基本了解对于阅读这本书的某些章节是很有帮助的。为此，我们使用通俗的语言详细地解释了一些重要的算法概念和概率模型。

为了给学生们提供更多的练习机会，除了在内容上有一些特殊安排外，我们还在每章的最后给出了很多练习题。大多数习题比较容易，可以作为学生作业的题目。个别较难的用星号标注。其中尚未解决的问题特别注明，可以作为研究生的研究课题。

本书共有六章。在第一章里，作者尝试使用比对图通俗地介绍生物序列比对的概念，动态规划算法，和打分矩阵。

第二章讲述有关同源搜索的各种重要的概念和快速比对算法。特别的是我们在该章最后一节系统地介绍了散核技巧。这是近年来提出的一个非常重要的设计比对快速算法的技巧。这是其它生物信息学教科书所不具有的一个特色。

在第三章里，我们主要讲述如何有效地比对多个生物序列。我们介绍了用于表示蛋白结构域序列保守性的各种概念和工具。我们还介绍了各种渐进式比对策略和常用的多序列比对程序。

第四章先介绍隐马尔科夫模型的三大基本问题和相应的算法。隐马尔科夫模型其实就是有限自动机的随机版本。在这一章，我们尝试从形式语言和自动机的角度来理解隐马尔科夫模型。我们希望这对计算机系的学生理解这一重要模型有帮助。然后，我们还讨论了它在蛋白质序列分析以及基因预测中的应用。

在第五章里，我们讲述进化树的基本数学性质、构建进化树的各种算法和两个实际问题。我们还简单地介绍了进化树应用在比较基因学和综合基因组学 (metagenomics) 中的两个例子：推断基因复制历史和构建一个生物特性的祖先状态。

第六章介绍了计算蛋白质组学的一些基本问题以及相应的算法和统计方法。计算蛋白质组学是生物信息学中近十多年来新兴的一个重要研究领域，主要是通过计算来分析蛋白质的质谱数据，从而推断出蛋白质的序列、修饰、

定量等信息。目前这个领域发展的非常快，而介绍这个领域的教材又非常少，无论是中文还是英文都是如此，所以这一章也是本书的另一个特色。

随着高通量 DNA 测序的成本降为每个人都可负担得起的程度以及蛋白质测序工具的日益成熟，本书所讲述的知识是每个生物信息工作人员所应据备的基础知识。本书大部分章节之间是相互独立的。在讲授生物信息学课程的时候，教师可以按照各自不同需求选择内容。第一至六章的授课时间分别是大约八、八、六、六、十和十小时。

最后，我们提醒读者生物信息学是一门仍在发展的学科。当这本书和读者见面的时候，也许某些章节所讲述的内容已经有了新的突破。为此，我们也特别建立了一个网站（http://www.math.nus.edu.sg/~matzlx/BioTextbook）。在这个网站上，我们会继续补充本书所涉及内容的最新进展，以供读者参考。另外，为方便制作课件，我们也给使用这本书的讲师们提供书中的所有插图。

第六章、§2.3 和 §2.7 由 BM 执笔，其余部分由 LXZ 完成。在长达两年的写作过程中，我们得到了许多朋友和学生们的帮助。在此，我们特别感谢高松、贺琳、林皓、宁康、姚玉华和徐魁。通过认真阅读这本书的初稿，他们帮助我们把它的内容更完美地呈献给读者。尽管我们作了最大的努力，本书难免还有一些由于我们疏忽造成的错误。如果发现任何错误，请你通过上述网站上的联系方式通知我们。除了在上述网站公布外，这些错误也将会在再版时加以更正。

马斌，滑铁卢
张洛欣，新加坡
2014 年 6 月

目 录