简明生物信息学

JIANMING SHENGWU XINXIXUE

主审 黄 健
主编 林 昊 郭锋彪 王 栋

电子科技大学出版社

简明生物信息学

JIANMING SHENGWU XINXIXUE

主审 黄 健
主编 林 昊 郭锋彪 王 栋

电子科技大学出版社

图书在版编目（CIP）数据

简明生物信息学 / 林昊，郭锋彪，王栋主编. —成都：电子科技大学出版社，2014. 11
ISBN 978-7-5647-2618-8

Ⅰ. ①简… Ⅱ. ①林… ②郭… ③王… Ⅲ. ①生物信息论 Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字（2014）第 207578 号

简明生物信息学

主 审 黄 健

主 编 林 昊 郭锋彪 王 栋

出 版：电子科技大学出版社（成都市一环路东一段 159 号电子信息产业大厦 邮编：610051）
策划编辑：高小红 李述娜
责任编辑：李述娜
主 页：www.uestcp.com.cn
电子邮箱：uestcp@uestcp.com.cn
发 行：新华书店经销
印 刷：成都蜀通印务有限责任公司
成品尺寸：170mm×240mm 印张 11 字数 215 千字
版 次：2014 年 11 月第一版
印 次：2014 年 11 月第一次印刷
书 号：ISBN 978-7-5647-2618-8
定 价：25.00 元

序　言

生物信息学是生命科学领域的“新兴学科”，当前，国内几乎每所开设生命科学相关专业的大学都开设了生物信息学课程，甚至专门开设了生物信息学专业，建立了生物信息学系或学院。基于电子科技大学较强工科背景的特点，我们专门为我院的生物技术专业撰写了此手稿，并在此基础上进一步编撰成书，方便本校学生学习。当然，如果本书能够被其他院校选用为教材或参考书，将是编者们莫大的荣幸。

在大多数的生物学家看来，生物信息学总是处于一种工具性的地位，无非是开发一些软件、方法，方便实验工作者使用。事实上，早期的生物信息学发展也是沿着这条路线进行的，比如最早的序列比对工具的出现；然而，随着测序技术的完善，科技发展已进入大数据时代，生物学正在从实验型向实验-理论结合型发展。理论生物学已逐渐走向生命科学的前台，并已受到越来越多的重视。实际上，早在薛定谔提出“生命是什么？”这一严肃性问题之时，理论生物学就已悄然诞生，甚至，他的小册子《What is life？》被誉为“分子生物学”领域的《汤姆叔叔的小屋》。

我国在理论生物领域的发展与世界几乎同步，从早期的“酶动力学”问题到“密码-序列-结构-功能”问题，到20世纪末参与的“人类基因组计划”，都凝聚了老一辈科学家的心血，也是他们把我们带入“生物信息学”这个领域，在此衷心地感谢他们。尽管，本书是作者为本科生上课的讲义升级版，但我们希望通过该书，不仅给学生们介绍一些必要的生物学软件，也给同学们讲述一些比较精妙的生物学理论。如果通过该书，能够吸引一些钟情于生物信息学研究的同学参与科研之中，便是编者最大的安慰。鉴于编者才疏学浅，文中错误在所难免，敬请读者指正。

编　者

2014年8月

目 录

第一章 绪论 1

第一节 基本概念 1
第二节 生物信息学的发展 2
第三节 研究任务 4
第四节 章节安排 5

第二章 序列比对与分子进化 6

第一节 相似性得分系统 6
一、基本概念 6
二、序列的相似性 7
三、序列比对得分矩阵 8
第二节 序列比对方法 14
一、点阵法 14
二、动态规划法 15
三、序列比对工具 17
第三节 数据库搜索 18
一、FASTA 18
二、BLAST 21
第四节 分子系统发生分析 28
一、分子钟学说与中性进化理论 28
二、进化距离 30
三、系统发生树构建方法 33
第五节 系统发生树构建软件使用 36
一、PHYLIP 使用 36
二、MEGA 使用 38

第三章 密码子与 DNA 序列信息 42

第一节 遗传密码的稳定性 42
一、四碱基的 DNA 序列 42
二、遗传密码的简并规则 44
三、氨基酸的不可替代性 49

第二节 DNA 的短程关联和长程关联 51
一、DNA 序列的短程关联 51
二、DNA 序列的谱分析与长程关联 54
三、DNA 序列阅读框架的非均匀性 55
第三节 DNA 的数学表示 56
一、DNA 序列的混沌图示法 56
二、DNA 序列的算术描述 57

第四章 基因组信息学 60

第一节 基因预测 62
一、真核基因预测 63
二、原核基因预测 66
三、病毒基因预测 70
四、元基因组的基因预测 72
五、微生物基因组的重新注释 74
第二节 基因组的组成及结构的分析及预测 75
一、真核生物基因组的 GC 等值区及 CpG 岛 75
二、细菌的基因组岛及复制起始位点的预测 80
三、密码子使用策略分析 84
第三节 人类疾病相关基因的识别 88
一、人类疾病相关基因的识别 88
二、致病菌毒力基因与耐药基因的识别 89
三、人类重要疾病数据库 91
第四节 非编码 RNA 的预测 92
第五节 利用 GO 注释基因功能 94

第五章 蛋白质信息学 98

第一节 蛋白质数据库 98
一、Uniprot 蛋白质序列数据库 98
二、PDB 蛋白质结构数据库 100
第二节 蛋白质结构预测 102
一、蛋白质结构 102
二、蛋白质二级结构预测 105
三、蛋白质三级结构预测 108
四、蛋白质结构相关预测 113
第三节 蛋白质功能预测 115

一、蛋白质亚细胞定位预测.....115
二、酶类型预测.....116
三、蛋白质翻译后修饰.....118
四、其他蛋白质功能预测问题.....119

第六章 基因芯片数据分析.....120

第一节 基因芯片平台简介.....120
一、基因芯片分类.....120
二、常见的基因芯片.....121
第二节 数据的预处理.....124
一、数据的提取.....124
二、数据的过滤.....125
三、缺失数据的处理.....125
四、数据的对数转换.....126
五、数据的标准化.....126
第三节 差异表达基因分析.....130
一、倍数法.....130
二、*t* 检验法.....130
三、SAM 法.....131
四、其他方法.....132
第四节 基因芯片数据的聚类分析.....132
一、聚类分析中距离（或相似性）的尺度函数.....133
二、聚类分析中的聚类算法.....134
第五节 基因芯片数据的分类分析.....137
一、*k* 近邻分类法.....138
二、决策树.....138
三、支持向量机.....140
四、Fisher 线性判别分析.....140
五、分类性能评价.....141
第六节 基因芯片数据库及常用分析软件.....142
一、基因表达数据库（Gene Expression Omnibus, GEO）.....142
二、ArrayTools.....143
三、SAM.....143
四、Cluster 和 TreeView.....143
五、R 语言和 BioConductor.....144

第七章　免疫信息学 145
第一节　免疫信息学源流 145
第二节　免疫信息学资源 147
一、免疫学数据库 147
二、单机软件与网络程序 151
第三节　免疫信息学的应用 151
一、表位预测 151
二、噬菌体展示 156
三、在抗体研究中的应用 160
四、在疫苗研究中的应用 164
五、在移植免疫中的应用 166
六、在变态反应防治中的应用 166

第一章 绪 论

生物信息学已成为生命科学领域重要的研究方向，随着高通量测序手段的不断进步，产生了海量的生物数据，生物信息学成为生物学家分析数据，解释实验结果，推断生物功能必不可少的手段。作为本书的第一章，将主要介绍生物信息学相关的背景知识，包括生物信息学的概念、起源、发展以及生物信息学研究的目的、内容，最后简单介绍本书的章节安排。

第一节 基本概念

什么是生物信息学？这是一个仁者见仁智者见智的问题。由于这个科学吸引了来自生物学、物理学、数学、化学、自动化、计算机等各个专业背景的科研人员，因此，从不同的研究背景给出了不同的定义。然而，由于各学科有不同的侧重点，因而给出的定义有一定的局限性。接下来，给出两个比较广泛使用的定义。

在维基百科中对生物信息学进行了如下定义："Bioinformatics is an interdisciplinary field that develops and improves on methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge."

在2000年7月NIH给出的定义："Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."

这里，基于我们自己的理解，给出一个比较概述性的定义："利用数学、物理、化学的理论、技术和方法，以计算机为工具，对生命现象加以研究，得到深层次的生物学知识。"

在许多实验生物学家的眼中，生物信息学是以开发生物数据分析工具为目的的，生物信息学仅以一种工程工具型的学科存在。而事实上，生物信息学之所以能独立成为一个崭新的、"独立发展"的学科，不仅在于它为实验学家提供了大量的实用工具，也在于生物信息学以揭示蕴藏在生物数据中的生物规律和内涵，给出生物的本质属性作为目标。生物信息学的任务是要搜集、储存和管理生物数据，并建立合适的算法来处理这些数据，在解释数据

背后真正的生物规律的同时，为新的生物学研究给予指导。

生物信息学的快速发展在科学发展的历史长河中有着举足轻重的作用。它是传统的实验生物学向理论生物学发展的重要一步，是从理论上认识生物的本质的必要途径，同时也为人类新一代的健康、医疗卫生的发展提供新的途径。

在生物信息学的发展过程中，还出现了一些与生物信息学类似的专有学科名词，一并列出，如下。

计算生物学，NIH 给出的定义："The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral and social systems." 计算生物学的研究重点更偏向算法的开发，与生物信息学经常混用。

理论生物学，顾名思义，就是所有非"湿"实验的研究都算作理论生物学，当然生物信息学可以算作理论生物学的一部分。当然，有许多生物信息学家认为生物信息学还是包括一些实验内容的，比如测序、chip-seq 等相关实验。在这里我们不做深究。

生物数学，利用数学的技巧和工具为生物学中的过程建模并进行分析。可以说生物信息学的大分部属于生物数学。

分子生物信息学，狭义的生物信息学，主要是研究 DNA 和蛋白质，也是本书着重讲述的内容。

第二节　生物信息学的发展

生物信息学的英文单词是 bioinformatics（bio+informatics），该单词最早出现的时间大约是在 20 世纪 80 年代。有资料报道著名的生物信息学家林华安（Hwa A. Lim）最早于 1987 年创造了 bioinformatics 一词，因而也被称为"生物信息学之父"。而荷兰理论生物学家 Paulien Hogeweg 在 PLoS Computational Biology 上发表了文章"The roots of bioinformatics in theroretical biology"，指出早在 1970 年，他就已经使用这词。但无论是谁最早创造了该单词，对于我们对知识的学习，已没有什么关系。

在国际上，生物信息学的研究可以追溯到 20 世纪 50 年代，我们称这个时期为生物信息学的储备时期。一些代表性的工作体现在 1951 年 Pauling 和 Corey 提出的蛋白质 α 螺旋和 β 折叠，以及在 1956 年 Watson 和 Crick 提出了 DNA 的双螺旋结构。1956 年，在美国的田纳西州盖特林堡召开的首次"生物学中的信息理论研讨会"上，产生了生物信息学的概念。

20 世纪 60～70 年代，是生物信息学发展的萌芽期，当时的研究热点聚

焦在序列比较方面。在这个时期，生物大分子携带信息成为生物学的重要理论，同源蛋白序列之间存在相似性引起了人们的注意。1962 年，Zucherkandl 和 Pauling 研究了序列变化与进化之间的关系，开创了一个新的领域——分子进化。1967 年，生物信息学家 Dayhoff（有人称她是生物信息学的鼻祖）搜集、构建了蛋白质家族序列数据，绘制了蛋白质序列图集，该图集后来演变为著名的蛋白质信息源 PIR（Protein Information Resource）。她在 1978 年对 71 个相关蛋白质家族的 1572 个突变进行研究，通过统计氨基酸的互相替换率得到序列比对中最广泛使用的矩阵之一——PAM 矩阵。Needleman 和 Wunsch 在 1970 年首次将动态规划引入到序列比对分析中，用于解决两条蛋白质氨基酸序列之间的全局比对问题，这就是著名的 Needleman-Wunsch 比对算法。尽管在过去的几十年中也开发了许多比对算法，但基于动态规划的比对算法依然是序列比对领域的“金标准”。

20 世纪 80 年代是生物信息学的形成期，例如专有 bioinformatics 的出现。而真正意义上的生物信息研究始于实验数据的共享和新算法。代表性的贡献有 EMBL、Genbank 和 DDBJ 三大分子数据库的国际合作，Smith 和 Waterman 在 1981 年开发的动态规划局部序列比对算法，用于识别两序列匹配的子序列，以及 Pearson 和 Lipman 于 1985 年开发的序列相似性搜索用具 FASTA。

生物信息学的高速发展得益于 1990 年启动的人类基因组计划。人类基因组计划是美国科学家于 1985 年率先提出的，后由美、英、法、德、中和日共同承担并完成了这份价值 30 亿美元的项目。我国承担了其中 1%的工作。该计划的目的是测定组成人类 24 条染色体中所包含的 30 亿个核苷酸序列的碱基组成。这一伟大工程与曼哈顿计划和阿波罗登月计划并称为三大科学计划。2001 年完成的人类基因组工作草图是人类基因组计划成功的里程碑，截至 2005 年，人类基因组计划的测序工作已基本完成。此外，在研究序列比对软件方面，也有了突破性的进展，时任 NCBI 主任的 Lipman 在 1990 年开发了目前使用最为广泛的生物信息学软件 BLAST。BLAST 能够方便快捷地实现两段核酸或者蛋白序列之间相似性的比较，快速地找到两段序列之间的同源序列，它已是整个生物信息学研究所必须掌握的一种工具。

国内几乎是同时开始着手生物信息学的研究的。国内最早的“准”生物信息学（称为理论生物学更贴切）研究主要表现在 20 世纪 70 年代与酶反应动力学相关的研究中，最具代表性的是对新的酶反应速率模型的提出，解决了底物分子还没有扩散到酶的活性中心处就会起反应的佯谬。当时，许多的物理、化学和数学学者投身于生物大分子的动力学和耗散结构研究。而“真正”的生物信息学的研究开始于 20 世纪的 80 年代，国内率先在中国科学院生物物理所、内蒙古大学、天津大学和云南大学等少数几个科研院所开展起来。直到 1996 年，我国第一个生物信息学中心（CBI）在北京大学成立，标志着中国生物信息学研究进入发展时期，然而，从事生物信息学的学者仍十

分稀少。1999 年华大基因研究中心正式成立，参与了人类基因组计划 1%的测序任务，随后又完成了水稻基因组计划、家蚕基因组计划等，大大地刺激了国内生物信息学的发展。此后，随着大量归国人员和年青一代的迅速成长，国内的生物信息学进入高速发展阶段。

第三节 研 究 任 务

科学研究有两个基本驱动力，一是实践需求驱动力，期望利用科技的发展改善人们的生活；二是好奇心驱动力，期望探索事物的本源，解决“为什么”。生物信息学是对生物数据进行处理，解释生物现象，其最终目标是揭示蕴藏在生物数据中的生物规律和内涵。因此，生物信息学不仅是要为人类的健康发展提供帮助，也要以解释“生命是什么”为最终目标。生物信息学作为技术，其任务表现在搜集和管理生物分子数据，对这些数据的处理与分析，最终提供实用性的研究工具。生物信息学作为科学，其任务表现在解释数据背后的生物机制，阐明涉及的生物内涵，给出生命的演化机理，最终给出生命的本质属性。

目前，尽管大量基因组已经测序，然而许多特殊要求的数据（如疾病相关数据）仍十分有限，因此也大大阻碍了生物信息学的发展。理论与实验相结合是生物信息学发展的必由之路，在相当长的一段时间里，在实验工作者眼里，生物信息学的研究是：Garbage in，Garbage out。然而，近些年，生物信息学提供的方法和观点正逐渐被实验生物学家所接受。事实上，无论是物理学还是化学的发展，都经历了从实验研究向理论研究再转向理论实验研究的阵痛。理论生物学的发展也不可避免，物理学的发展路线是否也能给生物学的发展提供一些启示呢（如图 1-1 所示）？

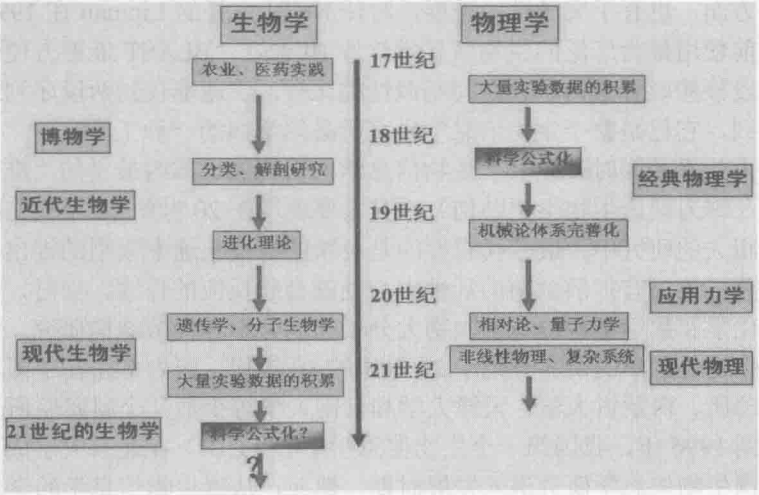

图 1-1 物理学的发展对生物学发展的启示

最后，希望用科学网上周达的话来结束该问题的探讨："也许在物理学家看来，数学家百分之九十九的工作都没有价值，但是就是那百分之一的工作，能引发物理科学的大踏步前进，而这些进步是本质的。"

第四节　章 节 安 排

本书的章节安排经过了多次讨论，制订了多种方案，最终采用以下形式。第一章为绪论，就是本章的内容，主要是对生物信息学的概念、发展以及任务做了阐述。第二章是经典的序列比对和进化问题，该章阐述了序列比对和分子进化的数学基础，而后给出了相关软件的使用，本章也是实验生物学家使用最多的部分。第三章和第四章是关于基因组方面的信息学，基因组信息学是生物信息学的龙头，也产生了大量的研究理论和方法，因此我们分用两章来讲解，前一章略偏理论基础，需要的数理知识较多，而后一章更偏重应用。第五章是蛋白质相关的信息学，主要涉及相关数据库和一些序列分析方法。第六章是对基因表达分析的讲述，涉及表达芯片的分析等内容。第七章是对免疫信息学做简单介绍，目的是给同学们一个概念，指出一个方向。

本书的撰写内容尽管不多，但已包括生物信息学的核心问题，适合于生物技术相关专业的低年级本科生学习。此外，本书也涉及一些前沿的理论生物学问题，适合高年级的同学和研究生作为参考书使用。

第二章　序列比对与分子进化

比较是科学研究中最常用的方法，通过比较研究对象在各个方面的相似程度，以确定同类事物可能具备的特征。本章将介绍生物信息学中最常用的比较方法——序列比对。序列比对是生物信息学的一个重要部分，在序列拼接、数据库搜索、分子进化和分子功能预测方面均有广泛的应用。本章首先将简要介绍序列比对所涉及的一些基本概念和算法；进而介绍序列比对在分子进化中的应用；最后，将主要介绍目前常用的序列比对工具和分子进化工具。

第一节　相似性得分系统

一、基本概念

生物序列通常是指 DNA、RNA 的碱基或蛋白质的氨基酸的排列顺序。对这些生物序列进行研究，可预测生物大分子的结构和功能，进而揭示生物调控、进化的内在机理，称为序列分析。序列分析问题涉及内容较多，主要包括序列比对、基因组序列分析、蛋白质序列分析和综合序列分析。其中，序列比对是基本但又十分重要的研究方法，在序列拼接、数据库搜索、功能预测和分子进化等方面是最重要的研究手段。

所谓序列比对，即将两条或多条 DNA 或蛋白质序列排列在一起，以一定的规则标明其相似之处。通过比对未知序列与已知序列的相似性，可以容易地预测出未知序列的功能。这种方法在大多数的情况下是成功的，然而，自然界还存在这样的情况，即两条序列的序列相似性很低，但其分子在空间的构象却十分相似，并可能拥有相似的生物功能，这就要求进行结构比对。进行序列比对的另一个目的是通过研究序列的相似性，判断序列间的同源关系。比对也是数据库搜索算法的基础。研究人员将研究序列与整个数据库的所有序列进行比对，从数据库中获得与其最相似序列的已有信息，通过这些已有信息推断该研究序列的可能的结构和功能。

在利用序列比对描述序列的进化关系时，会使用到“相似”（similarity）和“同源”（homology）这两个概念，这是两个极易被混淆的概念，但两者有一定的差别。相似是一个定量的概念，可以描述两条序列的相似性是 90% 或 60%；而同源是一个定性的概念，没有“度”的差别，两条序列要么同源，要么不同源。两者又互有联系，通常研究人员根据一定的标准来确定两条序

列是否同源，这个标准就是两条序列的相似性。一般情况下，如果两条序列的相似度达到 80%，基本可以确定它们是同源的；如果两条序列的相似度低于 25%，可以认为两者没有同源关系。接下来，我们将描述两条序列的相似性是如何确定的。

二、序列的相似性

相似性是定量描述两条序列的相似程度，根据计算规则的不同，可有不同的分值。根据不同的计算规则，相似性可有两种表达方式：编辑距离和相似性得分。

1．编辑距离

编辑距离可利用统计学中的距离公式或夹角余弦公式进行计算，常用的系数是：

欧氏距离

$$d_{ij}=\left[\sum_{k=1}^{n}(x_k^i-x_k^j)^2\right]^{1/2} \tag{2-1}$$

绝对值距离

$$d_{ij}=\sum_{k=1}^{n}|x_k^i-x_k^j| \tag{2-2}$$

马氏距离

$$d_{ij}=[(X_i-X_j)'C^{-1}(X_i-X_j)]^{1/2} \tag{2-3}$$

夹角余弦

$$r_{ij}=\frac{\sum_{k=1}^{p}x_k^i\times x_k^j}{\left[\sum_{k=1}^{p}(x_k^i)^2\sum_{k=1}^{p}(x_k^j)^2\right]^{1/2}} \tag{2-4}$$

四个公式均描述了两条序列之间的相似程度，对于公式(2-1)～式(2-3)，如果两条序列的距离越大，则两条序列越不相似；反之，亦然。然而，对于公式（2-4），其取值范围是[–1，1]，经过模标准化的内积数值在[0，1]之间，两条序列的夹角余弦值越趋近于 0，则两条序列越不相似；夹角余弦值越趋近于 1，则两条序列越相似。

编辑距离常用于衡量两条序列在组成上的相似性，例如可衡量两条蛋白质序列的 20 种氨基酸含量的相似性；进化树软件 CVTree 就是利用夹角余弦衡量两序列的 k 字串（k-string，k-mer，k-tuple，k-gram）的相似性，进而构建进化树。此外，还有其他可用于衡量序列相似性的编辑距离方法，主要有互信息、模糊数学理论、图论等。编辑距离可用于衡量序列的相似性，但并不是序列比对的理论基础。它主要用于从序列组成或其他信息方面构建进化树，也可用于生物数据预测分类问题，但没有在数据库搜索方面进行应用。

2．相似性得分

相似性得分是通过某种记分规则计算两条序列的相似性。记分规则是字符间两两比较的分值，与字符的位置无关。通过比较两条序列相应位置的字符，累加每个位置的得分即为两序列的相似性分值。图 2-1 给出一个计算两条 DNA 序列相似性得分的例子。从图中可以看出，使用不同的计分规则，可以获得不同的分值。另外，由于序列的差异包括替换、插入/缺失，后两者需在比对时引入空格。这种利用积分规则，考虑序列替换、插入/缺失的相似性得分即为序列比对。接下来的两部分将分别讨论序列比对的计分规则和如何考虑最优比对。

			未引入插入/缺失					引入插入/缺失					
			A	T	A	T	C	A	T	A	T	C	
			A	T	T	C	G	A	T		T	C	G
打分规则 1：	$p(a, a)=1$	位点得分	1	1	0	0	0	1	1	0	1	1	0
	$p(a, b)=0\ (a\neq b)$	总得分			2						4		
打分规则 2：	$p(a, a)=5$	位点得分	5	5	−4	−4	−4	5	5	−4	5	5	−4
	$p(a, b)=-4\ (a\neq b)$	总得分			−2						12		

图 2-1　不同规则相似性打分（无对应字符按照两字符不同处理）

三、序列比对得分矩阵

人们提出不同的打分规则来进行不同目的的序列比对研究。不同类型的字符替换，其代价或得分是不一样的。例如，在 DNA 序列中，A↔G，C↔T 比 A↔T，C↔T 嘧啶与嘧啶间或嘌呤与嘌呤间更容易发生替换。通常，在进行 DNA 序列比对时，得分系统较为简单；但蛋白质序列比对的得分系统相对复杂，这是由于相似理化性质的氨基酸替换通常不显著改变蛋白质的结构和功能所致。此外，如果两条序列的某一条在进化中发生了插入或缺失，还需要在比对时在相应位置引入空位，为补偿插入和缺失对序列相似性的影响，字符与空位的比对分值即为空位罚分。

1．核酸得分矩阵

①等价矩阵（unitary/identity matrix）

等价矩阵（如图 2-2a 所示）是最简单的一种打分矩阵。相同核苷酸的匹配分值为“1”，不同核苷酸的比对分值为“0”。由于该打分矩阵并未考虑核苷酸间理化性质的相似性，故很少使用。

②BLAST 矩阵（BLAST matrix）

BLAST 矩阵（如图 2-2b 所示）是目前最流行的打分矩阵。经大量实际比对发现，如果令两相同的核苷酸的得分为+5，不相同的核苷酸得分为–4，比对效果最好。

③转换/颠换矩阵（transversion/transition matrix）

四种核苷酸根据其化学结构，可分为嘌呤和嘧啶两类。嘌呤包括腺嘌呤 A 和鸟嘌呤 G，结构上都包括两个环；嘧啶包括胞嘧啶 C 和胸腺嘧啶 T，结构上都包括一个环。在立体化学上，相似的结构间更容易发生替换，因此，把嘧啶与嘧啶或嘌呤与嘌呤间的替换称为转换，把嘧啶与嘌呤间的替换称为颠换。研究发现，生物体转换发生的频率远高于颠换发生的频率。图 2-2c 给出了该矩阵的打分值。

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

a. 等价矩阵

	A	C	G	T
A	5	–4	–4	–4
C	–4	5	–4	–4
G	–4	–4	5	–4
T	–4	–4	–4	5

b. BLAST 矩阵

	A	C	G	T
A	1	–5	–1	–5
C	–5	1	–5	–1
G	–1	–5	1	–5
T	–5	–1	–5	1

c. 转换/颠换矩阵

图 2-2 核酸替换矩阵

④核苷酸 PAM 矩阵（PAM matrix）

States 等（1991 年）为提高序列数据库的相似性搜索的灵敏度，利用马尔科夫模型建立了核苷酸 PAM 矩阵。尽管该矩阵式建立在一定的进化模型上，但在实际 DNA 序列比对时，很少使用。

2．蛋白质得分矩阵

蛋白质得分矩阵较核苷酸矩阵更为复杂。组成蛋白质的 20 种氨基酸具有不同的物理化学性质，通常性质相似的氨基酸间更容易发生替换。基于不同的标准，已有很多氨基酸替换矩阵。以下均使用单字母法表示氨基酸。

①等价矩阵（unitary/identity matrix）

与核苷酸的等价矩阵类似，相同的氨基酸间替换分值为 1，不相同的氨基酸间替换分值为 0，该矩阵使用很少。