BIOINFORMATICS

第2版

The Machine Learning Approach

生物信息学

——机器学习方法

[法] 皮埃尔·巴尔迪（Pierre Baldi）
[丹麦] 索恩·布鲁纳克（Søren Brunak） 著
张东晖等 译
李衍达 朱宗涵等 审校

中信出版社
CITIC PUBLISHING HOUSE

BIOINFORMATICS
The Machine Learning Approach

生物信息学
——机器学习方法

[法] 皮埃尔·巴尔迪
[丹麦] 索恩·布鲁纳克 著
张东晖 黄 颖 蔡 军 孙应飞 夏慧煜 胡驰峰 计宏凯 朱宗涵 译
李衍达 朱宗涵 张东晖 审校

本书翻译工作得到《国家重点基础研究发展规划》课题（编号：2001CB51030）的支持

中 信 出 版 社
CITIC PUBLISHING HOUSE

图书在版编目（CIP）数据

生物信息学 /［法］巴尔迪等著；张东晖等译；李衍达等审校. —北京：中信出版社，2003.5

书名原名：Bioinformatics: The Machine Learning Approach

ISBN 7-80073-708-X

Ⅰ.生… Ⅱ.①巴… ②张… ③李… Ⅲ.生物信息论 Ⅳ.Q811.4

中国版本图书馆CIP数据核字（2003）第032756号

The MIT Press

http://mitpress.mit.edu

生物信息学——机器学习方法

SHENGWU XINXIXUE——JIQI XUEXI FANGFA

著　　者：［法］皮埃尔·巴尔迪　［丹麦］索恩·布鲁纳克

译　　者：张东晖　黄　颖　蔡　军　孙应飞　夏慧煜　胡驰峰　计宏凯　朱宗涵

审 校 者：李衍达　朱宗涵　张东晖

责任编辑：陈蕴真

出 版 者：中信出版社（北京市朝阳区东外大街亮马河南路14号塔园外交办公大楼　邮编　100600）

经 销 者：中信联合发行有限公司

承 印 者：北京忠信诚胶印厂

开　　本：787mm × 1092mm　1/16　　**印　　张**：26.75　　**字　　数**：342千字

版　　次：2003年7月第1版　　**印　　次**：2003年7月第1次印刷

京权图字：01-2002-0211

书　　号：ISBN 7-80073-708-X/Q · 1

定　　价：45.00元

译者序

2002年夏天，中信出版社交给我一本英文原著，是由皮埃尔·巴尔迪（Pierre Baldi）和索恩·布鲁纳克（Sφren Brunak）两位教授编写的《生物信息学——机器学习方法》（第2版），MIT出版社于2001年出版。出版社的编辑同志告诉我，鉴于本书的学术价值及其在生物信息学领域的重要性，出版社已购买了本书的中文版权，并准备作为社里的重点图书尽快在国内翻译出版。由于本书作者在生命科学、数学以及计算机科学等多个领域都有相当的造诣，加之本书同时涉及了生物信息学的理论基础和最前沿的实际应用，出版社走访了几位专家译者，他们都不愿意承担这一艰巨的翻译工作。我用了整整一个星期的时间，认真阅读了这本书的前言、目录和一些重要章节，深感本书分量之重。在此之前，我也曾经读过几本国内出版的生物信息学著作或译著，其中大部分是有关基因和蛋白质序列分析软件、算法以及相关网络资源的工具书，而真正涉及生物信息学基本理论和最前沿应用的著作还很少。我们在实际工作中经常会利用国外的一些生物信息学的数据库和软件，分析基因或蛋白质的序列和结构，但对于这些数据库和软件背后的理论、模型和算法却所知甚少。随着国内生物信息学和生命科学等相关领域研究工作的不断深入和发展，我们的研究方向已经从积累数据和追踪国外最近进展逐步转向前沿的基础研究和新的应用开发，而这些前沿领域的研究和开发要求我们了解和掌握生物信息学的主要理论、模型和算法。为了适应这些新的研究方向，越来越多的本科和研究生专业已经或将要开设生物信息学课程。因此，国内的生物信息学领域迫切需要一本足够深入的经典教材或参考书，而本书正好可以满足这一迫切需求。正如国外专家对本书的评论中所说的："仅靠这一本书或许很难掌握生物信息学的全部内容，但如果你想理解生物信息学，此书是不可不读的。"为此，我决定接受翻译此书这一艰巨的任务。

本书的内容涉及生命科学、数学、信息科学等诸多领域的最新进展，我深知仅靠

我个人很难在短期内完成全书的翻译工作，必须邀请相关领域的专家组成翻译小组，合作完成全书的翻译和审校工作。于是，我找到了我国著名的信息科学专家，清华大学信息学院院长、生物信息学研究中心主任李衍达院士，他欣然同意主持本书的翻译工作。我们还邀请到微软（中国）公司的资深软件设计工程师张东晖先生，以及清华大学信息学院的黄颖、蔡军等多位博士，共同组成翻译小组，几易其稿，又请了多位专家参与审校，最终完成了本书的中文译稿。整个过程的艰辛难以用语言表达。为此，要感谢李衍达院士的全力支持和翻译小组全体同仁付出的宝贵精力和时间，也要感谢中信出版社青年编辑陈蕴真同志的真诚合作。本书的翻译还得到了“国家重点基础研究发展规划”课题（编号：2001CB51030）的支持，北京市卫生局干部培训中心为翻译小组提供了良好的工作条件，在此一并致谢。

本书的作者是国际著名的生物信息学专家。其中皮埃尔·巴尔迪博士是美国加州大学医学院信息和计算机科学系教授、生物化学系教授，基因组学和生物信息学研究所所长。索恩·布鲁纳克博士是丹麦理工大学生物系教授，生物序列分析中心主任。他们在生物信息学领域发表了大量的论文和著作，涉及到许多生物信息学理论、模型和算法的前沿应用和探索。本书是他们多年研究和教学工作的积累，本书的早期版本曾作为几个国际重要的生物信息学研讨班的讲义。他们在本书中详细介绍了机器学习方法的理论基础——贝叶斯概率体系，并在此基础上着重讨论了神经网络、隐马氏模型、贝叶斯网络、概率图模型以及随机文法等不同方法在序列比对、基因建模与基因发现、系统进化树等生物信息学问题中的应用。书中还专辟一章介绍了DNA微阵列和基因表达，以及相关数据的分析方法。此外，本书还分类列举了大量相关网络资源的详尽网址，以及近600条参考文献和5个包含详尽数学推导的附录，这些参考资料无疑会给生物信息学的研究和教学工作者提供非常实际的帮助。

在翻译和审校过程中，我们发现本书有几个值得关注的特点。首先，本书试图利用贝叶斯概率理论的统一框架为机器学习方法在生物信息学领域的应用建立一套完备的理论基础。书中使用大量篇幅介绍了如何在概率理论的统一框架内理解神经网络、隐马氏模型、概率图模型以及随机文法等机器学习方法，并详尽地介绍了各种建模和学习算法。作者对理论完备性的追求无疑给读者提供了很好的背景知识和扎实的理论基础。第二，本书不仅介绍了机器学习的理论和算法，还介绍了大量的实际应用。最宝贵的是书中包含了作者在应用各种机器学习方法解决实际问题中所做的深入观察、富有创造力的假设、精细的建模、真实的实验结果和透彻分析，以及不断修正假设和模型的整个探索过程。与许多流行的教科书不同，作者不仅给我们展示了当今生物信息学大厦的缩影和构筑大厦的工具，更重要的是作者带领我们经历了如何构筑这个大厦的过程，如何搭建“脚手架”的经验无论对于修补这座大厦还是构建一座新的大厦都是非常重要的。第三，

正如许多专家的评论所指出的，本书在介绍相关理论和应用的同时，还提出了生物信息学前沿领域的许多重要问题。读者不仅可以了解生物信息学的前沿领域，还可以追随原作者探索这些问题的轨迹，开始自己对前沿问题的开创性研究。

本书主要针对两类读者：一类是生物学、生物化学和医学等领域的研究人员，他们可以通过本书了解更多数据处理和机器学习的有关算法；另一类是物理、数学、统计学和计算机科学等领域的学者，他们也可以通过本书了解机器学习方法在生命科学，特别是在生物信息学领域中的更多应用。本书也可以作为相关领域的大学本科和研究生教材或参考读物。

最后，我想指出尽管本书在建立生物信息学的理论基础方面可谓是一次成功的尝试，但生物信息学作为一门新兴的跨学科的科学还处在起步阶段。在本书的翻译过程中，我们深刻地体会到来自不同学科研究人员之间的密切合作和相互理解是多么重要。虽然本书包含许多数学公式和推导，但这并不意味着生物信息学排斥那些不熟悉数学公式的生物学和医学专家，如果失去了生物学和医学专家的合作与理解，生物信息学将失去继续发展的动力和应用的基础。为此，我们真诚地希望本书能够赢得来自生物学与医学领域专家的更多理解与关注。

虽然我们尽了很大努力，以确保翻译的质量和译文的准确，但是错误之处在所难免，希望广大读者批评指正。

朱宗涵

2003年3月20日

中文版序

我们很高兴看到自己的著作《生物信息学——机器学习方法》的第2版翻译成中文。中国在人类基因组计划中做出了重要贡献，水稻基因组的测序也给世界留下了深刻的印象，包括猪基因组测序在内的一系列国际交流合作都证明中国在基因组研究上达到了很高的水平。本书的出版更从另一个方面证明了这一点。作为一个拥有悠久历史的国家，中国在当今基因组研究的浪潮中具有自己独特的优势和发展机遇。现在，中国正迅速将先进的计算科学用于高通量的基因组和后基因组技术，并与那些传统的生物技术相结合。我们很荣幸此时能够为生物信息学研究思想的全球交流以及中国下一代计算生物学家的培养，尽自己的一份绵薄之力。现在中国的研究人员与世界具有密切的国际合作，我们希望本书的出版能够增进中国国内以及中国与其他国家之间的生物信息学理论和实验研究的协调合作。

皮埃尔·巴尔迪
索恩·布鲁纳克
2003年5月

前　言

本书第1版出乎意料的成功曾使我们深感欣慰。然而，由于生物信息学持续迅速发展，本书需要一个新的版本。在过去的3年里，随着果蝇基因组测序和人类基因组工程第一个草图的完成，全基因组测序研究蓬勃发展。除此之外，其他一些高通量/组合实验技术，如DNA微阵列（基因芯片）、质谱技术等，都取得了重大进展。这些高通量的实验技术能够快速产生10^{12}字节的实验数据，拥有传统生物学方法无法比拟的优势。这一切导致了今天对计算机、统计学和机器学习技术日益强烈的需求。

后基因组时代的生物信息学

在过去5到10年中，计算机在生命科学和医学的各个领域中发挥着前所未有的重要作用。计算机分析应用的第一个高潮主要出现在序列分析中，这个方面至今有许多非常重要的问题尚未解决；在目前以及未来的一段时期内，我们尤其需要关注那些极为多样化的数据的复杂集成关系。这些新的数据类型来源于能够在细胞、器官、生物个体甚至生物群体等不同层次获取数据的各种实验技术。

新的高效实验技术，主要是DNA测序技术，是以下转变的主要动力：新技术导致描述DNA、RNA和蛋白质的线性序列数据呈几何级数增长。其他新的产生数据的技术则是传统试验方法的高度并行版本。用DNA微阵列进行基因组范围的基因表达测定；基本上如同进行上万个RNA印迹实验（northern blots），这使得在实验设计、数据处理和结果解释等方面的计算机支持成为基本要求。而这一系列的发展极大地扩展了生物信息学的研究领域。

随着基因组和其他测序项目的不断进展，研究的重点正逐步从积累数据转移到如

何解释这些数据。在未来，生物学的新发现将极大地依赖于我们在多个维度和不同尺度下对多样化的数据进行组合和关联的分析能力，而不再仅依赖于对传统领域的继续关注。序列数据将与结构和功能数据、基因表达数据、生化反应通路数据、表现型和临床数据等一系列数据相互集成。在数据量呈几何级数增长的情况下，生物信息学的基础研究将致力于解决生命科学中与系统和集成相关的问题。

如此大量的数据，在生物信息的存储、获取、联网、处理、浏览以及可视化等方面，都对理论、算法和软件的发展提出了迫切的需求。而计算机科学也从生命系统中获得启示，产生了许多新概念，包括：遗传算法、人工神经网络、计算机病毒和人造免疫系统、DNA计算、人工生命以及VLSI-DNA混合基因芯片，等等。这样的学科交叉丰富了各个相关领域，这将在未来的几十年中得到进一步发展。事实上，基于“碳”的生物体信息处理和基于“硅”的电子化信息处理之间的界限，无论是在概念上，还是在实际中，都已开始逐渐淡化。[29]

用于序列分类、弱相似性探测、区分DNA序列中的编码区和非编码区、分子结构预测、转录后修饰和功能的预测，以及重构进化史的计算工具已经成为研究的基本组成部分。这些研究是我们理解生命和进化，以及发现新药物和新疗法的基础。生物信息学已成为在生命科学和计算机科学的前沿涌现出的一门具有战略意义的新学科，它将通过各种途径影响医学、生物技术以及社会的许多领域。

庞大的生物信息数据库对数据挖掘技术提出了许多颇具挑战性的问题，也提供了广阔的机遇，这些都需要研究人员提出新的思想和方法。在这方面，传统的计算机科学算法曾有用武之地，但面对许多最具重要意义的序列分析问题，它们越来越显示出不足。这一方面是由于进化不断修补基因，导致生物系统内在的复杂性；另一方面则由于我们尚缺乏一套在分子水平上理解生命组织的完整理论。而机器学习方法［例如神经网络、隐马氏模型、支持向量机、置信网络（belief network）］正适合这类数据量大、含有噪声模式并且缺乏统一理论的领域。机器学习方法的基本思想是通过推理、模型匹配或样本学习，从数据中自动学习理论。因此，机器学习方法是传统方法的重要发展。本书旨在从机器学习的角度对生物信息学进行广泛全面的介绍。

机器学习方法的计算量极大，因此在很大程度上得益于不断提高的计算机处理速度。值得注意的是，自20世纪80年代晚期以来，计算机的处理速度和序列数据量几乎以相同的速度增长，即大约每16个月增长1倍。而最近，随着人类基因组工程第一个草图的完成，以及诸如DNA微阵列等高效实验技术的出现，生物信息数据以更快的速度增长，每6~8个月就增长1倍，从而给生物信息学带来了更大的压力。在初学者看来，机器学习方法好像是一些彼此无关的技术的集合，其实并非如此。在理论方面，一个关于所有机器学习方法的统一的理论体系在20世纪80年代晚期已经产生，这就是用于建

模和推断的贝叶斯概率体系。实际上，在我们看来，机器学习方法与贝叶斯统计建模和推断之间，除了前者更强调计算机技术和大规模数据处理之外，几乎没有差别。正是由于数据、计算机和概率理论体系三者的交汇，才使得机器学习方法在生物信息学和其他领域获得了强劲的发展动力，并且不断扩展。客观地讲，生物信息学和机器学习方法已经开始在生物学和医学领域产生显著的影响。

即便您对数学的严格性缺乏敏感，生物数据的概率建模仍然具有重大意义。这一方面由于生物测量经常包含难以去除的噪声，例如目前的DNA微阵列或质谱数据等。另一方面，序列数据因其离散性质及重复测序的成本较低，并不受噪声约束。因此，测量噪声并非采用概率建模的惟一原因。对生物数据进行概率建模的真正需要来源于生命系统的复杂性和多样性，这一切来自于漫长的进化进程中生物体在复杂环境下历经的进化修补。这样的生命系统必然呈现很高的维度（dimensionality）。即使在能够同时测量数以千计的基因表达的微阵列试验中，我们也仅仅观察到相关变量的一个很小的子集，而其他绝大部分变量则仍然处于隐藏状态，我们必须依赖概率建模来确定它们。直接应用系统化的概率体系能够加速发现变量的过程，避免重复历史上序列分析所走过的弯路。概率模型作为正确的理论体系正是从序列分析这个过去几十年中充满荆棘的领域中逐步发展而来的。

机器学习技术经常受到的批评是，它们都是“黑箱”方法：我们总是无法确定一个复杂的神经网络或隐马氏模型是如何达到特定解的。我们已经尝试在全面的概率体系中以及从实践的角度解决这一问题。然而，我们需要看到，许多当代分子生物学的技术是完全基于经验的。例如聚合酶链式反应（PCR），就其实用性和灵敏度而言，在某种程度上仍然是一项黑箱技术，实验中许多参数调整仍然是通过尝试得到的。另一个例子是关于序列在胶体矩阵中的运动方式和机动性，这里人们更关心实际成功和可用性，而很少关心对其中物理现象细节的理解。同样，对大部分药物来说，其药理作用的分子基础目前在很大程度上尚属未知。理论最终需要实践检验。至此，我们已经简要地概述了机器学习方法的功能及其优势。

读者及预备知识

本书面向不同背景的学生以及高级研究人员。我们试图为具备较强数学、统计学和计算机科学背景的读者提供生物学基本概念和问题的阐述。同样地，本书内容的选择也考虑到生物学家和生物化学家的需要，他们的生物学知识超出本书的内容，但在理解生物数据处理的一些新算法方面需要更多的帮助。为了使读者能够实现本书中所介绍的算法或将算法应用于特定的问题，本书在提供相当深入的内容的同时试图保持足够的简练性。然而，我们并未涉及有关大型数据库和测序项目的管理，以及原始荧光数据处理

等方面的内容。本书对预备知识的要求包括大学本科水平的微积分、代数和离散概率理论等。任何关于DNA、RNA和蛋白质方面的知识都是有帮助的，但不是必需的。

内容提要

我们试图使本书成为一本全面深入且简练易读的介绍性著作。书中包括主要概念的定义和主要定理，它们至少是概略性。更多的技术细节可以在附录和参考文献中找到。本书的大部分内容基于我们在过去几年中发表的论文，以及在ISMB（Intelligent Systems for Molecular Biology）大会等会议上的讲义，在丹麦理工大学（Technical University of Denmark）、加州大学欧文分校（University California Irvine）以及在NIPS（Neural Information Processing Systems）会议期间组织的讨论班讲授的有关课程。尤其是作为本书核心的广义贝叶斯概率理论体系，曾在1994年之后的几届ISMB大会上讲解过。

本书主要介绍生物信息学领域的相关方法，而不是阐述这个迅猛发展的学科的历史。当我们引用相关文献的细节时，只将注意力集中于介绍相关技术以及一些通用的一般性思考方法。同时，我们试图用一些实验结果来说明每种方法，其中一些结果直接来源于我们自己的工作。

第1章 本章介绍分子生物学中的序列数据和序列分析。其中包括基因组和蛋白质组的概述，进化所创造的DNA和蛋白质数据，这些数据正逐步进入的这个领域的公共数据库。本章还包括基因组及其规模的概述，以及一些在其他教科书中很难找到的相关资料。

第2章 本章旨在建立整个机器学习技术的理论基础，并且介绍了存在不确定性的情况下如何进行推理，因此本章是有关理论的最重要的一章。本章阐述了序列问题的一般性思想方法：用于归纳和推理的贝叶斯统计理论体系。这一体系的主要观点是，概率理论语言是适合于处理机器学习及所有建模问题的语言。所有的模型必须是基于概率的。在科学地描述模型及其与数据间的关系时，概率理论是惟一需要的工具，这一点在本书的书名中已有所体现。本章简要涵盖了一些经典论题，如：先验分布、似然度、贝叶斯定理、参数估计和模型比较。在贝叶斯体系中，人们最关心的是与数据、隐变量以及模型参数等相关的高维空间中的概率分布。为了处理和逼近这些概率分布，需要尽可能地利用独立性假设以便进行简单的因子分解。图模型正是基于这一思想，模型中变量之间的依赖关系对应于图的连通性。一些易于求解的常用模型往往对应于相对稀疏的图。本章对图模型和其他一

些处理高维分布的技巧只做了简略介绍，更深入的内容参阅附录C。应用概率理论和（稀疏的）图模型必然成为各种方法的两个真正核心的思想。

第3章 本章用一些例子进一步说明广义贝叶斯概率体系，为以后的学习做准备。这里介绍了几个经典例子的处理细节，随后的几章中将用到它们。熟悉这些例子的读者在快速浏览本书时可跳过本章。本章中所有的例子都基于投掷一个或多个骰子从而生成序列的思想。骰子模型只是一个极为简单的模型，然而本书的主要部分，从第7章到第12章，都可以视为这个模型的不同推广。统计力学也被视为骰子模型在贝叶斯概率体系中的一个精彩应用。此外，统计力学在机器学习的许多方面为我们提供了深刻的启示。尤其在第4章，统计力学被应用于一系列算法中，如蒙特卡罗方法（Monte Carlo）和期望最大（expectation maximization，EM）等算法。

第4章 本章简要介绍了许多应用于贝叶斯推断、机器学习和序列分析的基本算法，这些算法大多用于计算期望值和优化代价函数（cost function）。这些算法包括各种形式的动态规划、梯度下降法和EM算法，以及一些随机算法，如马尔可夫链—蒙特卡罗算法（Markov Chain Monte Carlo，MCMC）。MCMC算法的一些著名应用，如吉布斯采样（Gibbs sampling）、Metropolis算法、模拟退火算法（simulated annealing）等，在本章中都有所涉及。在初次阅读时可以跳过本章，尤其是熟悉算法或者对算法的实现细节不感兴趣的读者。

第5章 第5~9章和第12章构成了本书的核心部分。第5章主要介绍神经网络的理论。其中包括基本概念的定义，反向传播学习算法的简要推导，以及神经网络作为广义函数逼近器的简单证明。更重要的是介绍了如何从第2章建立的一般概率体系出发，更好地理解神经网络这个经常被视为与概率理论不相关的方法。接下来，这种思想将用来指导神经网络结构设计以及机器学习中代价函数的选择。

第6章 本章列举了一些精心选择的应用神经网络技术解决序列分析问题的例子。我们并不想涵盖迄今为止的数百个应用例子，而只选择了一些由于方法论上的进展而显著改善了应用效果的范例。我们尤其关注那些序列分析中机器学习过程优化的问题，以及如何组合网络以构成更加全面有效的算法。本章中具体分析的方法包括：蛋白质的二级结构、信号肽内含子剪接位点和基因发现。

第7章 第7~8章是关于隐马氏模型（HMM），其内容安排与第5~6章相似。其中第7章包括对隐马氏模型的详尽介绍，相关的动态规划算法（前/后向算法和

Viterbi算法）和学习算法（EM算法、梯度下降法等）。生物序列的隐马氏模型可以理解为包含插入和删除操作的骰子模型的推广。

第8章 本章包括精心选择的隐马氏模型在蛋白质和DNA/RNA序列问题上的应用范例。这些例子示范了隐马氏模型的主要应用，即蛋白质家族建模、生成大规模多重序列比对、序列分类，以及在大型数据库中搜索完整或破碎的序列片断。对于DNA序列问题，我们介绍了隐马氏模型如何用于基因发现（启动子、外显子和内含子）和基因结构分析（gene-parsing）等任务。

第9章 尽管隐马氏模型非常有效，但它仍然存在一些局限性。第9~11章的内容可以看做隐马氏模型在不同方向上的扩展。其中第9章系统应用概率图模型的理论作为统一的概念，并从中导出几类新的模型，例如：隐马氏模型和神经网络相结合的混合模型，能够利用序列空间特征而不仅仅是时间特征的双向马尔可夫模型。本章还包括基因发现、DNA对称性分析和蛋白质二级结构的预测等应用。

第10章 本章介绍了系统进化树（phylogenetic tree）并将其纳入第2章建立的概率理论体系，由此导出进化的概率模型。本章讨论的模型以及本书的其他模型均可视为第3章中简单骰子模型的推广。我们特别指出：在了解这些方法所近似的内在概率模型的情况下，那些经常从非概率意义的角度阐述的系统进化树重构方法［如吝啬法（parsimony method）］，实际上只是广义概率体系的一个特例。

第11章 包括正则文法（formal grammar）和乔姆斯基层次（Chomsky hierarchy）。随机文法（stochastic grammar）作为隐马氏模型和简单骰子模型的推广，为生物序列提供了一类新的模型。其中随机正则文法（stochastic regular grammar）本质上等价于隐马氏模型。而上下文无关随机文法（stochastic context-free grammar）则有更强的表达能力，它大致对应于能够产生1对字符（而不只是1个字符）的骰子模型。本章简要回顾了随机文法的应用，尤其在RNA建模方面的应用。

第12章 本章主要集中于DNA微阵列的基因表达数据分析，并再一次推广了骰子模型。我们介绍了如何系统应用贝叶斯概率体系对微阵列数据进行分析。我们特别考虑了基因在不同条件下表达水平是否发生变化和基因聚类问题。本章还简要讨论了基因调控区的分析和基因调控网络的推导问题。

第13章 本章包括当前因特网上有关数据库资源和其他公共资源的概述，以及一个包含许多重要网站的网址目录和链接。由于这些资源变化很快，因此我们主要介绍一些定期更新信息的网站。当然，本章也给出了一些包含

其他相关网站链接的定期更新的网页。

本书的附录包含几节技术性较强的讨论，它们是深入理解本书内容的重要参考。

附录A 包括误差带（error bar）、充分统计量以及指数型分布族等统计学概念。

附录B 主要包括信息论以及熵、互信息（mutual information）、相对熵（relative entropy）等一些基本概念。

附录C 简要概述图模型、独立性和马尔可夫性，其中既包括无向图模型（随机马尔可夫域），也包括有向图模型（贝叶斯网络）。

附录D 关于隐马氏模型的一些技术问题，包括数域缩放（scaling）、环状构架（loop architecture）和可弯曲性（bendability）。

附录E 简要概述了两类相关且日趋重要的机器学习模型：高斯过程和支持向量机。

本书还附有许多练习题，从一些简单的证明到一些定理的扩展方法都有。

为了阐述方便起见，我们有时隐含了一些关于正定性或可微性的标准假设，但读者可以从上下文中清楚地知道这些假设成立。

第2版增加和删去的内容

在书中不同部分，我们增加了一些新的内容或者从一个新的角度对于原有内容进行阐述。例如，第3章中关于最大熵的讨论和关于波耳兹曼-吉布斯（Boltzmann-Gibbs）分布的推导；第8章中将隐马氏模型应用于序列片断、启动子、亲水性分布图（hydropathy profile）、可弯曲性分布图（bendability profile）分析；第10章中从概率论的角度分析吝啬法和高阶进化模型；第12章中关于芯片数据的基因差异表达的贝叶斯分析。另外，我们还给出了从自由能的角度看待EM算法。这种提法不为人熟知，根据我们得到的材料，这种方法最早是由尼尔（Neal）和欣顿（Hinton）在他们未发表的技术报告中提出的。

在本书第2版出版的过程中，我们从许多同事、学生和读者那里得到了大力的帮助和有益的反馈。书中许多地方都有不同程度的修正和更新，以便反映全基因组测序和其他高通量技术所引发的科学发现的迅速发展。此外，我们还在第2版中做了如下一些重大改变：

- 第1章中新增了介绍人类基因组序列的部分。
- 第1章中增加了关于蛋白质功能和可变剪接的内容。
- 第6章中列出了神经网络的一些新应用。
- 完全改写了第9章，其主要内容改为图模型的系统阐述及其在生物信息学中的应

用。本章还特别包含了有关基因发现，递归神经网络用于蛋白质二级结构预测的新内容。

- 增加了新的一章（第12章），专门讨论DNA微阵列数据和基因表达。
- 增加了一节新的附录（附录E），讨论支持向量机和高斯过程。

本书的材料组织和一些问题讨论反映了作者的个人偏好。由于篇幅所限，省略了一些相关问题的讨论。关于贝叶斯推断和贝叶斯网络的分析，在理论水平上尚待提高。如果从统一的角度出发和更有利于对问题进行抽象，本书的大部分内容实际上完全可以只用贝叶斯网络的思想加以组织写作。我们关于系统进化树、DNA微阵列和基因聚类的生物学讨论，还可以进一步扩充。无论如何，在相关问题具有合适的补充材料时，我们列出了丰富的参考文献。

词汇和表示法

诸如“生物信息学”（bioinformatics）、“计算生物学”（computational biology）、“计算分子生物学”（Computational molecular biology）以及“生物分子信息学”（biomolecular informatics）等词汇用以表示本书所关注的研究领域。为了用词灵活起见，我们在书中对于这些词汇并不加以区分，实际上读者必须注意前两个概念的范围更广，还包含免疫系统和大脑的计算机建模等本书没有讨论的研究领域。最近，计算分子生物学还被赋予了一个完全不同的含义，类似于“DNA计算”（DNA computing），这是一个用于描述利用生物分子——而不是硅片——制造计算设备的概念。本书中我们在使用神经网络的概念时，有时会在前面加上“人工的”这个形容词。这里，我们仅从模式识别算法的角度讨论人工神经网络。

最后提一句，本书所使用的大部分符号列于书的结尾处。我们一般不系统区分标量、向量和矩阵。诸如“D”这样的符号用于表示数据，但不考虑数据的复杂程度。必要时，向量都视为列向量。黑体字符通常用于表示概率概念，诸如概率（$\mathbf{P}$）、期望（$\mathbf{E}$）和方差（$\mathbf{Var}$）。如果X表示一个随机变量，我们使用$\mathbf{P}(x)$代表$\mathbf{P}(X=x)$，在不产生歧义的时候，还直接为$\mathbf{P}(X)$。实际的概率分布可以记为P、Q、R等符号。

本书中，我们主要讨论离散概率分布的情况，读者也应该了解如何在必要时将结论推广到连续概率分布的情形。手写体符号用于表示特殊函数，诸如能量（$\mathcal{E}$）和熵（$\mathcal{H}$）。此外，我们还必须经常考虑用许多下标标识的变量。例如，神经网络中连接权重所依赖的其所连接的神经元i、j和所在的隐层l，在学习算法迭代中的时间t，等等。在特定的分析中，仅仅那些最具相关性的变量需要在下标中标识。在极少数不会引起歧义的地方，我们会使用相同的符号代表两种不同的意义。（例如，D也代表隐马氏模型中的删除状态。）

致 谢

多年以来，本书的工作得到了丹麦国家研究基金会和国家卫生研究院的支持。SmithKline Beecham公司对Net-ID基因片断项目的部分工作提供了赞助。本书的部分内容是皮埃尔·巴尔迪在加州理工学院生物学院时完成的。我们要向Sun Microsystems公司和加州大学欧文分校（UCI）基因组和生物信息学研究所提供的支持表示感谢。

我们要感谢所有对于手稿的早期版本提供反馈的人们，尤其是Jan Gorodkin，Henrik Nielsen，Anders Gorm Pedersen，Chris Workman，Lars Juhl Jensen，Jakob Hull Kristensen，David Ussery以及Net-ID项目的Yves Chauvin和Van Mittal-Henkle。此外还有生物序列分析中心的所有成员，他们多年来在许多方面对这项工作提供了设备上的支持。

我们还要感谢Chris Bishop，Richard Durbin和David Haussler，他们邀请我们到剑桥的伊萨克·牛顿学院，我们在那里完成了本书的第1版，我们还要感谢学院的良好环境和盛情邀请。要特别感谢Geeske de Witte，Johanne Keiding，Kristoffer Rapacki，Hans Henrik Stærfeldt和Peter Busk Laursen，他们的巨大帮助使我们将原先的手稿改变成现在这本书。

关于本书的第2版，我们要向UCI的新同事和新学生们致谢，他们是Pierre-Francois Baisnée，Lee Bardwell，Thomas Briese，Steven Hampson，G.Wesley Hatfield，Dennis Kibler，Brandon Gaut，Richard Lathrop，Ian Lipkin，Anthony Long，Larry Marsh，Galvin McLaughlin，James Nowick，Michael Pazzani，Gianluca Pollastri，Suzanne Sandmeyer，Padhraic Smyth。我们还要向以下UCI以外的人员表示感谢，他们是Russ Altman，Mark Borodovsky，Mario Blaum，Doug Brutlag，Chris Burge，Rita Casadio，Piero Fariselli，Paolo Frasconi，Larry Hunter，Emeran Mayer，Ron Meir，Burkhard Rost，Pierre Rouze，Giovanni Soda，Gary Stormo，Gill Williamson。

我们还要向本丛书的编辑 Thomas Dietterich 以及MIT出版社的工作人员表示感谢，尤其是Deborah Cantor-Adams，Ann Rae Jonas，Yasuyo Iguchi，Ori Kometani，Katherine Innis，Robert Prior。还有Harry Stanton，他在我们开始写作时提供了许多帮助。最后，我们要感谢所有的朋友以及我们的家庭所给予的帮助和支持。

本书介绍了机器学习方法的主要内容及其在生物学数据处理中的应用。其中对机器学习技术的理论基础——贝叶斯概率体系进行了详细介绍，并在此基础上着重对神经网络、隐马氏模型以及概率图模型等方法在生物信息学中的应用作了详细分析。书中特别列出一章介绍了DNA微阵列和基因表达，以及相关数据的分析方法。

本书主要针对两个读者群体。一是生物学和生物化学研究人员，他们想了解基于数据处理的算法；二是物理、数学、统计、计算机科学等领城的学者，他们想知道机器学习方法在分子生物学研究中的应用。