北京大学生命科学译丛-1

生物信息学概论

T K Attwood & D J Parry-Smith 著

罗静初 等 译

北京大学出版社
北 京

图书在版编目(CIP)数据　图字：01-2001-0055

生物信息学概论/(英)谈莲莎(Attwood, T. K.), (英)史密斯(Parry-Smith, D.J.)著；罗静初等译. - 北京：北京大学出版社，2002.4

(北京大学生命科学译丛；1)

ISBN 7-301-05468-8

Ⅰ. 生…　Ⅱ. ①谈…②史…③罗…　Ⅲ. 生物信息学　Ⅳ. Q 811.4

中国版本图书馆 CIP 数据核字（2002）第 014224 号

书　　名：生物信息学概论

著作责任者：[英] T K Attwood & D J Parry-Smith 著

译作责任者：罗静初 等

责 任 编 辑：赵学范

标 准 书 号：ISBN 7-301-05468-8/Q·0089

出　版　者：北京大学出版社

地　　址：北京市海淀区中关村北京大学校内　100871

网　　址：http://cbs.pku.edu.cn

电　　话：出版部 62754962　发行部 62754140　编辑部 62752021

电 子 信 箱：zpup@pup.pku.edu.cn

排　版　者：兴盛达打字服务社　62549189

印　刷　者：北京大学印刷厂

发　行　者：北京大学出版社

经　销　者：新华书店

850 毫米×1168 毫米　32 开本　9.75 印张　280 千字

2002 年 4 月第 1 版　2002 年 4 月第 1 次印刷

定　　价：16.00 元

译　　序

众所周知，20 世纪 90 年代诞生的因特网及其随之而来的信息高速公路计划，意味着信息时代的到来。几乎与因特网同时诞生的人类基因组计划，是迄今为止耗资最为巨大、竞争最为激烈、意义最为深远的大型国际合作研究项目。完成人类基因组 30 亿个碱基对的全序列测定，只是人类基因组计划的第一步。搞清楚人类基因组全套遗传密码的全部含义，则是这一计划的目标。由此引发的结构基因组、功能基因组、药物基因组、蛋白组计划已经开始实施。DNA 序列测定开拓了一个测序仪制造和测序试剂生产的高新技术产业，以基因芯片技术为代表的一系列生物技术则是功能基因组、药物基因组研究和开发的基础。

基因组计划和因特网推动了生物信息学的诞生。它以核酸、蛋白质等生物大分子数据为主要对象，以数学、信息学、计算机科学为主要手段，以计算机硬件、软件和计算机网络为主要工具，对浩如烟海的原始数据进行存储、管理、注释、加工，使之成为具有明确生物意义的生物信息。并通过对生物信息的查询、搜索、比较、分析，从中获取基因编码、基因调控、核酸和蛋白质结构功能及其相互关系等理性知识。在大量信息和知识的基础上，探索生命起源、生物进化以及细胞、器官和个体的发生、发育、病变、衰亡等生命科学中重大问题，搞清它们的基本规律和时空联系，建立“生物学周期表”。

国内计算机网络的开通，特别是北京、上海等大城市网络基础设施的发展，为在我国实现生物信息共享提供了必要条件。我国人类基因组、水稻基因组等大型研究计划以及国家 863 计划、973 计划、攻关项目、自然科学基金项目的实施，对生物信息在生命科学和生物

技术领域的研究、开发和应用提出了迫切要求。国内一大批单位已经开始从事生物信息学研究。不少大专院校已经和正在准备开设生物信息学课程。为此，我们组织翻译了这本“生物信息学概论”，为国内读者，特别是生命科学和生物技术领域从事分子生物学和基因组研究的师生，提供一本入门教材和参考书。

需要特别说明的是，由于生物信息学和计算机网络正在以日新月异的速度发展，本书原作中有些素材已经过时，不少网址有所变更，无法一一注明，请读者见谅。此外，原书中加框的双色图（灰度及淡蓝色），译著中标以“图框”并排仿宋体字以示区分。限于译者水平，对原书的理解肯定有不少欠妥之处，译文中难免有不少错误，敬请读者批评指正。如能将错误疏漏之处用电子邮件方式发给译者（luojc@pku.edu.cn），将不胜感激。

本书初稿由以下译者执笔：罗洪：前言、第 1 章；曲红：第 2～3 章；方刚：第 4～5 章；李兵、刘翟：第 6～7 章；禺胄：第 8 章；黄弋：第 9～10 章；陈蕴佳：图表注释和校对；吴昕：词汇表。全书由罗静初统一修改、整理和审定。

本书翻译过程中得到作者 Teresa Attwood 和 David Parry-Smith 的帮助，特此致谢。郑伟谋、官山和北京大学选修“生物信息学概论”课的许多研究生在成稿过程中提出了许多宝贵意见，一并表示感谢。感谢北京大学出版社赵学范编审为本书出版所作的努力。

罗静初

2001 年 2 月于北京大学燕北园

致　谢

首先，我们谨向下表中列出的网站和有关作者深表谢意，感谢他们为本书提供了许多精美图表。

编号	出　处	作　者
图 1.2	http://www.biochem.ucl.ac.uk/bsm/pdbsum/	R. Laskowski
图 2.3	http://www.ebi.ac.uk/	R. Lopez
图 2.4	http://www.ebi.ac.uk/	T. Etzold
图 2.5	http://www.vtourist.com/webmap/maps.htm	B. Plewe
	http://www.angis.org.au/	T. Littlejohn
图 3.3	http://www.expasy.ch/	A. Bairoch
图 3.4	http://www.expasy.ch/	A. Bairoch
图 3.6	http://www.blocks.fhcrc.org/	J. Henikoff
图 3.7	http://www.blocks.fhcrc.org/	J. Henikoff
图 3.8	http://www.expasy.ch/	A. Bairoch
图 3.9	http://www.sanger.ac.uk/	R. Durbin
图 4.3	http://www.sanger.ac.uk/	R. Durbin
图 7.2	http://www.sanger.ac.uk/	R. Durbin
图 8.3	http://www.bioinf.man.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图 8.4	http://www.blocks.fhcrc.org/	J. Henikoff
图 8.5	http://www.expasy.ch/	A. Bairoch
图 9.10	http://www.bioinf.man.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图 9.11	http://www.bioinf.man.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图 9.12	http://www.blocks.fhcrc.org/	J. Henikoff
图 9.13	http://www.blocks.fhcrc.org/	J. Henikoff
图 9.14	http://www.blocks.fhcrc.org/	J. Henikoff
图 9.15	http://www.stanford.edu/identify/	C. Nevill-Mannine
图 9.17	http://www.biochem.ucl.ac.uk/bsm/cath/	C. Orengo
图 9.18	http://www.biochem.ucl.ac.uk/bsm/pdbsum/	R. Laskowski
图 9.19	http://www.biochem.ucl.ac.uk/bsm/pdbsum/	R. Laskowski
图 9.7	http://www.expasy.ch/	A. Bairoch
图 9.9	http://www.sanger.ac.uk/	R. Durbin
图 10.2	http://www.mrclmb.cam.ac.uk:80/pubseq/	R. Staden
图框 1.3	Rube Goldberg 公司和 United Media 公司	
图框 8.1	http://www.biochem.ucl.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图框 9.1	http://www.blocks.fhcrc.org/	J. Henikoff

感谢 Anne Parry-Smith 女士始终如一的支持，在写作本书的艰苦日子里，她对我们嘘寒问暖，关心备至。感谢英国剑桥 Drug Discovery 研究所 Jeremy Packer 为本书所做的十分细致的校对工作。感谢本书主要策划者 Addison Wesley Longman 图书公司的 Alex Seabrook 先生和编辑 Kate Henderson 女士。他们的耐心十分可贵，堪称楷模。最后，作者 Attwood 还要感谢伦敦大学 University College 生物化学系序列分析小组的 Phil Scordis、Julian Selly、Jane Mabey、Will Wright 和 Maria Karmirantzou。由于本书的写作，Attwood 博士无法顾及他们的研究课题，有时甚至因为写作不顺利而向他们发脾气，而他们却以微笑表示理解。每当在写作中遇到各种困难时，他们总是雪中送炭，帮助解决那些在关键时刻故意作弄人的难题。

真诚感谢所有为本书出版尽心尽力的同事、亲友和学生们！

原　序

过去的十年，对生命科学来说，是一个不同寻常的十年。“基于硅片的生物学”已经在向我们招手。它的出现，为我们开拓了生物学研究的新途径，使我们有可能对全基因组进行系统的研究和全面的比较。旨在阐明生物遗传密码和基因产物的基因组研究是当前生命科学的热点，其最终目的是要揭示生物进化的机制，搞清蛋白质折叠的机理，探索蛋白质结构功能关系等生命科学中一系列重大问题。

迄今为止，利用计算机模拟生物过程依然具有很大的局限性。究其根源，主要在于我们对生命科学中重大问题的了解仍然十分有限。必须承认，蛋白质折叠的机制尚未得到完全阐明；对某个特定的序列或折叠方式，究竟源于趋同进化还是趋异进化，尚难给出明确的结论；仅由单个蛋白质序列或结构，尚难孤立地推断其功能。只有真正懂得计算机并非魔术师手中的魔杖，才能更好地利用计算机，解决那些可以解决的问题。不懂得这一点，就会被某些计算机程序或软件所导致的错误结果引入歧途。

当今生物信息学领域的主要研究方向之一，是从研究蛋白质功能入手，进而阐明进化关系。常用方法可以分为两类：一类以序列为基础，用序列分析的手段以及所得结果推断生物大分子的功能；另一类则以结构为基础，其基本思路基于蛋白质分子的结构功能关系。第一类方法的基本出发点，是通过和数据库中已知功能的序列进行相似性比对，确定那些新测定的序列与已知序列之间的关系。也就是说，通过数据库搜索，找出可能只有若干相同残基的功能位点，由某个初看起来完全不同的蛋白质分子确定该未知蛋白的功能。此时，未知蛋白和已知蛋白序列整体相似性并不很高，之所以有相同功

能,很可能是趋同进化的结果。而那些序列、结构相似性较高的蛋白质分子,则可能是同一祖先趋异进化的结果。

应该肯定,利用计算机进行核酸和蛋白质序列分析的方法已经比较成熟。那么,由蛋白质序列识别其功能,似乎应该顺理成章。然而,实际情况并非如此乐观。由于生物学实验远远落后于基因组测序的进程,对于大量的序列数据无法及时用实验方法验证。截至1998年,所有已经测定的基因组序列中,约三分之一蛋白质序列的功能无法推断。序列数据库中很大一部分序列除了基因名称外,没有其他注释信息,或仅仅标上"假想蛋白质"(hypothetical proteins)。此外,有些具有相同结构的蛋白质,却具有不同的功能。典型的例子是溶菌酶(lysozyme)和 α-乳清蛋白(α-lactalbumin)。这两种蛋白质具有高达50%的相同残基,序列相似性分数达70%,其功能却完全不同。前者是催化分解细菌细胞壁多糖的水解酶,后者不具任何催化作用,而是一个起调控作用的蛋白质,其作用是将半乳糖(galactose)分子从 UDP-半乳糖(UDP-galactose)转化为 D-葡萄糖(D-glucose)。为什么两者的序列相似性程度如此之高,折叠方式又十分相似,它们的功能却大相径庭呢?这是因为,溶菌酶中起关键作用的两个催化残基谷氨酸(glutamic acid)和门冬氨酸(aspatic acid)在 α-乳清蛋白中却没有。相反,α-乳清蛋白中起关键作用的酸性钙结合序列模体只在个别溶菌酶分子中存在。这一实例并不意味着利用计算机进行序列分析和结构模拟毫无用处,而是强调这种分析和模拟必须与基本的生物学知识相结合,才能得出较为可靠的结论。

不言而喻,蛋白质分子的序列和结构是决定其功能的基础,两者相辅相成,缺一不可。如果用简单的数学公式表示,不妨写成:"功能 = 序列 + 结构"。可以将蛋白质的各种折叠单元看做各种不同的基本部件,用来整合各种不同序列模体,以实现各种不同功能。可以用房子作比喻,形象地说明这个问题。把一间空房比做基本单位,就像蛋白质的某种折叠单元。若在这间房子里放入一张桌子和一把椅

子，还很难猜测这间房子的用途，最多只能说它不大可能是浴室。如果在桌子上放上一台计算机，就可以推测这间房子可能是工作室而不是餐厅。当然，这只是问题的一个方面。这里并没有考虑这间房子所处的环境。如果它是家庭里的一个书房，那么这台计算机可能用来管理家庭账务；如果是公司或学校的计算机机房，那么这台计算机就可能用来进行数据库管理。显然，只有掌握基本单元的主要属性及所处环境，才有可能对其用途做出推断；只有对房间中的家具和计算机等其他附件以及房间所处位置有所了解，才能推测其功能。

不妨用类比的方法对蛋白质分子的序列、结构和功能之间的关系加以分析。β-折叠桶是蛋白质分子中常见的一种折叠方式，就像上面提到的房子一样，随处可见。β-折叠桶的功能多种多样，取决于它所包含的序列。显然，像 β-折叠桶这样的折叠单元是较为稳定的蛋白质结构，这种稳定结构可能来自于不止一种蛋白质，而是不同种类的蛋白质趋同进化的结果。而上面提到的溶菌酶和 β-乳清蛋白，则是趋异进化的典型实例。这两种蛋白质的序列和结构都很相似，就像房间和家具，初看起来大同小异。此时，不同蛋白质的不同功能仅取决于某几个特殊氨基酸残基，就好像同是一台计算机，家庭书房中的计算机可能是一台微机，而公司或学校计算机机房中的那台计算机，则可能是功能更为强大的服务器。

显然，对于自然界中错综复杂的生命现象，我们的认识还相当肤浅。用现有的计算机方法，尚不能轻而易举地揭开其中的奥秘。迄今为止，计算机程序模拟不能取代生物学实验，而只能提供生物系统的某种模型。判断不同序列或不同结构是否相似，判断它们源于趋同进化还是趋异进化，判断相似序列或相似折叠方式是否具有相同功能，即使对一个经验丰富的生物学家来说，也并非易事。没有一蹴而就的捷径，也没有能给出现成答案的计算机程序。计算机为我们提供的，仅仅是大量的数据；计算机能够做到的，仅仅是缩小选择答案的范围。而只有计算机的使用者，也就是我们自己，才能从序列分

析和结构模拟中得出具有生物学意义的结论。

综上所述，序列分析和结构模拟是生物信息学研究的两个主要组成部分，两者都包括许多分支，包含的内容很多，涉及的范围很广。假如能从序列和结构两个方面同时入手研究生物大分子的功能，显然是最为理想的途径。遗憾的是，与大量的序列数据相比，可用的结构信息依然十分有限。我们知道，蛋白质的三维结构远比序列保守，不同的序列可能形成相似的折叠模式。这对序列分析方法是一个很大的挑战。面临浩如烟海的序列数据，这一挑战将有增无减。本书旨在介绍用生物信息学方法和工具，从大量错综复杂的序列数据中，探索序列、结构、功能之间的关系，并指出这些方法的可能性和局限性。

绪　论

打开这本《生物信息学概论》，读者一定急于了解：此书包括哪些内容？它们是如何安排的？这里将回答这两个问题，以便读者对本书全貌有个基本了解。

首先必须说明，本书不是关于蛋白质结构的教科书，不打算介绍蛋白质结构分析，不涉及蛋白质二级结构或三级结构预测。关于蛋白质结构分析和结构预测的书籍，读者可以找到不少很好的权威著作。其次，本书也不是一本生物学教科书，不打算介绍生物学的基本知识。尽管如此，本书中介绍的许多分析结果必须结合生物学背景并最终用生物学实验验证。这是本书的基本出发点，强调生物学实验重要性的思想贯穿于全书。

本书主要介绍核酸和蛋白质序列的计算机分析方法，试图探讨利用现有的计算机程序，从现有的数据库中能够获取什么，而不能够获取什么。必须指出，序列分析并不能给出结构、功能或进化关系的最终结论。计算机程序只能提供一些线索，而我们的任务则是提出好的分析策略，以便最有效地获取生物学知识，进而为生物学实验提出具有参考价值的建议。如果应用得当，序列分析可以成为现代分子生物学研究的有用工具。

基因组计划的实施及伴随而来的序列数据激增，使序列分析成了计算机在生物学中应用的热点。研究新的计算机方法，从序列数据提取有用的生物信息，已经成了当务之急。通过序列比较，确定新测定序列与数据库中已知结构和功能的序列间的相似性关系，从而以足够的可信度确定新序列的结构和功能信息，是本书将要讨论的分析策略之基本出发点。序列之间的相似性关系并不都十分明显，

有时仅在某一很小范围内才能略见端倪。从浩如烟海的生物数据中发现进化关系，就好像从大量噪声中提取有用信号，是一个很大的难题。为了解决这一问题，近年来人们提出了各种不同方法，试图从不同角度探索数据挖掘的新方法。面对各种方法，初学者往往会不知所措。本书并不打算对这些方法逐一介绍，而是为大家提供目前最常用的数据库和数据库搜索工具的指南。因此，希望读者不要把注意力集中在某个特定的数据库或某个特定的分析工具上，而应该着重学习如何运用各种不同方法对各种不同数据库进行全面的搜索，并对各种不同分析策略做出全面的评价。

之所以反复强调这一点，完全是根据当前的实际情况。迄今为止，没有一个数据库能够包罗万象。仅靠一个特定的数据库，无法实现序列分析任务。某些看起来很相似的数据库，其内容实际上并不完全一样。仅仅利用一个数据库的信息而忽略其他数据库，往往会使分析结果大打折扣。另外，也没有一个绝对可靠的数据库搜索和序列比对算法。利用某一方法所得结果，应该用另一方法来检验其可靠性。有些数据库和模式识别方法还处在研究阶段，尚未经过实际应用的检验，其实用性尚待推敲。由于研究经费等种种问题，有些尚在初始阶段的研究项目不得不半途而废。同时，在飞速发展的生物信息学领域中，新方法、新软件、新程序层出不穷，常常令人目不暇接，不知所措。本书所强调的重点，是生物信息学的基本概念，特别是序列分析的基本方法。

不言而喻，序列分析必须利用大量的数据库和各种分析方法。本书从第 2 章开始，逐步介绍各种常用数据库，介绍各种丰富的生物信息资源和各种分析方法。书中不厌其烦地讲述各种不同类型数据库的内容、格式、特点，以及数据库搜索和序列分析的方法，以便读者在实际序列分析中，对不同类型的数据库和各种分析工具了如指掌，使用时得心应手。在熟悉了基本数据库、掌握了基本分析方法后，用一个基于 Web 的交互式数据库搜索实例，详细介绍如何利用各种数

据库和分析工具，由始至终、循序渐进，对一个实际序列进行分析，并得到一定的结果。从这个意义上说，本书可以看做一本详细的生物信息学网络教程。

图 i 表示本书基本内容和结构。图中给出全书各章节、各部分间的关系。从图中可以看出，本书编排上的一个主要特点，即在介绍了有关数据库和序列比对基本概念的基础上，在第 9 章中给出一个实例，说明如何设计一个数据库搜索和序列分析方案。

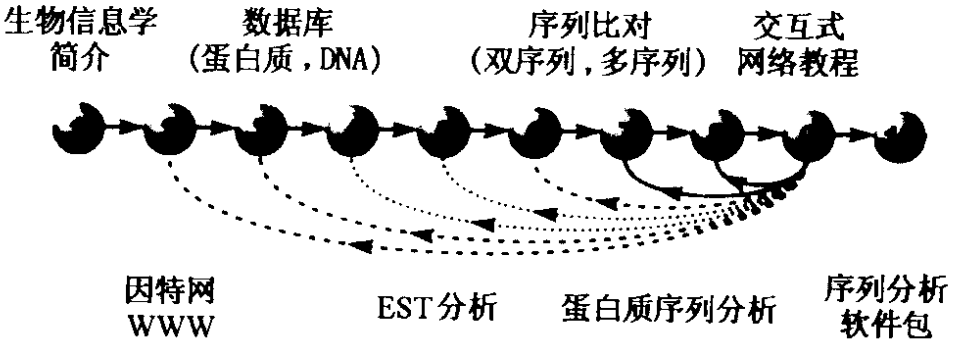

图 i　本书概貌，指出各章节之间的关系及阅读顺序

尽管第 2～8 中的内容在第 9 章数据库搜索实例中均会用到，但相关程度有所不同。图中从第 9 章出发指向第 2～8 章的弧形线条，有的用实线(第 7～8 章)，表示与第 9 章的关系最为密切；有的用点线(第 4～5 章)，表示与第 9 章的关系较密切；而指向第 2、3、6 章的是虚线，则表示与第 9 章联系的密切程度不如其他各章。

为便于读者对全书有一个基本了解，下面概要说明每一章的主要内容。本书大体包括 4 个主要部分：第 1 部分介绍核酸和蛋白质数据库的内容和格式；第 2 部分介绍双序列比对和多序列比对的方法以及它们的局限性；第 3 部分介绍序列模式及其识别方法，包括单位点识别、多位点识别和序列谱分析等，并指出这些识别方法的局限性；最后说明如何把以上方法结合起来，设计一个有效的搜索和分析方案。

1. 概论

第 1 章是全书概论，提出并回答什么是生物信息学、生物信息学的重要意义等问题。为了更好地介绍这些基本概念，本章简要回顾生物信息学发展的历史，说明从耗时费力的蛋白质序列测定到信息技术所带来的革命性进步。仅仅在十多年前，测定一个蛋白质全序列需要用一年甚至更多的时间，而今天，由核酸序列翻译所得的蛋白质序列正以惊人速度递交到蛋白质序列数据库中。序列数据的爆炸性增长，使结构数据在数量上无法与其相比。寻找仅从序列数据本身提取功能信息的有效方法，已经势在必行。本章还简要介绍蛋白质二级结构和三级结构预测的进展和局限性，介绍同源性、相似性等基本概念。

2. 信息网络

第 2 章介绍因特网、万维网，介绍国际生物信息资源中心和服务机构，介绍用于查询和搜索世界各地生物信息数据库的网络浏览器和计算机软件，并给出一些重要生物信息机构的网址。

3. 蛋白质信息资源

第 3 章主要介绍一些重要的蛋白质数据库，详细讨论一级结构、二级结构和三级结构等不同层次的蛋白质数据库以及它们的格式，如 SWISS-PROT、PROSITE 等。同时说明构建复合数据库和二次数据库的意义。

4. 基因组信息资源

第 4 章介绍核酸序列数据库，包括世界上三大核酸序列数据库 GenBank、EMBL 和 DDBJ，介绍某些基因组信息资源。限于篇幅，只介绍可以通过因特网查询的一些常用数据库，并提供因特网上众多的基因组信息资源网址。作为第 3 章中介绍的蛋白质数据库格式的对照，本章介绍核酸序列数据库 GenBank 的格式。

5. DNA 序列分析

第 5 章介绍 DNA 序列分析的目的意义和基本方法，介绍基因

组信息的层次，介绍由 cDNA 库快速测序得到的表达序列标签 EST。由于 EST 在基因和药物研究中的重要性，本章将重点介绍 EST 分析方法，讨论 EST 序列分析中的一些特殊问题，给出一个实例和三种常用方法。

6. 双序列比对

双序列比对是序列分析中的基本方法。双序列比对为数据库搜索提供了识别相似序列的基础，并由此推测它们属于同源序列的可能性。本章将说明分析序列比对结果时常用的同一性和相似性概念，给出局部相似和全局相似的定义。

7. 多序列比对

找出一对序列之间的关系只是序列分析的第一步。我们的兴趣往往集中在构成基因家族的一组序列上。为了确定这一基因家族的保守性特征，必须追溯各个序列之间的关系。多序列比对可以有效地提高序列比对的信噪比，并最终找出具有显著生物学意义的序列片段，它们可能与结构和功能有关。因此，本章对各种多序列比对方法进行了概括，包括完全用计算机程序实现的自动比对方法和基于交互式编辑的手工比对方法。

8. 二次数据库搜索

在对双序列比对和多序列比对有了一个基本了解后，本章介绍二次数据库搜索方法。二次数据库的格式在第 3 章中介绍，本章集中讨论存储于二次数据库中特殊信息模式的搜索方法，如正则表达式、序列谱、序列指纹图谱、序列模块和隐马氏模型等。二次数据库搜索的基本策略是从不同角度进行多序列比对，以确定它们是否属于同一个蛋白质家族。不同方法所得结果往往有一定区别，本章分别指出各种搜索方法的长处和短处。各种二次数据库之间的内容并不完全涵盖，不同的模式识别方法均有其局限性，完善的搜索策略应该包括所有二次数据库搜索的方法。

9. 数据库搜索实例

在介绍了一次和二次数据库搜索的基本概念后，可着手构建数据库搜索的具体方案。本章从构建一个数据库搜索方案实例出发，介绍如何通过万维网进行实际操作，用交互方式说明数据库搜索的实际步骤。各种不同的程序所得结果往往以不同格式输出，有的不很直观。本章重点说明如何分析搜索结果，确定所得结果是否具有生物学意义，剔除那些假阳性结果。根据上述指导思想，本章给出一个实际的网络教程，说明如何搜索一级结构、二级结构以及结构分类数据库，确定一段未知 DNA 序列片段的功能。这一实例给出了数据库搜索的具体步骤，辅以详尽的文字说明、图表、流程图等有关信息。需要特别强调的是，本章只是提出数据库搜索的基本原则，以帮助读者设计自己的数据库搜索方案。

10. 序列分析软件包

在介绍了大量的数据库、数据库搜索方法、序列分析基本策略以及数据库搜索网络教程后，本章介绍一些常用序列分析软件包，包括 GCG、Staden，以及最近几年发展起来的基于 Web 的序列分析软件 CINEMA 等，说明这些软件的基本功能，它们的开发过程和发展趋势。限于篇幅，本书只能对当前常用软件的主要特点作简要说明，而不可能对它们作详细介绍。读者可通过因特网对它们作进一步了解，包括这些软件的使用许可协议等。

生物信息学是一门新兴学科，初学者对有些专用词汇术语尚不很熟悉。为此，书末词汇表列出一些常用术语、短语和缩略词，并给出了简单解释，便于读者查询。

阅读本书时，希望读者能够注意以下几点：

第一，不要轻信数据库，它们所提供的信息有时可能会有误差，甚至完全错误。据估计，目前核酸序列数据库的错误率约占 0.1%～4.0%，由此翻译而得到的蛋白质序列数据库中氨基酸序列的错误率则可达 5%。至于序列数据库的注释部分，很难用定量标准估计

其错误率。有的科学家认为，目前所用的数据库自动注释程序，会使原有的错误进一步扩大，甚至可能造成灾难性的后果。

第二，不要轻信计算机程序，它们给出的结果有时可能会造成误导，甚至使我们得出完全错误的结论。这样的实例并不少见。

第三，不要轻信万维网，它所提供的信息也可能会有误导。即使是最著名的国际生物信息中心的网页制作者，也可能会犯这样那样的错误。这一点在编写本书的过程中已经发现。

第四，不要轻信书本，即使是公开发表的文献，也常常会有这样那样的错误。本书引用的某些文献，也难免有一些错误。

总之，不要做一个幼稚的计算机用户，而要善于思索、勇于提问，要对所获得的信息提出自己的见解。尽可能地掌握全貌，而不是被那些看起来似乎很有价值的凤毛麟角所迷惑。只有当一些线索从杂乱无章的背景中浮现出来，并开始形成合乎逻辑的结果时，你的分析才开始步入正轨，你得到的结论才会比较可靠。