Bioinformatics

生物信息学：

序列与基因组分析

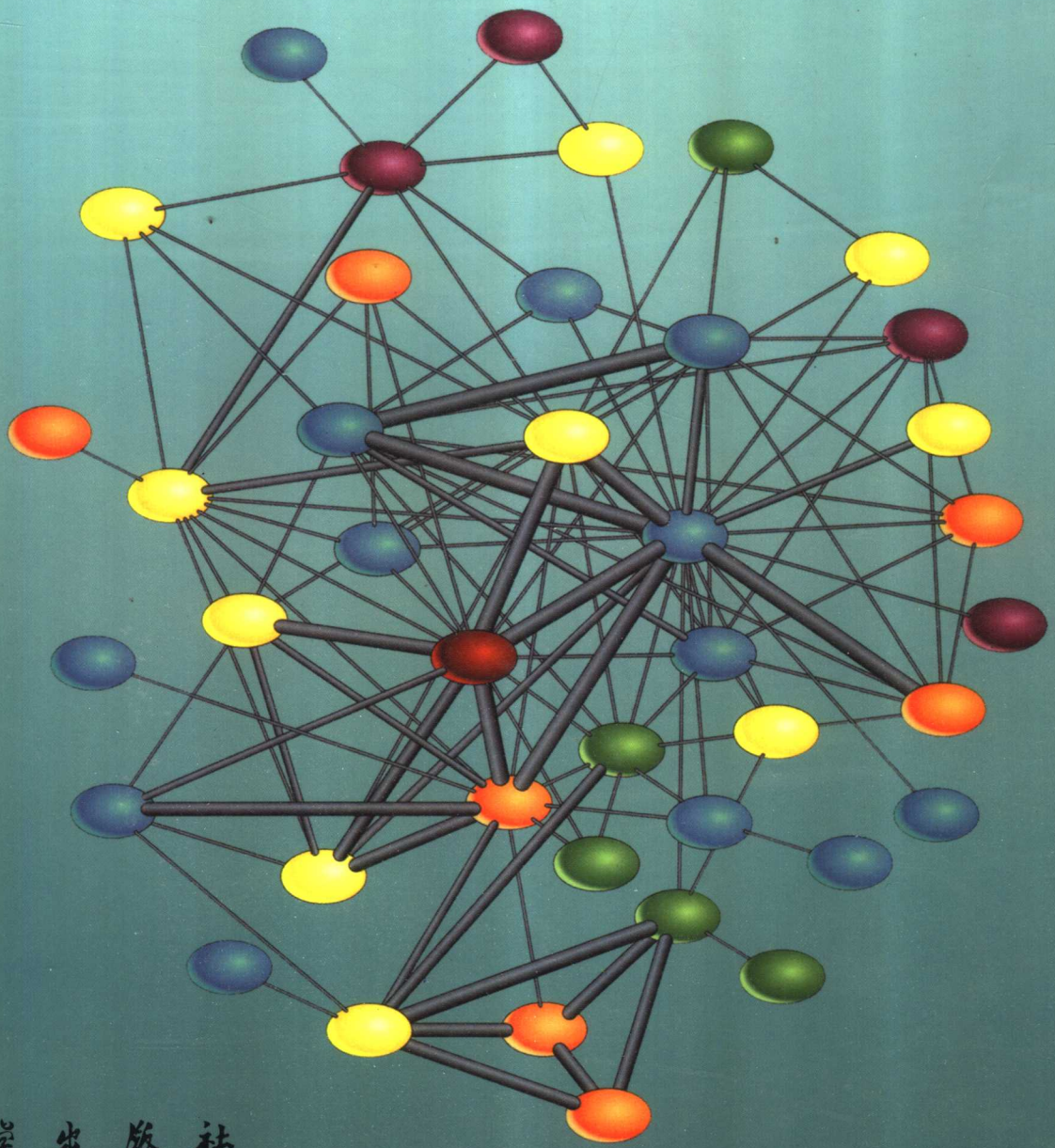
Sequence and Genome Analysis

第二版

Second Edition

David W.Mount

曹志伟 编译



科学出版社

www.sciencep.com

COLD SPRING HARBOR LABORATORY PRESS

生 物 信 息 学

序列与基因组分析

（第二版）

Bioinformatics

Sequence and Genome Analysis

(Second Edition)

David W. Mount

曹志伟 编译

科学出版社

北 京

图字:01-2005-3665

内 容 简 介

当前生物信息学研究重点是对基因组序列、蛋白质组学和数组技术所产生的大量数据的计算分析。本书对 DNA、RNA 和蛋白质数据的计算提供了丰富的演算方法,并指出了在解决生物学问题中这些方法的优缺点及应用策略。

本书的第一版是在 Mount 博士讲稿的基础上进行整理出版的,在全球范围内用作教材。第二版对内容进行了全面的修订,由专业教师提供导读,最大程度地适用本科生和研究生教学。

本书为高等院校生物信息学专业本科生和研究生提供理想的学习材料。同时,本书也适宜科研人员、信息专家自学使用。

Bioinformatics: Sequence and Genome Analysis
Second Edition
David W. Mount
ISBN: 0-87969-712-1

图书在版编目(CIP)数据

生物信息学/(美)芒特(Mount, D. W.)著;曹志伟编译. —2 版. —北京:科学出版社,2006
(国外生命科学优秀教材)
ISBN 7-03-017640-5

Ⅰ. 生… Ⅱ. ①芒…②曹… Ⅲ. 生物信息论-教材 Ⅳ. Q811.4

中国版本图书馆 CIP 数据核字(2006)第 078363 号

责任编辑:甄文全 周 辉
责任印制:张克忠/封面设计:卢秋红

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

*

2006 年 10 月第 二 版 开本:890×1240 1/16
2006 年 10 月第一次印刷 印张:37
印数:1—3 500 字数:1 102 000

定价:75.00 元 (含光盘)

(如有印装质量问题,我社负责调换〈路通〉)

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。

D. W. M.

2004 年 5 月

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 clastalW 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)

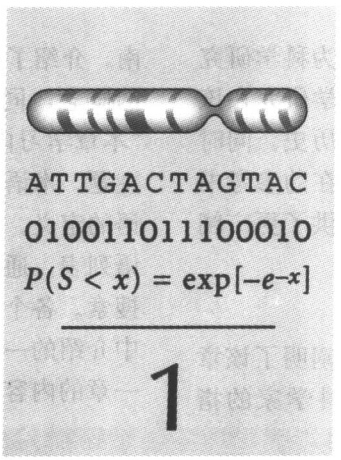
此学科令我感受深刻的一个特点就是大多数研究者，尤其是本领域的前沿学者都愿意与同事分享他们的研究成果。我有幸结识了几位这样的先驱，尤其是 David Lipman，Hugo Martinez（我与其共度了学术休假年）和 Temple Smith。这些学者取得的巨大成就因为愿意和同行免费共享他们努力的成果而更加备受称赞。正是因为这样，他们为学术界和商业界在序列分析领域的成功起到了非常重要的作用。

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，Foger 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

目 录



历史简介和概论

目录

简　介

本章介绍了生物信息学是如何演变为科学研究的一个新领域，描述了生物学和计算科学研究在该领域所扮演的角色，并简要记述了发展历史。同时还提供第二版各章节的概述。早期和现在的参考书籍、文章、综述以及杂志也为该领域提供了更广阔的视角。

本书各章节结构

每一章节以介绍段落开篇，简要地阐明了该章节的目的，接着是分别给生物和计算科学家的指南，介绍了在他们各自的研究领域中可能并不熟悉的概念，同时也提供该章节主题的介绍性指导。“本章学习内容”列出章节中重要的概念和实际的主题。术语表包括各章节中介绍部分以外的主要术语的定义。在各个章节最后的“本章检索词”是术语列表，通过它们能够对本章引用的网址进行引擎搜索。各个章节还总结了习题集，用于探究该章节中介绍的一些概念和步骤。因为是介绍和概述，第一章的内容比其他各章更总括，也没有习题集。

生物学家指南

在这章中，我们描述了实验室中获得的 DNA 序列在计算机文件中通常如何保存，这些文件和普通的文本文件非常相似。序列文件包括序列的信息，如生物体、实验室来源、文献引用、序列名（如果是已知基因），一个或多个唯一数字，这是特定序列的主记录号或其他标识号。由于序列的数量非常大，序列文件就能够组成数据库，如 GenBank 数据库，经过索引，它可以基于文件序列信息非常容易地获取特定的序列；比如获取特定生物体的所有序列。数据库的格式，称为关系数据库，由相互索引的或主键表格组成。我们使用数据管理系统构建数据库，可以将数据存储在里面，也可以从中提取数据。虽然序列是文本格式保存，常用的商业文本编辑器却不能检查或处理这些文档，因为这些编辑器在文本中引入了一些控制符反而破坏了序列文件。在网络和本地计算机系统上有许多计算机程序可用于数据存取的工作。在本书中，这些程序能够从数据库中显示或获取部分或所有的序列，并以适当的格式保存，通常是有一个新的数据的库。

本章的其余部分介绍了比较和分析序列的方法，同时也为本书其他各章的做了简要的概述。主要介绍的概念是序列联配，这一方法是设法比较不同句子之间的每个字母或词，确定它们是否有相同的顺序，以表明它们是否有相关的主题。能够以这一方式联配的 DNA 和蛋白序列有相同生物学功能和共同的进化起源。序列联配是非常困难的计算问题，因为要进行非常多的比对，并且在包括词间空位的情况下有很多可能的不同的联配方式。动态规划是解决这一问题的一种计算方法，它将问题分解成很小的单元，只对序列开始部分做比对分析。一旦在序列开始部分搜索到最好的匹配，比对就逐步延伸到整条序列。相同的方法也被用于搜索 RNA 序列中可形成碱基对的互补区域，或是比较蛋白质的三维结构。

在基因组时代，一个主要的研究目的是为了分析基因的功能，阐述他们在特定生物体中的相互作用。在实验室中，模式生物的遗传学操作，测量基因表达的基因芯片，以及用于蛋白分析的蛋白质组都被用于收集新的生物学数据。来源于这些实验的信息需要组织成合适的数据格式，并且需要开发计算机程序，最好能通过网络来存取相关信息并进行数据分析。生物信息学家已经编写了不少这类程序用于序列的处理和分析。

目前，编写程序已经被简化成编写一些小的，被称为“对象”的模块程序用于处理各类简单的任务。许多对象组成的程序库能够处理几乎所有的任务，这样的库被大程序所使用就能够处理更具体的序列分析任务。比如，一个模块可用于从 GenBank 数据库中提取序列，而另一个模块可以将序列转化成为特定的序列格式。BioPerl 对象库就是这类对象编程的一个例子。BioPerl 的对象可用 Perl 语言编写。使用这些已经开发的模块，只要接受一些训练或是在学生或专业程序员的帮助下就能够很容易开发任何序列处理或分析的应用程序。但是，要真正的精通这些知识，有生物学背景的学生需要通过一些课程加强自己的生物学

基础，如分子生物学和生物化学的实验方法课程、种群和进化遗传学的进化分析课程。他们还必须学习数学、概率和统计、数据处理、数据挖掘和建模工具、计算机算法设计和编程，以及在可能的情况下学习基因组分析和生物信息学专题的高等课程。

计算科学家指南

生命的基本单元是细胞，细胞外侧是保护性的细胞膜，它包围着一些细胞器（亚细胞结构）以及能够支持细胞结构、提供细胞能量和进行自我复制的大而复杂的生物分子。在植物和动物中，独立的细胞相互合作形成多细胞组织和器官系统，实现生物体的生物学功能。本书主要涉及在细胞和生物体中调控生物学过程的生物序列的分析，以及指导生物体发育过程中细胞组织的指令信息的分析。

序列存贮在长化学链中，称为DNA，由A、G、C、T（分别代表腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶）四种不同的信息字符或碱基以及糖和磷酸骨架组成。DNA紧密地缠绕成染色体，在显微镜下就能够看见这样的亚细胞结构。除了如单细胞细菌这样简单的生物体外，染色体位于可见的细胞核中，细胞核周围包裹着细胞质。组织细胞有两套染色体，这是被称为二倍体的一种遗传结构，两条染色体分别来自于母本和父本。在有性发育过程中，形成称为配偶子的性细胞（精子细胞或卵细胞），它拥有一套染色体（单体型）。单体型细胞中的染色体组成了生物体的基因组。在有性复制过程中，来自于两个亲本的染色体重新分类配对。接着，拥有一套新配对染色体的配偶子重新生成，并最终传递到子代中去。

染色体的DNA序列编码了制造细胞蛋白质的指令。蛋白质是有化学活性的20个氨基酸组成的线性链。每个蛋白质特定的氨基酸序列由染色体上的DNA序列确定。蛋白质的形成使生物体能够形成结构以及行使生物学功能。在转录的生物学过程中，细胞通过搜索特定的序列模式（如启动子）“阅读”DNA序列，这些特定的序列模式标志着遗传信息单位——基因的起始位置。从该点开始，在生物分子上沿着一定的化学方向读序列，直到遇到标志着基因结束的序列模式才停止。转录生成的化学长链称为信使RNA或mRAN，对应于将要生成的蛋白质的氨基酸序列。mRNA分子在结构和化学上和DNA分子非常相似，但是他们通常是单链，并且以新的碱基尿嘧啶（U）代替了胸腺嘧啶（T），在主链上也连接着不同的糖。mRNA分子也有特定的序列模式，标记了蛋白质编码开始的位置。细胞质中称为核糖体的大细胞器能够结合在蛋白质编码开始的位置，按定义的化学方向移动，每次读三个碱基位置（一个密码子）确定一个氨基酸。接着，该密码子对应的氨基酸就添加到形成蛋白质的氨基酸链上。就这样，氨基酸逐个添加，直至碰到几个终止密码子之一。每个蛋白质的序列就这样基本与染色体上的原始编码序列线性一致。

蛋白质一旦形成，将迅速由线性字串折叠成简单的螺旋和折叠单元（即二级结构），随后这些单元将形成特异的三级结构。产生的蛋白质分子作为组成构成单元形成组织或者参与特定的化学活动。一个生物体的结构及其生物学功能取决于蛋白质组，及一个生物体产成的所有蛋白。并非所有的基因都被翻译成为蛋白质——某些在RNA水平上仍保持着4个字母序列的基因，调控着许多重要的细胞进程。简单的生物体拥有数以千量的基因，而较为复杂的生物体则拥有12 000～35 000左右的基因。在一些生物体中，尤其是植物类生物体，可能会存在巨大数量的重复基因或重复染色体，这些多余的基因逐代复制下去，有时生成包含100 000甚至更多基因的基因组。

有个概念对了解生物学家如何看待基因非常重要，就是所有生物体看起来都通过进化过程相互关联。即使是完全不同的生物体，譬如单细胞细菌和多细胞动植物，他们之间也可能有一些共同基因。这些基因通常并不是每次由新的生物起点产生，而是从先前的基因复制得来，在复制中DNA序列在有限程度上随机地产生突变。生物体可以通过二元树划分成组，在外面的分支代表联系更加紧密、进化时代较近的生物体，在内部的分支则代表更加原始，通常比较简单的生物体。这些树的构建可以基于生命信息（生物体的结构和行为），但是最近随着模式生物基因组的测序，基因组上基因的互补信息可以用来构建二元树。因此，很多的生物信息学研究关注对比基因和被翻译的蛋白质，将他们视为由4个字母（DNA）或20个字母（蛋白质）组成的线性字串。当某个生物体编码的一个蛋白质与另一生物体编码的一个蛋白质很容易比

对上的话，这就告诉生物学家们这些基因来自共同祖先并且会有相同的功能。

学习基因和蛋白质还有一个重要概念。为产生一项新的生物学功能生物体似乎运用三个策略而不是构造一个全新的基因。第一个策略是，复制已有的基因，其中的某个基因通过随机突变逐渐改变而发展出一个新的生物功能。在极少数情况下，这个新功能对生物体本身有利，自然选择就在这个生物中扮演了基因催生的角色。这个过程可以建立起一个相互关联的、功能重要的、可以通过进化传递的基因家族。第二个策略是，生物体将现有的基因部件（功能域）结合在一起从而形成新的基因，这些部件本身代表蛋白质三维结构或者生物功能的基本单位。生物信息通过序列比对，运用统计方法寻找序列的共有模式，做了大量的基因、蛋白质家族和序列功能域发现的工作。生物体内已有基因功能分化的第三个策略是非常少见的从不同、不相关的生物体内转移基因。通常情况下基因都是由父系向子系传递，生物学家把这个过程称作垂直传递，因为遗传沿着树状家谱的分支传递。另一种遗传方式是一段外源 DNA 随机地转入细胞并掺入已有染色体从而增加了基因。因为贡献基因的生物来自通常并不发生基因交换的不同种群，所以这种遗传被称作横向转移。在远古生命中，甚至认为整个细胞都能融合来创造新的物种。横向转移仍在单细胞细菌的进化中起主要的作用，尤其是在发展对特殊抗生素的抵抗力方面。所以，所有生物体都共有许多基因或基因的组成部分，现在这些都能在生物体基因组中被识别。

目前生物学家面临的两大挑战是如何发现基因的生物功能以及理解基因相互之间的作用是如何调节生命活动的。如果得不到其他物种的同源基因的信息，可以使用以下实验方法：①破坏基因序列（对序列进行改变，或插入，或删除）。②引入改变形式的基因的 mRNA 拷贝，它能导致细胞 mRNA 拷贝的降解，以一种半持久的、可遗传的方式（外成性改变）改变基因的结构，使得基因不能转录，从而关闭基因的表达。通过一整套生物学实验，可从这些遗传性或表观性改变的影响来推断基因的功能。除了遗传分析以外，生物学家还通过 DNA 微阵列技术和蛋白质分析实验（蛋白质组学）来跟踪许多细胞基因和蛋白质的表达。这些技术帮助发现与生物功能相关的基因表达和蛋白质发生的模式，比如癌细胞中的非正常模式。

生物信息学专业学生的基本要求

- 知道序列、蛋白质和基因组数据的存贮位置以及保存方式。
- 具备编程技术能从网络，数据库中提取数据，将它们整理，重新储存成合适的数据库格式。
- 掌握足够的生命科学指示，熟悉序列和基因组信息与基因功能及蛋白质结构功能之间关系，并知道如何适当保存这些数据。
- 了解在不同物种找到功能相似基因的最新序列基因组分析方法，达到基因调控分析，蛋白质结构功能预测的目的。
- 懂得如何处理海量的数据如基因表达微阵列，蛋白质组数据，种群中的序列变异。
- 能够整合分散的数据，并利用已有的工具来进行数据挖掘，在数据中发现新的关系。
- 与计算机科学家、数学家、统计学家沟通，发展所需要的新的分析工具和模型。

术语表

alignment（联配） 通过查找两个甚至更多序列中相同次序的字符串或字符串模式来对序列进行比较。

algorithm（算法） 一种数据分析的方法，强调时间及空间效应，可从数学上证明其正确性。例如用于序列比对的动态规划算法。

annotation（注释） 基因组序列上的基因定位，包括编码蛋白质，RNA 的基因，同时提供了编码蛋白质和 RNA 分子的序列和位置。

codon（密码子） DNA 中的三核苷酸，被细胞翻译成蛋白质中的氨基酸。在 64 个可能的密码子中，有 61 个通常被翻译成 20 个氨基酸中的一种，其余 3 个是终止子。mRNA 以一系列密码子的形式携带蛋白质序列信息。

comparative genomics（比较基因组学） 比较不同物种的基因组中基因数目、位置、基因的功能，目的在于识别在特定生物中起特殊功能的家族基因。

database（数据库） 存储数据的组织系统，通常是一系列相互关联的表格，被称为关系数据库。

distance score（距离分值） 指相关序列的联配中，比对字母不同或很少找到的那些位点，通常（但并不总是）不考虑缺口位点。

DNA 由4个核苷酸A，T，G，C组成的双链螺旋分子，A总是与另一条链的T相配对，而G总是和C配对，核苷酸之间的化学相互作用将两条链连接在一起。两条链分开后，其中每一条链作模板都可以基于A/T，G/C配对原理合成新链。这条适用于所有生物繁殖的分子机理是由James D. Watson和Francis Crick发现的。DNA分子序列可以用DNA测序机以每次500～800个的速度阅读出来。

dot matrix analysis（点矩阵分析） 以图形化的方式比较两条序列。一条序列从图的顶部（或底部）开始水平书写，另一条沿左侧从上至下，在两条序列中都出现的核苷酸的交叉点被标记，对角线意味相似性。

dynamic programming（动态规划） 允许配对，错配和空缺的序列联配算法。算法首先从序列开始位置找到最佳匹配，然后在每条序列上逐渐增加最佳匹配直至完成。

extreme value distribution（极值分布） 一些诸如序列比对分值之类的数据服从一种“长尾”分布，在高值端衰减得比正态分布慢得多。其中一种缓慢衰减的分布称为极值分布。不相关序列或随机序列之间的比对分值就是一个例子。这些分值能达到很高的值，尤其在进行数据库相似性搜索进行了大量比对时。特定分值的概率可通过Gumbel型的双负指数函数的极值分布准确预测。

functional genomics（功能基因组学） 评估从基因组比较中发现的基因功能。新识别的基因常通过引入突变并观测后代生物性状改变的方法来判定功能。

gene（基因） 与某生物功能相关的一段DNA，通常是蛋白质的氨基酸序列。与合成DNA新链相同，DNA运用碱基互补原理复制到mRNA分子。

gene expression microarrays（基因表达微阵列） 代表一个物种中大量基因的显微镜载玻片上的DNA微点阵列。微阵列用来比较一个生物样本中整套基因的mRNA水平，通过合成化学性质不稳定、带有荧光标记的与mRNA互补的DNA（cDNA）来探测。

genome（基因组） 一个生物全部的遗传信息，包括特殊的蛋白质，RNA分子以及其他的序列。他包含体细胞中一半的染色体，性染色体的全部。

global sequence alignment（全局序列联配） 一种包含所有序列的比对方法。

homologous（同源性） 基因因相似的序列被证实来自相同的祖先基因。

local sequence alignment（局部同源联配） 通过最高密度匹配方法进行序列区域比对的方法。Maximum parsimony tree一最节省树：多序列比对中已观察到的变化的图形表示，使得在树分支上变化数量的和达到最小。

model organism（模式生物） 可被遗传操纵，从后代生物的变化中找到具有特殊功能的基因的物种。从这样的模式生物研究中得到的结论也可应用于其他生物。模式生物有果蝇，酵母，斑马鱼，老鼠以及植物阿拉伯芥。

motif（protein） 可以代表蛋白质结构中的活性中心，功能区域的一小段氨基酸。

multiple sequence alignment（多序列联配） 排列三个或更多的序列，试图将功能或进化相关的序列位点放在同一纵栏，允许错配，空缺。

orthologous（同源） 形容来自不同生物体，两个甚至更多的基因，因为相似被预测拥有相同的生物功能。

orthologs 同源的基因。

paralogous（旁系同源） 形容一个生物体内一群相似基因，被预测来自于同一祖先的基因复制。

paralogs 旁系同源的基因。

phylogenetic analysis of sequences（序列系统进化分析） 试图应用进化树来找出一系列相似序列的进化关系。

position-specific scoring matrix（PSSM）（位置特异记分矩阵） 表示从一些列关联基因的多序列排列实验中找到的差异的表格。表格纵列对应排列中的纵列，行对应每列序列特征的出现频率。这个频率通常除以序列中特征频率以得到奇数，为了方便，奇数用其对数形式表示。

protein（蛋白质） 由一条长链氨基酸组成的分子，它的序列由基因的密码子序列决定，长链折叠形成三级结构，唯一对应具有某生物功能的蛋白质。

什么是生物信息学?

过去，生物信息学被定义为包括生物学、计算机科学、数学、统计学的交叉学科，它研究生物序列数据、基因组内容及排列、预测生物大分子的结构与功能。随着基因组时代的来临，现在生物信息学在生物学和医学研究中起着更重要的作用，论文发表量逐年上升（Luscome et al. 2001）。

生物信息学与信息学及计算生物学等研究领域有交叉。过去，信息学是这样的学科：数学家、计算机科学家、统计学家、工程师共同发展在诸如健康等领域的信息管理的支持技术。现在，生物信息学也参与进来，它组织与基因组相关的数据，旨在将这些信息应用到农业、制药业和其他商业中。有两种类型的生物信息：序列信息，以及来自基因组和由实验获得的基因产物结构-功能分析等内容。针对人类基因组，生物信息学的作用是收集约35 000个的人类基因的生物信息，以发现那些基因在人类疾病中起着最显著的作用。利用现代技术，例如基因表达芯片，细胞和组织中基因遗传操作，以及蛋白质结构与功能的快速评估，可以收集到许多基因功能方面的新数据。

对于上述问题，密切相关领域的计算生物学也提供了计算方面的支持，但与生物信息学所提供的支持是有所不同的。一方面，计算生物学一般关注发展一些能解决诸如多序列比对、基因组碎片拼接等困难问题的新型高效算法。而生物信息学更多地注重开发数据管理与分析的实用工具，如基因组信息的显示和序列分析，而不注重于对高效性和可证明的正确性的研究。在很多情况诸如进行多序列比对和数据库相似性搜索时，得不到一个适当的序列改变的模型，或者合理定义的序列分析问题太复杂而不能在合理的时间内解决。在这种情况下，生物信息学能提供满足当前需求但没有理论证明基础的计算方法。因此，今天的生物信息学领域支持包括确定数据的生物学意义、提供组织数据的专门技术、发展用于新数据挖掘的实用计算工具等很大范围的研究工作。当信息具有实用的重要性时，生物信息学协助实际应用诸如识别药物疗法的新蛋白药靶等。然而不可否认地，生物信息学、信息学和计算生物学都在快速发展并将在基因组数据的利用方面起着不同的作用。

首先被收集的蛋白质序列

由于蛋白质测序方法的发展（Sanger 和 Tuppy 1951），一些具有代表性的常见蛋白质家族得以测序，如不同物种的细胞色素蛋白。早在 20 世纪 60 年代，Margaret Dayhoff（1972，1978）和她在位于华盛顿的国立生物医学研究基地（National Biomedical Research Foundation，NBRF）的合作者们首先建立了蛋白序列数据库，这一数据中心最终发展成了著名的蛋白信息资源库（Protein Information Resource，PIR；http://watson.gmu.edu:8080/pirwww/index.html）。NBRF 自 1984 年起维护这些数据库。1988 年，作为 NBRF 的合作伙伴，PIR-国际蛋白序列数据库（PIR-International Protein Sequence Database，http://www-nbrf.georgetown.edu/pir）、慕尼黑蛋白质序列中心（Munich Center for Protein Sequences，MIPS）、日本国际蛋白质信息数据库（Japan International Protein Information Database，JIPID）相继成立。

Margaret Dayhoff

Dayhoff 和她的合作者们基于序列的相似程度将蛋白质组织成家族和超家族。由此衍生出了反映一群紧密相关的蛋白质序列的氨基酸频度变化的表格。为了保证观察到的氨基酸变化是一次性变化而非两次连续的变化，仅那些序列差异性小于 15%的蛋白质被选来作研究，这样可以避开多次改变的嫌疑。根据序列比对的结果可生成系统发育树，以图形的方式显示出那些由于密切相关而位于系统发育树同一分支上的序列。这些树可用于对来源于不同物种的蛋白质序列在进化中形成的氨基酸改变打分（图 1.1）。

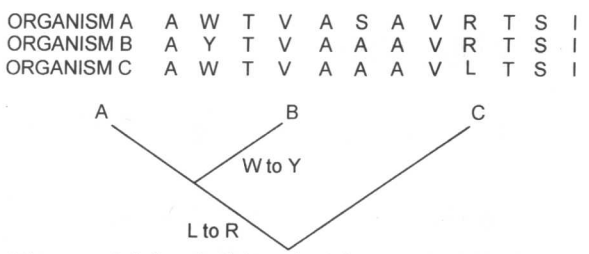


图 1.1 对进化上相关的蛋白质序列预测系统发育关系和可能的氨基酸改变的方法

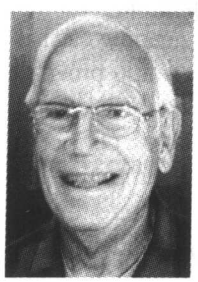
图中显示了来自 3 个不同物种的相同蛋白质的 3 条高度保守的序列（A，B，C）。这些序列非常相似，在进化中每个位置只有一次改变。蛋白质序列只有一个或两个位置的替换，由此可构建如图所示的树。一旦树形成后对应的氨基酸改变也就确定了。图中显示的特定氨基酸改变的发生概率要高于随机替换过程出现的改变

评估系统发育树分析结果的规则是：两条序列中相同和保守的氨基酸越多，它们进化自共同祖先的可能性越大。如果两条序列非常相似，蛋白质可能有相同的生化功能和三维立体结构。因此，可以构建一组称为 PAM（percent accepted mutation）表的矩阵，来描述进化选择所接受的氨基酸突变百分比。这些表格给出了系统发育树中由一个氨基酸变成另一个氨基酸的概率，因此能显示出两个序列中哪些对应位置的氨基酸是最保守的。PAM 表还可以用来测度蛋白质序列间的相似性以及在数据库中搜索匹配序列。

通过建立第一个蛋白质序列数据库以及建立蛋白质序列比较的 PAM 表，Dayhoff 和她的同事们在许多方面对现代生物序列分析做出贡献。氨基酸置换表成为序列比对和数据库相似性搜索的常规工具，它们在这方面的应用将在第 3 章和第 6 章讨论。

始于 20 世纪 80 年代早期的 DNA 序列数据库

新墨西哥州的 Los Alamos 国立实验室的 George I. Bell 于 1974 年建立的理论生物学与生物物理研究组最先收集了 DNA 序列，并将其存放到 GenBank 数据库中。这些物理学家试图为实验室的工作，主要是免疫学方面的研究提供理论背景。该小组的研究扩展到了计算生物学和生物信息学的领域，并且在 1982～1992 年期间由 Walter Goad 和其同事开发了 GenBank 的第一个版本。Goad 早在 1979 年就已经开始对 GenBank 数据库原型进行构思。由 DNA 翻译的蛋白质序列也被华盛顿的国立生物医学研究基地的蛋白质信息资源数据库（PIR）收录。其他相关的数据库，如欧洲分子生物学实验室（EMBL）数据库于 1980 年成立（http://www. ebi. ac. uk），日本 DNA 数据库（DDBJ，http://www. ddbj. nig. ac. jp）于 1984 年成立。GenBank 现在由国家生物工程信息中心（NCBI，http://www. ncbi. nlm. nih. gov）管理。GenBank，EMBL 和 DDBJ 现在已形成国际核酸序列数据库联盟（http://www. ncbi. nlm. nih. gov/collab），旨在进行每日的数据交换。PIR 也进行着同样的工作。



Walter Goad

每年第一期的《核酸研究》杂志会刊登许多序列数据库类型。GenBank 数据库中序列数目的增长可在 http://www. ncbi. nlm. nih. gov/GenBank/genebankstats. html 查到。

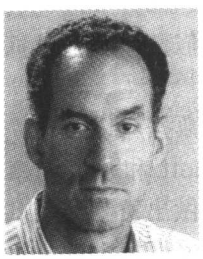
最初，这些数据库中的一条序列记录包括其计算机文件名以及 DNA 或蛋白质序列文件。最终，这些记录扩展了更多的序列信息，如功能、突变、编码蛋白质、调节位点和参考文献。这些注释信息与序列一起存放成一种易于搜索的数据格式。类似的数据库和格式将在第 2 章讨论。

核酸序列数据库 GenBank 和 EMBL 的记录每天都有大量更新。注释所有这些新序列是耗时、艰苦的过程，有时还易出错。近年来，这个过程变得更自动化，但是又引起了准确性和可靠性的问题。GenBank 在 1997 年 12 月有 1.26×10^9 碱基，到 2004 年 4 月增长到 39×10^9 碱基。尽管储存的序列数成指数增长，有效的搜索方法的应用保证对这些序列快速公开的检索。

为减少数据库搜索所得到的匹配数目，建立了非冗余数据库对于相同序列只列出一条记录。NCBI Refseq 就是这样的数据库。然而，许多序列数据库，如 NCBI 用于 BLAST 搜索的非冗余 NR 数据库，仍然包括了大量相同的基因和蛋白质记录，它们可能来源于不同数据库的序列碎片、专利、重复序列以及其他类似的序列记录。

从公众数据库中方便的获取序列

提供序列数据库访问的重要一步是开发查询页面，这一工作形成了一些重要的序列数据库（GenBank，EMBL 等）。这项技术一个例子是在 NCBI 早期由 D. Benson、D. Lipman 及其同事开发了称为 GENINFO 的菜单驱动程序。这个程序快速搜索已建立索引的序列数据库，为生物学家的查询提供相匹配的记录。NCBI 随后派生了有简单窗口界面的 Entrez 程序，并最终开发成网页界面，形成 NCBI 网站（http://www.ncbi.nlm.nih.gov/Entrez）。



David Lipman

这些程序的想法就是用关键字搜索标准记录字段，以灵活的方式为序列数据库提供易于使用的界面。一些主要数据库中的序列还记录了序列的附加信息，如索引号、序列名称和别名，相关基因名称，调节序列类型，物种来源，参考文献，已知突变。Entrez 程序能访问这些信息，因而能够快速搜索整个序列数据库，查对到一个或几个特定匹配的术语。

这些程序也能根据预先的相似性比较定位相似序列（Entrez 称为“邻居”）。当对一项或多项术语执行查询时，简单的模式搜索程序只能找到精确匹配结果。相反，Entrez 能够容易的进行相似或相关术语的搜索，或是进行多个组合选项的复杂搜索，并将搜索结果按与原查询的相似程度排序。Entrez 最初允许直接访问 DNA 和蛋白质序列数据库及它们的参考文献，甚至允许索引不同或同一个数据库中的相关记录或相似序列。最后，Entrez 还提供了对所有 Medline 数据库的访问，这是一个位于华盛顿区的国家医学图书馆全文文献数据库。另外也提供对许多其他数据库的访问，如物种系统发育数据库、基因组数据库和蛋白质结构数据库。NCBI 任何用户—个人，政府、工业或是研究机构提供免费访问。他们所做出的这一决定对生物学研究的促进作用是不可低估的。NCBI 每天都要处理数百万个访问事件。

注意事项：Entrez 这种数据库查询程序推动了数据库和日益增长的序列数据以及生物化学杂志保持同步。然而，如同任何自动化方法一样，必须注意数据库搜索也许不能够提取出所有的相关材料，重要的记录可能会遗漏。每条数据库记录在某个阶段都需要手工编辑，这会造成少部分不可避免的拼写错误和其他问题。有时在数据库中应有的记录没有被检索到，可能因为搜索术语在相关数据库记录中被误拼，或这条记录未出现在数据库中，或是存在其他更复杂的问题。如果各种搜索尝试都失败了，将此问题报告给程序编写者或系统管理员将有助于解决这一问题。

序列联配程序的发展

DNA 的测序是通过排列一套由测序仪检测的来自测序凝胶的峰值（A，G，C 或 T），这是一个很易出错的过程，主要依赖于数据的质量。为了能够正确地收集数据，华盛顿大学的 Phil Green 和他的同事编写了 phred（Ewing and Green 1998，Ewing et al. 1998）和 phrap 程序，用以帮助读取和处理测序数据。phred 和 phrad 现在由 CardonCode Corporation（http://www.codoncode.com）发布。这些序列读取工具对更准确地收集序列数据有很大帮助。

20 世纪 70 年代末，随着序列数据的增长，人们也增加了使用各种方式开发计算机程序分析序列的兴趣。在 1982 和 1984 年，《核酸研究》出版了两期专辑刊登计算机在序列分析方面的应用，包括大型机和微机程序。不久后，J. Devereux 在威斯康星组建了遗传学计算机研究组（Genetics Computer Group，GCG），提供了一套在 VAX 计算机上运行的分析程序。同一时期还创立了其他的公司，包括 Intelligenetics，DNAstar 等，他们为序列分析提供可在微机上运行的程序。实验室也开发并在免费或低费用的基础上共享计算机程序。

1977 年 Maxam、Gilbert 及 Sanger 等发明了

DNA 序列分析的方法。Sanger 方法在第 2 章开头将详细介绍。

这些商业化及非商业化程序仍被广泛使用。而且，有许多网站为各种序列分析提供平台，它们对学术机构免费，商业用户可以适当的价格购买。以下是对序列分析方法发展的简单综述。

序列比较的点阵法或图示法

1970 年，A. J. Gibbs 和 G. A. McIntyre (1970) 提出了氨基酸和核酸序列比较的新方法。可以画图说明这一方法，将其中一条序列横排在页面中，另一条序列竖排在左下方的位置。当相同的字母在两条序列中同时出现时，在序列对应的两个字母的交点位置上放置一个点（图 1.2）。通过扫描图中一系列的点，能够在两条序列中找到揭示相似性或有相同字符的字串的对角线。长序列的比较通过使用较小的点也可以在一页纸内实现。

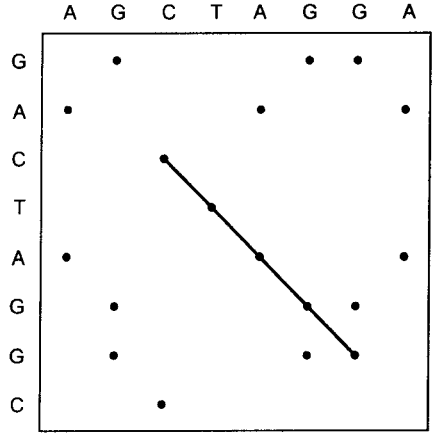


图 1.2 两条 DNA 序列 AGCTAGGA 和 GACTAGGC 的点阵比较图

对角线上的点表明在两条序列中同时出现的一连串序列 CTAGG

由于点阵的方法找到的对角线在水平和垂直方向上的平移是根据序列间差异而得到的，所以很容易反映序列间的缺失和插入。单条序列的自比较能够揭示序列中的相同（正向重复）或相反方向（反向重复或回文）的重复区域。序列自比较的方法能够揭示一些特征，如染色体之间的相似区域、串联基因、蛋白质序列中的重复区域、有重复碱基组成的低复杂度序列，或是 RNA 中能够配对形成双链结构的互补序列。

由于序列间的弱相似性对角线可能无法在图中找到，Gibbs 和 McIntyre 为了找到显著的匹配，计算了图中所有可能出现的对角线，同时与一些随机序列比对时出现的对角线数目相比较。这套随机序列提供了不相关序列之间联配分值的分布范围。如果两个测试序列之间对角线分值有显著性，即对角线分值应该超过随机序列比对的对角线分值，就可以证明序列已成功联配。同时，高于随机序列的对角线分值的程度提供了一个统计结果，该结果能决定测试序列对角线的显著性。随后，Maizel 和 Lenk（1981）开发了各种过滤和彩色显示配置，极大增加了点阵方法的有效性。以点阵展示序列比对在 DNA 和蛋白质序列相似性分析以及基因和染色体上重复序列的分析中一直起着重要的作用，这一内容将在第 3 章介绍（第 79 页）。

动态规划方法在序列联配中的运用

虽然点阵方法可以检测序列中的相似性区域，但是不能很好的解决由于序列间的弱匹配或缺失插入区域的存在而中断序列相似性匹配的问题。因此，人们寻找新的方法希望通过点阵找到一条曲折路径，在两条序列中发现最有可能的联配，或称为最优联配。如图 1.3 所示，将其中一条序列连续的横排在页面中，另一条序列排列在其下方，相同碱基排列在同一栏中，如果同一栏中有不匹配的碱基则被视为错配或加入空位表示插入（在另一条序列中是缺失）。对于一条长度为 300 的蛋白序列，如果要在所有可能的匹配中找到最优的联配，同时考虑插入和缺失的匹配情况，需要进行 10^{88} 比对，这在计算上有很大的困难（Waterman 1989）。

```
SEQUENCEA  A  G  Δ  Δ  C  D  E  V  I  G
SEQUENCEB  A  G  E  Y  C  D  Δ  I  I  G
```

图 1.3 双序列比对显示匹配、不匹配和空缺（▲）

最好的或最合适的比对需要允许所有这三种变化

为了简化上述工作，Needleman 和 Wunsch (1970) 将这个问题分解为逐个氨基酸的比较，以渐进的方式建立起序列联配。序列比对从两条序列的结尾开始，而后同时向前移动一个氨基酸，允许匹配、错配或是在一条序列中出现插入或缺失氨基酸的各种组合形式。这一方法在计算机科学中称为动态规划。Needleman 和 Wunsch 的方法产成了以

下结果：①可能存在的序列联配，即包括了匹配、错配或是单碱基插入或缺失的各种可能组合的序列联配。②序列联配的打分系统。其目标就是通过判断最高分以确定最好的联配结果。其中每对匹配的碱基记 1 分，错配记 0 分，空位给予一定的罚分。而后沿着序列将分值相加得到联配的总分，拥有最高分值的联配被定义为最优联配（图 1.4）。

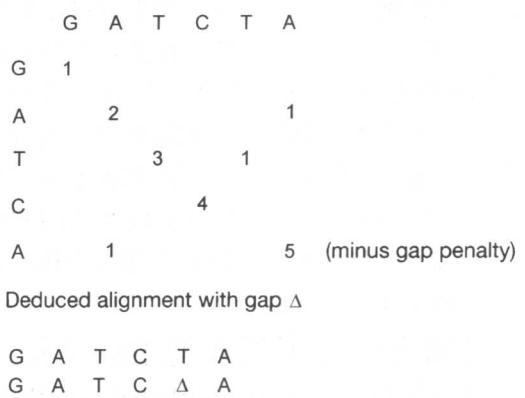


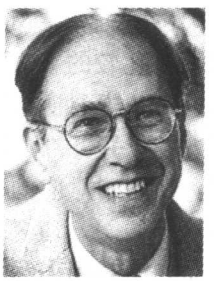
图 1.4　使用 Needleman-Wunsch 的方法联配 GATCTA 和 GATCA 序列

首先，在本例中我们定义所有匹配的碱基记 1 分，错配记 0 分（未显示）。接着将对角线上的分值连续相加，在本例中总分值为 4。这时横行因无法延伸到另一对相匹配的碱基对而不能得到 5 分。但是，如果在 GATCA 中插入空位生成 GATCΔA 就可以继续延伸，在这里 Δ 表示一个空位。引入空位后，需要从目前行列匹配的总分值 5 分中减去空位罚分值。最优联配的寻找开始于最高分值对应的碱基，并回溯对这一最高分值有贡献的所有位置。

如图 1.4 所示，所有可能的联配产生的过程开始于其中一条序列的末尾位置，并在一个与点阵非常相似的矩阵中有序地移动（图 1.2）。允许从序列任意位置开始，并且包含匹配、错配、插入和缺失的情况，寻找最高分值并放置在矩阵的每个位置中。而后通过寻找图中的高分位置并回溯这些路径就能找到最优序列联配。对应于这些路径的碱基相互匹配最终序列也得以联配。

搜索序列间的局部联配

上面提及的方法是在整条序列中寻找最优联配，被称为全局联配。Smith 和 Waterman（1981a，b）意识到在 DNA 和蛋白质的序列中许多有生物学意义的区域是匹配较好的子区域，而其他弱相关的部分匹配不显著。因此，他们对 Needleman-Wunsch 算法做了重要的修改，开发了称为局部联配或 Smith-Waterman（或 Waterman-Smith）的算法用于类似局部区域的比对。他们还认识到序列之间不同长度的插入和缺失与序列进化上的变化有关，因此他们调整了该方法以适应这样的变化。最后，他们还从数学上证明了动态规划能够保证在序列间找到最优的匹配结果。第 3 章有相关算法的具体介绍（第 83 页）。



Mike Waterman

Temple Smith

研究者开发了相似性分值和距离分值两种互补的方法用于两条序列联配打分。如图 1.3 所示，在联配序列中有三种情况——一致匹配，错配和空位。为了获得相同的分值，使用简单的打分机制将匹配设定为 1；匹配的分值累加除以匹配以及错配残基的总数（空位通常是忽略的）。这一序列相似性打分机制是生物学家所熟知的，由 Needleman 和 Wunsch 设计，Smith 和 Waterman 也采用这一方法。另一种打分机制，即距离分值，是一条序列转变成另一条序列所需要的替换次数的累加值。这一分值在系统进化分析时对于蛋白质或基因序列进化距离的预测十分有用。该方法是数学家特别是 P. Sellers 的贡献。距离分值计算联配序列中的错配残基总数，并除以匹配和错配残基的总数。距离分值代表了在不考虑空位的情况下，从一条序列转化成为另一条序列所需要的替换次数。在图 1.3 的例子中，有 6 个匹配的残基，1 个错配。对于这一联配的相似性分值是 6 个匹配除以 7＝0.86，距离分值是 1 个错配除以 7＝0.14。一般来说，相似性分值和距离分值相加为 1。如果没有空位出现，联配的长度应该和序列的长度一致。如果出现空位，联配的长度要比序列的长度更长。可以注意到序列的长度等于两倍的匹配和错配的和值，再加上插入或缺失的数目。这样，在我们的例子中，可以这样计算 8＋9 ＝2×(6＋1)＋3＝17。通常更复杂的打分机制用于生成有意义的联配，联配用似然值和几率分值（第 3 章）评估，但是相似性分值和距离分值之间

的相反关系还是保持着的。

在序列联配中存在一个困难问题，既判断特定的序列联配是否有显著性。我们需要知道序列联配的分值真能够揭示两条序列是相似的吗？或是这样的分值在两条不相关的序列中（或是计算机生成有相同的组分的随机序列）也能容易的找到吗？这一问题由 S. Karlin 和 S. Altschul（1990，1993）提出，在第 4 章中做了具体的描述（第 136 页）。

Karlin-Altschul 分析了不相关或是随机序列的分值，发现这一分值通常比常规分布的期望值要高。分值的分布符合正偏态，也称为极值分布，对随机或不相关序列联配分值的分析提供了一种方法，该方法可以用于评估一段联配的分值在相同长度的不相关或随机序列中出现的概率。正如第 6 章中讨论的，这一发现对于评估查询序列和数据库中序列的比对特别有用。另外，一段特定联配的分值的评估必须考虑数据库搜索时相似序列的数目。假设使用随机序列联配分值的极值分布来评估目标蛋白质序列和数据库中蛋白质序列匹配的分值，有 10^{-7} 的概率和两条不相关序列的联配分值一样高。如果数据库中有 80 000 条蛋白质序列，将进行 80 000个不同的比较。每次比对都会有不相关序列联配的分值与 10^{-7} 概率分布下的分值一样高。如果做80 000次比对，就增加了80 000次这样的机会。这些序列中的任意一条能够得到 10^{-7} 概率分布下的比对分值的概率是 $10^{-7} \times 8 \times 10^{-4} = 8 \times 10^{-3} = 0.008$（称为期望值）。该值介于 0.02～0.05 之间被认为有显著性。但是，在基因组比较时，当数据库中的序列与查询序列进行联配时，低期望值（如 10^{-20}）更有意义。即便在这样的期望值下找到了联配，也需要仔细检查序列联配中的缺陷，如不切实际的氨基酸匹配，连串重复的氨基酸都会降低序列联配的置信度。

多序列联配

多序列联配是用于同时比较三条或多条序列的方法（早期的例子见 Johnson 和 Doolittle 1986）。它采用运算密集的方法，通常基于连续联配最相似的成对序列。常用的工具有 GCG 软件中的 PILEUP，CLUSTALW（Thompson et al. 1994），以及 T-COFFEE（Notredame et al. 2000）。软件的最新发展在第 5 章中有具体描述。一旦生成一套相关的生物大分子序列（一个家族）的多序列联配，可以确定特定家族中高度保守的区域（Gribskov et al. 1987），也可以用于确定家族新成员。序列谱和位置特异性矩阵（PSSM）是两种替代多序列联配的重要的计算工具。

多序列联配也是系统发育建模的第一步。检查序列联配的每一列，接着构建最可能的系统发育关系或系统发育树，这样的系统发育关系或树基于在序列联配中观察到的改变的残基。

多序列联配的另一种用途是在一套蛋白质或 DNA 序列中的搜索模式（Stormo et al. 1982；Staden 1984，1989；Stormo 和 Hartzell 1989；Lawrence 和 Reilly 1990）。对于蛋白质序列，模式定义为结构或功能的保守组件。对于 DNA 序列，模式可能特指调控蛋白质在启动子区域的结合位点，或是 RNA 分子上的修饰信号。统计或非统计的模式搜索方法在这一领域有广泛的应用。实际上，这些方法都是排列序列，而后当序列联配后，试图在序列上定位一系列能够有最高匹配数目的连续的字符。神经网络、隐马尔科夫模型、期望极大化以及吉布斯采样的方法（Stormo et al. 1982；Lawrence et al. 1993；Krogh et al. 1994；Eddy et al. 1995）都可用于该分析。第 5 章有这些方法相关的解释和示例。

预测 RNA 二级结构的几种方法

使用计算机预测 RNA 二级结构的方法发展得也很早。在 RNA 分子中，如果在 RNA 序列的下游重复并且有化学方向相反的互补序列，这一区域可能组成碱基对，形成发卡结构，如图 1.5 所示。Tinoco（1971）等人在寡核苷酸分子中合成这样的对称结构，并且尝试着用一些方法预测其的稳定性，这些方法用能量值表估算模型中堆积的碱基对、自由能的相关性以及环结构的失稳效应值（Tinoco et al. 1971；Salser 1987）。单链环以及其他不匹配的区域都会降低预测的能量。