上　篇

多序列比对基础篇

第 1 章　生物多序列比对

1.1　生物信息学

1.1.1　生物信息学的起源

自从 1990 年美国启动人类基因组计划以来，人与模式生物基因组的测序工作进展极为迅速。迄今已完成了约 40 多种生物的全基因组测序工作，人基因组约 3×10^9 个碱基对的测序工作也接近完成。至 2000 年 6 月 26 日，被誉为生命“阿波罗计划”的人类基因组计划，经过美、英、日、法、德和中国科学家的艰苦努力，终于完成了工作草图，这是人类科学史上又一个里程碑式的事件，它预示着完成人类基因组计划已经指日可待。截至目前，仅登录在美国 GenBank 数据库中的 DNA 序列总量已超过 70 亿个碱基对。在人类基因组计划进行过程中所积累起来的技术和经验，使得其他生物基因组的测序工作可以完成得更快捷。可以预计，今后 DNA 序列数据的增长将更为惊人。生物学数据的积累并不仅仅表现在 DNA 序列方面，与其同步的还有蛋白质的一级结构，即氨基酸序列的增长。此外，迄今为止，已有 10 000 多种蛋白质的空间结构以不同的分辨率被测定。基于 cDNA 序列测序所建立起来的 EST 数据库，其记录已达数百万条。在这些数据基础上派生、整理出来的数据库已达 500 余个。这一切构成了一个生物学数据的海洋。可以打一个比方来说明这些数据的规模。有人估计，人类(包括已经去世的和仍然在世的)所说过的话的信息总量约为 5EB(1EB=10^{18}B)，而如今生物学数据信息总量已接近甚至超过此数量级。这种科学数据的急速和海量积累，在

人类的科学研究历史中是空前的。

数据并不等于信息和知识，但却是信息和知识的源泉，关键在于如何挖掘它们。与正在以指数方式增长的生物学数据相比，人类相关知识的增长(粗略地用每年发表的生物、医学论文数来代表)却十分缓慢。一方面是巨量的数据；另一方面是我们在医学、药物、农业和环保等方面对新知识的渴求，这些新知识将帮助人们改善其生存环境和提高生活质量。这就构成了一个极大的矛盾。这个矛盾就催生了一门新兴的交叉科学，这就是生物信息学。

1.1.2　生物信息学的概念

美国人类基因组计划实施五年后的总结报告中，对生物信息学做了以下定义：生物信息学是一门交叉科学，它包含了生物信息的获取、处理、存储、分发、分析和解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具，来阐明和理解大量数据所包含的生物学意义。生物信息学这一名词的出现仅仅是几年前的事情，但是计算生物学这一名词的出现要早得多。鉴于这两门学科之间并没有或难以界定严格的分界线，在这里统称为生物信息学。它是当今生命科学和自然科学的重大前沿领域之一，同时也是 21 世纪自然科学的核心领域之一。其研究重点主要体现在基因组学(genomics)和蛋白组学(proteomics)两方面，具体说就是从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。

1.1.3　生物信息学的主要研究内容

生物信息学主要包括以下几个主要研究领域，但是限于篇幅，这里仅列出其名称并只做简单介绍。

1. 序列比对(alignment)

基本问题是比较两个或两个以上符号序列的相似性或不相似性。

序列比对是生物信息学的基础，非常重要。两个序列的比对有较成熟的动态规划算法，以及在此基础上编写的比对软件包—— BLAST 和 FASTA，可以免费下载使用。这些软件在数据库查询和搜索中有重要的应用。有时两个序列总体并不很相似，但某些局部片断相似性很高。Smith-Waterman 算法是解决局部比对的好算法，缺点是速度较慢。两个以上序列的多重序列比对目前还缺乏快速而又十分有效的算法。

2. 蛋白质三级结构比对

蛋白质三级结构比对是生物信息学的重要研究领域。蛋白质的功能由蛋白质的三级结构决定，蛋白质三维空间结构的相似性比较是分析蛋白质结构和功能的重要手段，因此比较蛋白质的三级结构可以了解它们之间的相互作用和进化关系。

研究蛋白质的结构意义重大，分析蛋白质结构、功能及其关系是蛋白质组计划中的一个重要组成部分。研究蛋白质结构有助于了解蛋白质的功能，了解蛋白质如何行使其生物功能，认识蛋白质与蛋白质或其他分子之间的相互作用，这无论是对于生物学还是对于医学和药学都是非常重要的。对于未知功能或者新发现的蛋白质分子，通过结构分析可以进行功能注释、指导设计进行功能确认的生物学实验。通过分析蛋白质的结构确认功能单位或者结构域，可以为遗传操作提供目标，为设计新的蛋白质或改造已有蛋白质提供可靠的依据，同时为新的药物分子设计提供合理的靶分子结构。

3. 蛋白质二级和三级结构预测

从方法上来看，有演绎法和归纳法两种途径。前者主要是从一些基本原理或假设出发来预测和研究蛋白质的结构和折叠过程，分子力学和分子动力学属这一范畴。后者主要是从观察和总结已知结构的蛋白质结构规律出发来预测未知蛋白质的结构，同源模建和指认(threading)方法属于这一范畴。虽然经过 30 余年的努力，蛋白质结构预测研究现状还远远不能满足实际需要。

4. 计算机辅助基因识别(仅指蛋白质编码基因)

基本问题是给定基因组序列后，正确识别基因的范围和在基因组序列中的精确位置。这是最重要的课题之一，而且越来越重要。经过 20 余年的努力，提出了数十种算法，有 10 种左右重要的算法和相应软件(网上提供免费服务)。原核生物计算机辅助基因识别相对容易些，结果好一些。从具有较多内含子的真核生物基因组序列中正确识别出起始密码子、剪切位点和终止密码子，是一个相当困难的问题，研究现状不能令人满意，仍有大量的工作要做。

5. 非编码区分析和 DNA 语言研究

在人类基因组中，编码部分只占总序列的 3%～5%，其他通常称为“垃圾”DNA，其实一点也不是“垃圾”，只是暂时还不知道其重要的功能。分析非编码区 DNA 序列需要大胆的想象与崭新的研究思路和方法。DNA 序列作为一种遗传语言，不仅体现在编码序列之中，而且隐含在非编码序列之中。

6. 分子进化和比较基因组学

早期的工作主要是利用不同物种中同一种基因序列的异同来研究生物的进化，构建进化树。既可以用 DNA 序列也可以用其编码的氨基酸序列来做，甚至可通过相关蛋白质的结构比对来研究分子进化。以上研究已经积累了大量的工作。近年来由于较多模式生物基因组测序任务的完成，为从整个基因组的角度来研究分子进化提供了条件。可以设想，比较两个或多个完整基因组这一工作需要新的思路和方法，当然也渴望得到更丰硕的成果。这方面可做的工作是很多的。

7. 序列重叠群(contigs)装配

一般来说，根据现行的测序技术，每次反应只能测出 500 个或更多一些碱基对的序列，这就有一个把大量的较短的序列全体

构成重叠群，逐步把它们拼接起来形成序列更长的重叠群，直至得到完整序列的过程，称为重叠群装配。拼接 EST 数据以发现全长新基因也有类似的问题。已经证明，这是一个 NP 完备性算法问题。

8. 遗传密码的起源

遗传密码为什么是现在这样的？这一直是一个谜。一种最简单的理论认为，密码子与氨基酸之间的关系是生物进化历史上一次偶然的事件造成的，并被固定在现代生物最后的共同祖先里，一直延续至今。不同于这种“冻结”理论，有人曾分别提出过选择优化、化学和历史等三种学说来解释遗传密码。随着各种生物基因组测序任务的完成，为研究遗传密码的起源和检验上述理论的真伪提供了新的素材。

9. 基于结构的药物设计

人类基因组计划的目的之一在于阐明人的约 10 万种蛋白质的结构、功能、相互作用以及与各种人类疾病之间的关系，寻求各种治疗和预防方法，包括药物治疗。基于生物大分子结构的药物设计是生物信息学中的极为重要的研究领域。为了抑制某些酶或蛋白质的活性，在已知其三级结构的基础上，可以利用分子对接算法，在计算机上设计抑制剂分子，作为候选药物。这种发现新药物的方法有强大的生命力，也有着巨大的经济效益。

10. 代谢网络的分析

代谢网络涉及生化反应途径、基因调控及信号转导过程(蛋白质间的作用)等。后基因组时代将研究大规模网络的生命过程，称为“网络生物学”研究。与代谢分析直接相关的便是系统生物学研究，它将是后基因组时代最为突出的研究方向。

11. 其他

如基因表达谱分析、基因芯片设计和蛋白质组学数据分析等，逐渐成为生物信息学中新兴的重要研究领域。

1.2　序列比对的概念及其发展历史

1.2.1　序列比对的提出与基本概念

随着人类基因组计划的实施和科技的发展，生物学数据呈爆炸式增长，这些海量的生物学数据必须通过生物信息学手段进行收集、分析和整理后，才能成为有用的信息。而如何有效分析和处理这些大型序列数据(即序列分析)成为生物信息学的首要任务。序列比对是生物序列分析的主要方法，也是生物信息学中挑战性的问题之一。序列比对在序列装配、序列注释、基因和蛋白质的结构和功能预测以及系统发育和进化分析等方面均有广泛应用，因此对它的研究一直以来都是热点。

序列比对就是在两个或更多序列的相同区域寻找最大相似性的任务，其基本思想是找出检测序列和目标序列的相似性。比对过程中需要在待比对序列中引入空位，以表示插入或删除。根据需要比对的序列个数，序列比对分为双序列比对和多序列比对。

序列比对问题就是通过向待测序列中插入空格，使得每条序列的长度一样，并尽可能保证每列的字符具有最大的相似性。随着参与比对序列的条数增加，比对的难度及复杂度急剧增加。

1.2.2　序列比对的目的和意义

当所考察的序列不同时，保守的残基往往是维持稳定结构或生物学功能的关键残基。多序列比对可以揭示关于蛋白质结构和功能

的许多线索。

序列比对的目的是发现相似的序列，得到保守的区域，寻找序列之间的功能、结构或进化上的关系。

序列比对的意义如下：

(1) 用于描述一组序列之间的相似性关系，以便了解一个基因家族的基本特征，寻找 motif、保守区域等。

(2) 用于描述一个同源基因之间的亲缘关系的远近，应用到分子进化分析中。

(3) 定量估计序列间的关系，并由此推断它们在进化中的亲缘关系，可以通过计算完全匹配的残基数目或计算完全匹配残基和相似残基的数目得到这种定量关系。该方法还可以用来评估比对质量。

(4) 对于系统发育，相等的残基相当于具有共同的进化祖先；对于结构生物学，相等的残基与一组蛋白质中同源折叠的类似位置相关；对于分子生物学，相等的残基在其相应的蛋白质有同类功能作用。在每一种情况下，比对提供了潜在的演化、结构，或简洁直观地表达蛋白质家族功能限制表征的鸟瞰视图。

(5) 如果是对多个蛋白质或核酸同时进行比较分析，就有可能寻找到这些有进化关系的序列之间共同的保守区域、位点和 profile，从而就能够探索到导致它们产生共同功能的序列模式。

(6) 序列比对在预测和治疗疾病方面有着非常重要的应用，如白血病。大量的试验数据和临床数据表明，DNA 序列中包含的一些基因诱发了白血病，如由原癌基因的转录因子的不适当表达和异常表达造成。于是可以通过比对一些检验者的 DNA 序列数据和某些特殊的序列数据，之后通过比对结果就可以从检验者 DNA 序列数据中找出具有生物意义的特征，从而提供一些有价值的信息，这些信息为诊断白血病提供了一定的依据，从而对白血病的治疗方面提供一些指导性的作用。

1.2.3　国内外研究现状

多序列比对(multiple sequence alignment，MSA)问题是生物信息学中一个尚未解决的问题，被列为一个 NP 完全的组合优化问题，想要找到复杂性为多项式的精确算法是不可能的。多序列比对是生物序列分析的主要方法，也是生物信息学中挑战性的问题之一，Chuong B, Do 和 Katoh K(2008)在文中引用 258 个文献对多序列比对做了非常全面的综述，根据多序列比对的研究现状得出结论：虽然多序列比对问题已经研究了几十年，但是多序列比对的研究一直蓬勃发展，每一年国内外都有数十个描述多序列比对新方法的文章发表[如 McClure M A, et al.(1994), JD Notredame C(2002), Julie D. Thompson(2005), Chuong B Do, et al.(2008), C Kemena, et al.(2009), Orobitg Miquel, et al.(2013), 张鹏帅&霍红卫(2005), 葛宏伟&梁艳春(2006), 杨凡&朱清新等(2010)]。许多最近的研究表明，在提高比对的精度和比对处理范围可扩展性方面都取得相当大的进展，且在该方面仍有很大的发展空间。

1.2.4　多序列比对面临的挑战

同源性分析中常常要通过多序列比对来找出序列之间的相互关系。和 BLAST 的局部匹配搜索不同，多序列比对大多采用全局比对的算法。这样对于采用计算机程序的自动多序列比对是一个非常复杂且耗时的过程，特别是序列数目多且序列长的情况下。

多序列比对是一个 NP 完全问题，解决此问题的传统算法是渐进算法或迭代算法，但是随着序列长度和条数的增多，时空复杂性急剧上升，设计一个具有高敏感性、高精度且低复杂度的算法，成为解决生物多序列比对的瓶颈问题。

1.3　多序列比对的基本原理

1.3.1　多序列比对的相关概念

数据库搜索的基础是序列的相似性比对，而寻找同源序列则是数据库搜索的主要目的之一。在序列比对中需要注意区分以下几个基本概念。

1. 相似性、同源性

相似性(similarity)是指序列比对过程中用来描述检测序列和目标序列之间相同 DNA 碱基或氨基酸残基顺序所占比例的高低。例如，A 序列和 B 序列的相似性是 80%，这是个量化的关系。

同源性(homology)是指从一些数据中推断出的两个基因或蛋白质序列具有共同祖先的结论。例如，A 序列和 B 序列只有是同源序列或非同源序列两种关系，属于质的判断。

相似性和同源性是两个完全不同的概念。序列之间的相似程度是可以量化的参数，而序列是否同源则需要有进化事实的验证。任何序列之间均存在相似，只有当序列是从一个共同祖先进化分歧而来的，它们才是同源的。因此，相似的序列有可能同源，同源的序列也常常具有相似的生物学功能，但基因复制机制又使得同源序列进化出不同的功能。一般来说，序列间的相似性越高，它们是同源序列的可能性就越大，当相似度高于 50%时，比较容易推测检测序列和目标序列可能是同源序列；而当相似度低于 20%时，就难以确定或者根本无法确定其是否具有同源性，所以常常通过序列的相似性来推测序列是否同源。

(1) 序列相似性比较：将待测序列与 DNA 或蛋白质序列库进行比较，用于确定该序列的生物属性，也就是找出与此序列相似的已知序列是什么。完成这一工作只需要使用两两序列比较算法，常用的程序包括 BLAST、FASTA 等。

(2) 序列同源性分析：将待测序列加入到一组与之同源但来自不同物种的序列中进行多序列同时比较，以确定该序列与其他序列间的同源性大小。这是理论分析方法中最关键的一步。完成这一工作必须使用多序列比对算法，常用的程序包括 Clustal、MAFFT 等。

2. 直系同源、旁系同源

在考虑序列相似性时，还必须认识另外两个概念：直系同源(ortholog)和旁系同源(paralog)。分子进化不仅使不同的物种发生差异，在某个物种的基因组内部也可能发生进化事件。来自共同祖先的基因成为同源基因。

直系同源基因是指在不同物种中有相同功能的同源基因，它是在物种形成过程中形成的。旁系同源基因是指一个物种内的同源基因。一般情况下，一个生物物种的基因组中，两个基因或可读框在各自全长的 60%以上范围内，同一性不少于 30%时，称为同源基因。研究直系同源基因之间或旁系同源基因之间的功能关系，可以为基因组分析提供很大的帮助。直系同源基因由共同的祖先演化而来，从而具有序列相似性。而旁系同源是种内基因倍增的结果。当序列相似性高时，直系同源可以暗示功能性同源，而旁系同源一般会有相似但并不相同的功能。

1.3.2　序列比对的分类

1. 按比对的序列数量可分为双序列比对和多序列比对

双序列比对就是对两条序列进行比对，通过比较两个序列之间的相似区域和保守性位点，寻找二者可能的分子进化关系。双序列比对是序列分析的基础。然而，对于构成基因家族的成组的序列来说，要建立多个序列之间的关系，这样才能揭示整个基因家族的特征。

多序列比对就是对三条及以上的序列进行比对。多序列比对可用来区分一组序列之间的差异，但主要是用来描述一组序列之间的相似

性关系，以便对一个基因家族的特征有一个简明扼要的了解，在阐明一组相关序列的重要生物学模式方面起着相当重要的作用。

2. 按比对的序列长度可分为短序列比对和长序列比对

序列长度单位通常定义为 bp，一般规定长度不超过 100bp 的序列为短序列，超过 100bp 的序列为长序列。不同类型的测序数据适用于特定的实验应用领域：近似 100bp 长的双末端序列适合重测序研究；染色质免疫沉淀测序用于研究转录因子结合位点和组蛋白修饰位点，其序列长度不超过 50～75bp；对于基因组序列未知的物种测序，短序列组装通常是最基本的分析步骤，通过将短序列定位到参照基因组序列上进行后续分析。长序列比对对于特定基因组学问题来讲仍然是非常有意义的，例如基因组较大的物种基因组序列从头组装、重测序和结构变异检测等，目前也已经有很多小组基于 PacBio 的测序数据开展了一些相关研究并取得了非常好的效果。

3. 按比对的序列范围可分为全局比对和局部比对

全局比对考虑序列的全局相似性，是对给定序列全长进行比较的方式。运用全局比对的主要优势是对具有高度同源性的序列进行优化，这在以已知三维结构的同源性序列为基础对未知序列的三维结构进行预测的模型构建过程中是很有用的。

局部比对考虑序列片段之间的相似性，仅能获得特定序列在数据库中配对好的亚区。局部比对适合于那些在全长中具有局部的小同源性片段的序列比较，一般用于特定序列位点、结构域及其他类型重复序列的搜索，同时在发现数据库中待分析序列的同源序列过程中也具有重要意义。

1.3.3　多序列比对的数学定义

一条长度为 T 的序列是 T 个字符组成的字符串，字符取自于字

母表$\sum=\{A,G,C,T\}$，分别代表 DNA 的四个核苷酸类型。对于 DNA 序列，给定包含 N 个序列的序列集 $S=\{S_1,S_2,\cdots,S_N\}$，$N\geqslant 2$，$S_i=S_{i1}S_{i2}\cdots S_{il_i}(1\leqslant i\leqslant N)$，$S_{ij}\in\sum(1\leqslant j\leqslant l_i)$，$l_i$ 是第 i 条序列的长度，则一个序列比对可定义为一个矩阵 $\boldsymbol{A}=(a_{ij})$，其中 $1\leqslant i\leqslant N$，$1\leqslant j\leqslant l$，$\max(l_i)\leqslant l\leqslant\sum_{i=1}^{N}l_i$，如图 1.1 所示。矩阵必须满足下列三个条件：

(1) $a_{ij}\in\sum\cup\{-\}$，其中“－”代表空位。

(2) 矩阵中的第 i 行去掉“－”后，即得到字符串 S_i。

(3) 矩阵中不包含字符全是空格的列。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
I	Y	Y	D	G	G	A	V	—	E	A	L	C	A	M
II	Y	D	D	—	G	A	L	V	E	A	L	C	A	M
III	F	D	E	G	G	—	L	V	Q	A	—	C	F	—
IV	F	Y	E	G	G	I	V	V	Q	A	V	—	—	M
V	Y	D	D	G	G	I	L	V	—	—	L	C	A	N
VI	Y	Y	E	—	G	A	—	V	Q	A	V	C	E	M

图 1.1　多序列比对

1.3.4　多序列比对的打分方法

序列分析的目的是揭示核苷酸或氨基酸序列编码的高级结构或功能信息，而序列比对打分的目的则是提供一个比较两(多)序列之间相似性的量度，从而使人们有可能迅速区分有着微妙不同的任意两个序列的比对结果。常见的打分标准有以下几种。

1. 目标函数

目标函数是用来考查多序列比对结果好坏的一种度量标准，所有的多序列比对方法都依赖于一个目标函数来说明比对结果的好坏，从而反映出此方法的精确度和有效性。当前有三种主流的目标函数：比对和函数(sum-of-pairs functions)。一致性函数(consensus

functions)和树函数(tree functions)，其中使用最普遍的是比对和函数(简称为 SP 函数)。SP 函数需要设置两个重要的参数：替换矩阵(substitution matrix)和空位罚分(gap penalties，其中包括起始空位罚分和延续空位罚分)。

sum-of-pair(SP)函数的计算公式均可用 $score=\sum residue-\sum penalty$ 表示。其中，*score* 定义为正数，分值越高，比对效果越好；$\sum residue$ 是比对后的蛋白质序列中氨基酸残基的总分，定义为 $\sum residue>0$；$\sum penalty$ 是插入空格产生的总罚分，定义为 $\sum penalty>0$。

氨基酸残基的总分公式为

$$\sum residue=\sum_{h=1}^{L}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}cost(S_{ih},S_{jh})$$

式中：

$$cost(S_{ih},S_{jh})=\begin{cases}S_{aa}=score(a,a) & \text{如果两个非空位残基相同(匹配)；}\\ S_{ab}=score(a,b) & \text{如果两个非空位残基不同(不匹配)；}\\ S_{a-}=score(a,-)=0 & \text{残基与空位的分数为0(空位罚分另计)。}\end{cases}$$

式中：L 为比对序列长度；m 为参加比对的序列条数；S_{ih} 为第 i 条序列第 h 个残基。匹配和不匹配的分数通常由替换计分矩阵给出。

2. 替换计分矩阵

对于蛋白质序列，计分矩阵主要用于记录在做序列比对时两个相对应的残基的相似度。一旦这个矩阵定义好了以后，比对程式就可以利用这个矩阵，尽量将相似的残基排在一起，以达到最好的比对。常用的替换计分矩阵有可接受点突变矩阵(point accepted mutation，PAM)和模块替换矩阵(blocks substitution matrix，BLOSUM)。

1) PAM 矩阵

基于进化的点突变模型，如果两种氨基酸替换频繁，说明自然界接受这种替换，那么这对氨基酸替换得分就高。一个 PAM 就是一

个进化的变异单位，即 1%的氨基酸改变，但这并不意味 100 次 PAM 后，每个氨基酸都发生变化，因为其中一些位置可能会经过多次突变，甚至可能会变回到原来的氨基酸。对应于一个更大进化距离间隔的突变概率矩阵，可以通过对初始矩阵进行适当的数学处理得到(Dayhoff 等，1978)，如常用的 PAM250 矩阵。PAM250 相似性分数矩阵相当于在两个序列之间具有 20%的残基匹配，如图 1.2 所示。主对角线上分数值是指两个相同残基之间的相似性分数值，有些残基的分值较高，如色氨酸 W 为 17，半胱氨酸 C 为 12，说明它们比较保守，不易突变；有的残基的分值较低，如丝氨酸 S、丙氨酸 A、门冬酰胺 N 三种氨基酸均为 2，这些氨基酸则比较容易突变。不同氨基酸之间的分数值越高，它们之间的相似性越高，进化过程中容易发生互相突变，如苯丙氨酸 F 和酪氨酸 Y，它们之间的相似性分数值是 7；而相似性分数值为负数的氨基酸之间的相似性则较低，如甘氨酸 G 和色氨酸 W 之间为−7，它们在进化过程中不易发生互相突变。此外，图 1.2 所示矩阵中把理化性质相似的氨基酸按组排列在

C Cys	12																		
S Ser	0	2																	
T Thr	-2	1	3																
P Pro	-3	1	0	6															
A Ala	-2	1	1	1	2														
G Gly	-3	1	0	-1	1	5													
N Asn	-4	1	0	-1	0	0	2												
D Asp	-5	0	0	-1	0	1	2	4											
E Glu	-5	0	0	-1	0	0	1	3	4										
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4									
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y Tyr	0	-3	-3	-5	-3	-5	-2	4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

图 1.2　PAM250 相似性分数矩阵

一起，如碱性氨基酸组氨酸 H、精氨酸 R 和赖氨酸 K。突变数据矩阵的产生基于相似性较高(通常为 85%以上)的序列比对，那些进化距离较远的矩阵(如 PAM250)是从初始模型中推算出来而不是直接计算得到的，其准确率受到一定限制。序列分析的关键是检测进化距离较远的序列之间是否具有同源性，因此突变数据矩阵在实际使用时存在一定的局限性。

针对不同的进化距离采用 PAM 矩阵：

序列相似度　=　40%　50%　60%

打分矩阵　=　PAM120　PAM80　PAM60

PAM250→14%～27%

2) BLOSUM 矩阵

模块替换矩阵(BLOSUM)以序列片段为基础，基于蛋白质模块数据库 BLOCKS。Henikoff 夫妇(Henikoff 和 Henikoff，1992)从蛋白质模块数据库 BLOCKS 中找出一组替换矩阵，用于解决序列的远距离相关。在构建矩阵过程中，通过设置最小相同残基数百分比将序列片段整合在一起，以避免由于同一个残基对被重复计数而引入的任何潜在的偏差。在每一片段中，计算出每个残基位置的平均贡献，使得整个片段可以有效地被看作单一序列。通过设置不同的百分比，产生了不同矩阵。由此，高于或等于 80%相同的序列组成的串可用于产生 BLOSUM80 矩阵；那些有 62%或以上相同的串用于产生 BLOSUM62 矩阵，依此类推。BLOSUM 与 BLOCKS 对于同样的序列比对产生的结果在局部有所不同，可能是一个认为不相似不可以替换，而另一个认为相似可以替换。必须说明，如果比对这两个序列高度相似，这些细微的差别对整个序列比对结果的影响不大，但在序列比对的边界区可能产生显著影响，此时增强微弱信号以探测远距离相关变得十分重要。

此矩阵与 PAM 矩阵的不同之处在于：

(1) 用于产生矩阵的蛋白质家族及多肽链数目，BLOSUM 比

PAM 大约多 20 倍。

(2) PAM：家族内成员相比，然后把所有家族中对某种氨基酸的比较结果加和在一起，产生“取代”数据(PAM-1)，PAM-1 自乘 *n* 次，得 PAM-*n*；BLOSUM：首先寻找氨基酸模式，即有意义的一段氨基酸片断(如一个结构域及其相邻的两小段氨基酸序列)，分别比较相同的氨基酸模式之间氨基酸的保守性(某种氨基酸对另一种氨基酸的取代数据)，然后，以所有 60%保守性的氨基酸模式之间的比较数据为根据，产生 BLOSUM-60；以所有 80%保守性的氨基酸模式之间的比较数据为根据，产生 BLOSUM-80。

(3) PAM-*n* 中，*n* 越小，表示氨基酸变异的可能性越小。相似的序列之间比较应该选用 *n* 值小的矩阵，不太相似的序列之间比较应该选用 *n* 值大的矩阵。PAM-250 用于约 20%相同序列之间的比较。BLOSUM-*n* 中，*n* 越小，表示氨基酸相似的可能性越小。相似的序列之间比较应该选用 *n* 值大的矩阵，不太相似的序列之间比较应该选用 *n* 值小的矩阵。BLOSUM-62 用来比较 62%相似度的序列，BLOSUM-80 用来比较 80%左右的序列。

BLOSUM-62 矩阵如图 1.3 所示。

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	-2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

图 1.3　BLOSUM-62 计分替换矩阵

应用 BLOSUM-62 矩阵(为示例表述清晰，只截取部分矩阵内容)计算双序列比对残基分数的示例见图 1.4。

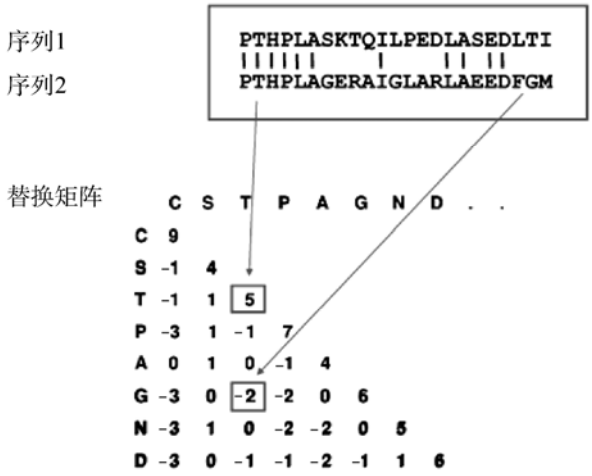

图 1.4　残基分数计算图

在这个例子中，共有两条序列，即 m=2，每个序列有 23 个残基，且长度相等，即 L=23，对照图 1.3 所示矩阵，有 $S_{\mathrm{TT}}=5$，$S_{\mathrm{TG}}=-2$ 等比对数据，则

$$\sum residue=\sum_{h=1}^{L}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}cost(S_{ih},S_{jh})=\sum_{h=1}^{23}\sum_{i=1}^{1}\sum_{j=2}^{2}cost(S_{ih},S_{jh})=48$$

3. 空位罚分

在比对过程中需要在检测序列或目标序列中引入空位(gap)，以表示插入和删除。一般情况下，参与比对的多条序列不完全相同，为了将其对齐，就需要插入空位。为了使整条序列而不仅仅是空位插入区域产生可比较的模式，空位的插入是必需的。不插入空位，序列比对过程就无法进行。序列对齐后，才能提出最初的同源性假设，才能确定插入缺失置换(transition)和颠换(transversion)等事件是

否发生以及发生位置和发生频率。

在多序列比对阶段，如果选择插入大量的空位，那么任意两条随机的不相关的序列都可以对齐，但是这样的比对结果可能毫无生物学意义，因此必须在一定程度上限制空位的数目以产生有生物学意义的比对结果，为此采用一个打分策略：匹配的残基获得正的分值，而空位获得负的分值或者叫罚分(Hall，2001)，这样获得最大净分值的比对结果即为比对程序所寻找的最优结果。

空位罚分包括两部分：起始空位罚分和延伸空位罚分。一般起始空位罚分高于延伸空位罚分。所谓起始空位，是指序列比对时在某一序列中插入的第一个空位。所谓延伸空位，是指在引入一个或几个空位后，继续引入下一个连续的空位。空位罚分 $\sum penalty$ 有以下两种计算方法。

1) 线性空位罚分

线性空位罚分(constant gap penalty)是最简单的罚分方式，对于每一个空位罚同样的分数，则总空位罚分的计算公式为 $\sum penalty = n \times gap$ 。其中，n 是空位的个数；gap 是空位的罚分。

对于图 1.5 示例，假设空位罚分为 5，则该例的总空位罚分是 11×5=55 分。

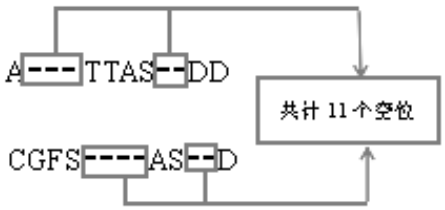

图 1.5　线性空位罚分计算

2) 仿射空位罚分

大多数的比对算法采用仿射空位罚分策略，这种策略设定起始空位罚分加权起始空位，并以一个较小的延伸空位罚分来加权延伸

空位，起始空位罚分(gap open penalty)和延伸空位罚分(gap extend penalty)的值域分别为 0～100 和 0～10。研究表明，增加起始空位罚分会使插入空位频率降低，增加延伸空位罚分则使空位变短。起始空位后的延伸空位被赋予较小的罚分值，用以鼓励长的且更加连续的空位而不是大量单个空位的插入。

起始空位罚分和延伸空位罚分的分值直接影响到序列比对的结果，如表 1.1 所示。

表 1.1　起始空位罚分和延伸空位罚分对比对的影响

起始空位罚分	延伸空位罚分	说　　明
大	大	插入和删除极少，适用于相似性较高的序列比对
大	小	少量的大片空位插入
小	大	大量小块插入，用于相似性较低的序列比对

仿射空位罚分根据生物定义将空位分为起始空位(gap open penalty)和延续空位(gap extend penalty)，其产生的罚分分别简写为 *GOP* 和 *GEP*，$\sum penalty = N_{GOP} \bullet GOP + N_{GEP} \bullet GEP$。其中，$N_{GOP}$ 是 *GOP* 的个数，N_{GEP} 是 *GEP* 的个数，且 *GOP*>*GEP*。具有生物意义的仿射罚分是当前最常应用的罚分方式。图 1.5 示例以仿射空位罚分计算，则如图 1.6 所示。

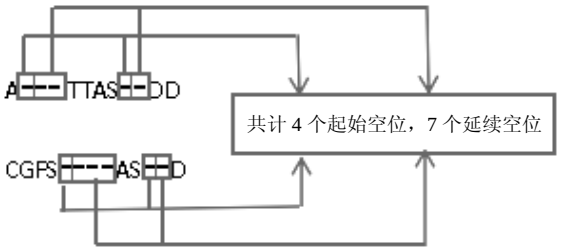

图 1.6　仿射空位罚分计算

假设起始空位罚分为 5，延续空位罚分为 1，则该例的总空位罚分是 4×5+7×1=27 分。

1.4　多序列比对方法

1.4.1　比对方法

1. 手工比对

手工比对方法就是通过辅助编辑软件(如 BioEdit、SeaView、GeneDoc 等软件)的不同颜色显示不同残基，靠分析者的观察，手工改变比对的状态。手工比对方法在文献中经常看到。因为在手工比对过程中难免会有一些主观因素，通常被认为有较大的随意性。

通常使用不同颜色表示具有不同特性的残基，以便判别序列间的相似性。颜色的选择非常重要，如果使用不当，则看起来不够直观，就会丢失一些比对结果中的有用信息。反之，如果选择恰当，就能从比对结果中快速找到某些重要的结构模式和功能位点。颜色的选择可以根据主观愿望和喜好，但是最好和常规方法一致。表 1.2 给出了氨基酸分组方法和代表性颜色。

表 1.2　氨基酸分组方法和代表性颜色

残基种类	残基特性	颜色
Asp(D)，Glu(E)	酸性	红色
His(H)，Arg(R)，Lys(K)	碱性	蓝色
Ser(S)，Thr(T)，Asn(N)，Gln(Q)	极性	绿色
Ala(A)，Val(V)，Leu(L)，Ile(I)，Met(M)	疏水性，带支链	白色
Phe(F)，Tyr(Y)，Trp(W)	疏水性，带苯环	紫色
Pro(P)，Gly(G)	侧链结构特殊	棕色
Cys(C)	能形成二硫键	黄色

2. 计算机程序自动比对

通过特定的算法(如同步法、渐进法等算法，详见 1.4.2 节)，由计算机程序自动搜索最佳的多序列比对状态。

(1) 同步法：将序列两两比对时的二维动态规划矩阵扩展到三维矩阵；即用矩阵的维数来反映比对的序列数目。这种方法的计算量很大，对于计算机系统的资源要求较高，一般只有在进行少数的较短的序列比对时才会用到这种方法。

(2) 渐进法：最常见的是 Clustal 所采用的方法，其基本思想就是基于相似序列通常具有进化相关性的假设。

1.4.2　多序列比对算法

在生物多序列比对方面，现有的算法基本分为三大类：精确比对算法、近似比对算法和基于图论的比对算法。

1. 精确比对算法

精确比对算法是基于数学理论基础上的一种动态规划算法。利用动态规划思想求解序列比对问题的方法最早由 Needleman 和 Wunsch 于 1970 年联合提出，并将该算法用于求解两条蛋白质序列的全局比对问题，因此该方法也称为 Needleman-Wunsch 算法，被视为经典的双序列比对全局动态规划算法。1981 年，Smith 和 Waterman 在改进 Needleman-Wunsch 算法的基础上提出了经典的双序列比对局部动态规划算法—— Smith-Waterman 算法，该算法适用于亲缘关系较远、整体上不具有相似性但在一些较小的区域上存在局部相似性的两条序列。动态规划算法的思想核心是将原问题分解为子问题，基本步骤如下：

(1) 最优分的递归计算。

(2) 存储子问题的最优分的动态规划矩阵。

(3) 子问题最优解矩阵的填充过程。

(4) 寻找最优比对路径的回溯方法。

精确比对算法的优点是比对非常精确，不会重复算或漏算。通过此方法虽然能得到理论上的最佳值，可是现实中并不常采用此方法，因为此算法的空间和时间的复杂度都会随着序列条数和序列的长度成指数级的速度增长。由于动态规划消耗的时间开销太大，因此一般只用于双序列比对问题的求解中。动态规划算法是生物信息学中一个最流行的解决方法，序列的比较、基因的识别、蛋白质序列的重排以及蛋白质结构和功能的分析等很多生物信息学中的问题都可以通过动态规划的方法解决。但是基本动态规划方法的时间和空间复杂度为 $O(m\times n)$，满足不了实际的需要，所以，后人在此基础上提出了各种各样的基于动态规划思想的改进算法。下面介绍经典的两类双序列动态规划算法。

1) Needleman-Wunsch 算法

Needleman-Wunsch 算法是用在蛋白质序列的双比对问题上的经典全局动态规划算法，该算法从整体水平上分析不同的两条序列之间的进化关系或者同源关系，即考虑序列总长度的比较，用类似于使整体相似最大化的方式，对序列进行比对。如果参与比对的是长度互不相同的几条序列，则需要采用一些方法在序列的某些位置插入空格，从而使序列的长度达到一致。Needleman-Wunsch 算法的基本思想是：使用递归方法计算出两条序列所有可能的比对结果的相似性得分，同时将该得分存储在某个矩阵中，该矩阵称为得分矩阵。如果将矩阵中每一个分值所在单元格的位置称为一个单元，则从一个单元移动到另一个单元的路径或者说方向最多有三种：向上、向左和向左上。回溯时从右下角开始，每次都选择相邻的最大元素，并用箭头对所选的路径做出标记，直到到达矩阵的左上角时才算结束。这时根据标出的路径，通过动态规划的方法回溯，则能寻找出最优的相似性比对。

设序列 S 和 T 的长度分别为 m 和 n。考虑 S 和 T 的两个前缀 $s[0,i]$ 和 $t[0,j]$，$i,j>1$。假设已知序列 $s[0,i]$ 和 $t[0,j]$ 所有较短的连续子

序列的最优比对，即已知：

(1) $s[0,i-1]$和$t[0,j-1]$的最优比对；

(2) $s[0,i-1]$和$t[0,j]$的最优比对；

(3) $s[0,i]$和$t[0,j-1]$的最优比对。

则$s[0,i]$和$t[0,j]$的最优比对一定是上述三种情况之一的扩展，即：

(1) 替换(s_i,t_j)或匹配(s_i,t_j)，这取决于s_i是否等于t_j；

(2) 删除$(s_i,-)$；

(3) 插入$(-,t_j)$。

令$M(s[0,i],t[0,j])$为序列$s[0,i]$和$t[0,j]$对比的得分，可根据式(1.1)所示的递归算式计算最大值：

$$M(s[0,i],t[0,j])=\max\begin{cases}M(s[0,i-1],t[0,j]+\sigma(s_i,-))\\M(s[0,i-1],t[0,j-1]+\sigma(s_i,t_j))\\M(s[0,i],t[0,j-1]+\sigma(-,t_j))\end{cases}\tag{1.1}$$

其初值为
$$\begin{cases}M(s[0,0],t[0,0])=0\\M(s[0,i],t[0,0])=M(s[0,i-1],t[0,0]+\sigma(s_i,-))\\M(s[0,0],t[0,j])=M(s[0,0],t[0,j-1]+\sigma(-,t_j))\end{cases}\tag{1.2}$$

按照这种方法，对于给定的打分函数$\sigma(s_i,t_j)$，两条序列所有前缀的比对得分值定义了一个$(m+1)\times(n+1)$的得分矩阵：

$$\boldsymbol{D}=(d_{i,j})\tag{1.3}$$

式中，$d_{i,j}=M(s[0,i],t[0,j])$。

对于一个长度为 n 的序列，有$n+1$个前缀(包括一个空序列)，所以得分矩阵的大小为$(m+1)\times(n+1)$。其中矩阵的纵轴方向自上而下对应于第一条序列 S，横轴方向从左到右对应于第二条序列 T。矩阵横向移动表示在纵轴序列中加入一个空位，纵向的移动表示在横轴序列中加入一个空位，而斜对角向的移动表示两序列各自相应的字符进行比对，各轴第一个元素的索引下标为 0。

$d_{i,j}$ 的计算公式为

$$d_{i,j}=\max\begin{cases}d_{i-1,j}+\sigma(s_i,-)\\ d_{i-1,j-1}+\sigma(s_i,t_j)\\ d_{i,j-1}+\sigma(-,t_j)\end{cases}\tag{1.4}$$

$d_{i,j}$ 最大值的三种选择决定了各矩阵元素之间的关系，如图 1.7 所示。

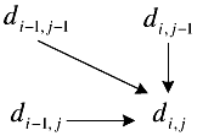

图 1.7　矩阵各元素的关系

矩阵右下角元素即为期望的结果：$d_{m,n}=M(s[0,m],t[0,n])=M(s,t)$。其计算过程描述如下：首先初始化得分矩阵 $\boldsymbol{D}$，然后计算 $\boldsymbol{D}$ 矩阵的其他元素。计算过程从 $d_{0,0}$ 开始，可以按行计算，每行从左到右，也可以按列计算，每列从上到下。任何计算过程，只要满足在计算 $d_{i,j}$ 时 $d_{i-1,j}$、$d_{i-1,j-1}$、$d_{i,j-1}$ 都已经被计算这个条件即可。在计算 $d_{i,j}$ 后，需要保存 $d_{i,j}$ 是从 $d_{i-1,j}$ 或 $d_{i,j-1}$ 中的哪一个推进的，或保存计算的路径，以便于后续处理。上述计算过程到 $d_{m,n}$ 结束。

与计算过程相反，求最优路径或最优比对时，从 $d_{m,n}$ 开始，反向前推。假设在反推时到达 $d_{i,j}$，现在根据保存的计算路径判断 $d_{i,j}$ 究竟是根据 $d_{i-1,j}$、$d_{i-1,j-1}$、$d_{i,j-1}$ 中的哪一个计算而得到的。找到这个点以后，再从此点出发，一直到 $d_{0,0}$ 为止。走过的这条路径就是最优路径(即得分最大路径)，其对应于两条序列的最优比对。

根据算法原理可以计算出 Needleman-Wunsch 算法在双序列比对的时间复杂度为 $O(m\cdot n)$。三条序列比对可以理解为将双序列比对的二维空间扩展到三维空间，类似于数学中的三维坐标系，即在原来的二维平面上增加一条坐标轴，此时算法的时间复杂度就变成

了 $O(m \cdot n \cdot p)$，p 是第三条序列的长度。依此类推，当用 Needleman-Wunsch 算法解决 N 条序列比对的时间复杂度为 $O(l_1 \cdot l_2 \cdot l_3 \cdot \cdots \cdot l_N)$，$l_N$ 是第 N 条序列的长度。

2) Smith-Waterman 算法

在生物学中，残基的功能点是由较短的序列片段组成的，亲缘关系较远的两条蛋白质序列可能只存在一些相同的基因片段，因此寻找出具有最大相似度的片段对于物种亲缘关系的测定具有重要意义。1981 年，Smith 和 Waterman 在改进 Needleman-Wunsch 算法的基础上提出了 Smith-Waterman 算法。该算法是经典的双序列局部比对动态规划算法，适用于亲缘关系较远、整体上不具有相似性但在一些较小的区域上存在局部相似性的两条序列。

Smith-Waterman 算法的思想与 Needleman-Wunsch 算法基本相似，仍然采用动态规划思想，记分矩阵的方式在识别局部相似性时灵敏度非常高。只不过在 Smith-Waterman 的记分矩阵中元素是根据式(1.5)所得，且三个记分函数的值也有相应的变化。

$$M(s[0,i],t[0,j]) = \max \begin{cases} M(s[0,i-1],t[0,j] + \sigma(s_i,-)) \\ M(s[0,i-1],t[0,j-1] + \sigma(s_i,t_j)) \\ M(s[0,i],t[0,j-1] + \sigma(-,t_j)) \end{cases} \tag{1.5}$$

其初值为

$$\begin{cases} M(s[0,0],t[0,0]) = 0 \\ M(s[0,i],t[0,0]) = 0 \\ M(s[0,0],t[0,j]) = 0 \end{cases} \tag{1.6}$$

因此，Needleman-Wunsch 算法和 Smith-Waterman 算法的不同点在于前者的回溯从记分矩阵的右下角开始，回溯到左上角终止；后者的回溯从记分矩阵中的最大元素开始，终止于第一个遇到的 0 元素。

2. 近似比对算法

近似比对算法又称启发式算法。目前国际上最具代表性的启发式算法有两大类：渐进比对算法和迭代比对算法。

1) 渐进比对算法

渐进比对算法又称贪心算法，最开始是由 Hogeweg 提出的，Feng 和 Doolittle 等改进了此算法，并开发出了现在经常使用的序列比对程序软件包 Clustal、ClustalW 和 T-Coffee。这种算法的思路是：首先将多个序列两两比对构建距离矩阵，反映序列之间的两两关系；然后根据距离矩阵计算系统进化指导树，对关系密切的序列进行加权；最后从关系最紧密的两条序列开始，逐步引进临近的序列并不断重新构建比对，直到所有序列都被加入为止。渐进比对算法由于简单快速的特点，使它成为多序列比对中最常用的方法之一，但由于渐进比对算法的本质为贪心算法，一次比对的部分结果不能随着比对过程中更多序列的加入而改变，因此由于渐进过程中部分比对结果是“冻结”的，从而导致“局部最小化”问题的产生。

ClustalW 是一个最常用、最经典的基于渐进比对算法的多序列比对程序。ClustalW 采用近邻法生成向导树，此算法对树的每一个分枝长度的估算更准确，这样依据分枝长度计算的序列权重也就更加准确。另外，ClustalW 所采用的空位罚分策略比较复杂，影响空位罚分的因素有残基特异性、序列长度、比对时采用的记分矩阵、序列相似程度、空位位置等，这就使得空位的产生更加合理，从而提高了比对的准确度。T-Coffee 也是一个采用渐进比对算法的多序列比对程序，它与 ClustalW 最大的不同在于：前者采用 SP 记分函数作为目标函数，而 T-Coffee 采用 Coffee 记分函数作为目标函数，这样可以有效减少比对初期的错误，使得 T-Coffee 可以快速、准确地进行序列比对。测试结果表明 T-Coffee 比对的准确度高于 ClustalW，但其比对速度低于 ClustalW。Iain M.Wallace 提出了一种 M-Coffee 的

方法，此方法是在 T-Coffee 的基础上进行的。Katoh K 提出的 MAFFT 是一个引入快速傅里叶变换的快速多序列比对程序，它将每一条序列转化为包含每一个残基的体积和极性值的序列，然后采用快速傅里叶变换算法来寻找序列间的同源模块，从而减小动态规划的矩阵空间。MAFFT 中的 FFT-NS_2 基于渐进比对算法实现，测试结果表明其比对速度快于 ClustalW。Edgar R C 提出的 MUSCLE 方法在时间复杂度和空间复杂度上进行了改进。

2) 迭代比对算法

迭代比对是另一类有效的应用广泛的多序列比对策略，与渐进比对算法不同，它基于一个能产生比对的算法，并通过一定的进化策略，不断迭代替换来精细多序列比对，直到比对结果不再改进为止。这类算法的缺点是不能提供获得优化比对结果的保证，速度也不能和渐进比对算法相比。优点是将目标函数和优化过程在概念上进行了分离，具有鲁棒性、对于序列比对个数不敏感等特性。

近年来，迭代算法也被越来越多地应用到序列比对中去，如遗传算法、蚁群算法、隐马尔科夫等。Cedric Notredame 首先提出用遗传算法解决多序列比对问题，在遗传算法中对 22 种交叉变异算子应用了一个自动调度机制，并证实了此算法的有效性，其准确度与 ClustalW 结果相似。Notredame 提出了将遗传算法应用在 RNA 序列比对上，并取得了很好的比对结果。Thomsen 研究发现，使用自动调度策略得到的结果并不比以同等概率选择遗传算子好，因此自动调度策略的使用并非改善其算法的准确性，反而使得算法的复杂度增高。Goondro C 认为初始化种群的质量直接影响到算法的收敛速度，适应度值高的种群能够很快地收敛到接近最优解的解，因此他提出了一种新的初始化种群的方法，以增高初始种群的适应度值。Fernando Jose Mateus da Silva 提出了将局部最优搜索融入遗传算法中的新算法，提高了算法的准确度。胡桂武提出了一种基于遗传算法与星比对算法的多序列比对混合算法，这种算法是先通过星比对

算法得到一个多序列比对，然后将这个比对作为种群中的一个个体，并结合遗传算法的一种混合算法。张维梅提出了一种基于遗传算法和蚁群算法的多重序列比对算法，这种算法是将蚁群算法作为局部搜索的一种算法。司秀华提出了一种多搜索策略的多生物序列比对自适应遗传算法，这种算法是通过调整遗传算法中的交叉率和变异率从而避免算法找到局部最优而提出的。司徒浩臻提出了基于遗传算法的多序列比对算法。还有许多的求解多序列比对的混合算法，如遗传算法和蚁群算法相结合的求解多序列比对的方法、遗传算法和 GLOCSA 相结合的序列比对方法、并行混合遗传算法，还有一些以遗传算法为主的序列比对算法。除了这些算法外，还有许多其他的方法，如 Taheri J 提出的两种求解多序列比对的方法 RBT-L 和 RBT-GA。Fan H 提出了智能算子在遗传算法中的应用。邹权和郭茂祖等综述了常见的启发式方法，除了遗传算法和粒子群算法，还有许多其他的启发式方法，尽管在应用到多序列比对问题上做了许多尝试，但中间还存在一些十分难处理的问题，因而还没有形成基于这些方法的主流软件。

根据迭代算法的特点，本书以迭代算法作为解决多序列比对的主要方法，包括遗传算法、粒子群算法、量子粒子群算法以及结合 HMM 的粒子群算法等方法，在后面将详细介绍这些迭代算法的基本原理与应用。

3. 基于图论的比对算法

应用图论解决多序列比对问题的思想是近年来提出并发展起来的一种新方法，它与动态规划算法、渐进比对算法和迭代比对算法有着根本上的不同。

这种算法的基本思路是通过将序列中所有的片段转化成一个 DeBruijn 图，将序列装配问题转变成欧拉路径问题，这种方法称为“欧拉比对”。利用图论方法进行多序列比对最重要的优势是在比对

过程中所需要的时间和空间与序列的长度成线性关系，并且可以提高比对精度。全局多序列比对问题实际上就是多个极端片段的装配问题，如图 1.8 所示，如果几乎所有的片段都来自于基因组的相同区域，欧拉比对方法就会输出该区域的一致性序列。对于多序列比对，如果一致性序列是与所有给定序列最接近的一个，那么就希望通过某种记分机制提高序列的一致性。欧拉比对将原始的复杂图转化成有向无环图(directed acyclic graph，DAG)，并且在转化过程中尽可能记住这些序列的 k 元组，而且 DAG 图的边的权重记录的是序列的一致性的最大值。

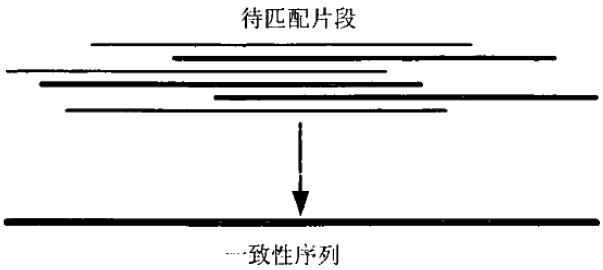

图 1.8　序列比对问题转化为序列装配问题

通过这两步，寻找一致性序列的问题就成了寻找最大路径的问题，寻找的主要流程是：①用序列中需要比对的 k 元组构造有向 DeBruijn 图；②将有向 DeBruijn 图转换成 DAG；③根据 DAG 各条边的权重求出一致性路径；④在一致性路径和每个输入序列之间作一次快速双序列比对；⑤根据双序列比对构造最后的多序列比对结果。

基于图论的多序列比对方法的主要代表就是 Lee 和 Grasso 等提出的偏序比对方法(partial order alignment，POA)。POA 能直接根据动态的双序列比对程序自动地进行比对，避免了将 MSA 降低维数的麻烦，从而保证每一个新序列与 MSA 中的序列的最优比对都能被考虑到。在 POA 算法中，新的编辑操作——同源重组在多

区域序列中起着重要作用，以有向无环图(DAG)的表示方式取代了过去 30 多年来用线性行列表示多序列比对的方式。在行列表示的比对方法中，图总是单一的有向路径，而 POA 方法扩展了图的结构，使得它成为一个有向无环图，打破了这个限制，在多序列比对领域中打开一个全新的视角。后来 Yuzhen Ye 和 Adam Godzik 把偏序结构又进一步应用于蛋白质序列比对中。Benjamin Raphael 等在 2004 年提出了 ABA(A-Bruijn alignment)算法。ABA 算法与以前的方法不同点在于：ABA 以有向图的方式表示一个比对，在这个图中允许环的存在。这种表示方式使得 ABA 算法比传统的比对矩阵甚至偏序比对(POA)更具有灵活性，尤其适用于含有交错或重复区域结构的蛋白质序列。允许构建包含下面三类区域的蛋白质序列：①不是在所有蛋白质序列中都出现的区域；②在不同的蛋白质中以不同顺序出现的区域；③在某些蛋白质序列中重复出现的区域。ABA 在求解包含重复和倒位的 DNA 序列比对问题时非常适用。霍红卫提出了用于进行全局 DNA 多序列比对的基于最大权值路径算法的 DNA 多序列比对方法。

现有的多序列比对程序大多基于上述算法思想或多种算法思想的结合并且选择不同的目标函数。表 1.3 列出了部分研究成果。

现有的多序列比对程序各有其优缺点，有文献对部分比对程序做了比较，所得结论为不存在唯一最好的比对程序，在进行序列比对时，根据问题特性，选用不同的序列比对程序，以期得到最满意的比对结果。而对于所得到的比对结果也不能简单地做出“正确或错误”的结论，因为多序列比对的方法建立在某个数学或生物学模型上，所以只能认为所使用的模型在多大程度上反映了序列之间的相似性关系以及它们的生物学特性。

表 1.3　多序列比对算法统计表

程序名称	算法
MSA	准确比对，基于 Carrillo-Lipman 算法
DCA	准确比对，基于分治法
CLUSTALW	渐进比对，基于动态规划方法，SP 记分函数
T-Coffee	渐进比对，基于动态规划方法，COFFEE 记分函数
Praline	渐进比对和迭代比对结合
IterAlign	迭代比对
Prrp	随机迭代比对
HMMT	随机迭代比对，基于 Markov 模型
SAGA	随机迭代比对，基于遗传算法，SP 和 COFFEE 记分函数可选
PHGA	随机迭代比对，基于并行遗传算法，SP 记分函数
MAFFT	其中 FFT-NS-2 采用渐进比对，FFT-NS-1 采用迭代比对
MUSCLE	渐进比对
Align-m	渐进比对
PROBCONS	渐进比对
POA	图论比对
ABA	图论比对

1.5　多序列比对常用数据库

随着生物数据的爆炸增长，目前记录的核苷酸序列已有成千上万条，蛋白质序列超过 10 万条。如此巨大数量的资源，必须应用电子数据库的存储和计算机分析。在对生物序列进行分析的过程中，除了由生物学实验直接得到实验数据信息外，还可以通过

查询相关的生物学数据库来得到相关的数据信息，特别对于数据分析方法的确立和研究，大量已知的数据信息是必须和非常重要的。虽然目前有很多不同类型的数据库，但这里将结合本书的主题，根据数据库的用途进行分类，分别介绍两种常用的生物序列数据库，通过应用数据库可以得到核酸序列和氨基酸序列的基本信息查询和序列比对算法的验证，为基因与蛋白质的深入分析提供可参考的信息。

1.5.1　综合性数据库

这类数据库中包含的生物数据信息非常齐全，通常应用于搜索查询相关的生物数据。以下是两个常用的数据库中心，它们提供了100多种不同数据库的链接。

1. NCBI(美国国家生物技术信息中心)

NCBI(http://www.ncbi.nlm.nih.gov)全称为National Center of Biotechnology Information。参议员Claude Pepper意识到信息计算机化过程方法对指导生物医学研究的重要性，发起了在1988年11月4日建立国立生物技术信息中心(NCBI)的立法。NCBI是NIH的(美国)国立医学图书馆(NLM)的一个分支。NLM是因为它在创立和维护生物信息学数据库方面的经验被选择的，而且这可以建立一个内部的关于计算分子生物学的研究计划。NCBI的主要任务是发展新的信息学技术来帮助对那些控制健康和疾病的基本分子和遗传过程的理解。NCBI有核酸、蛋白、基因名、基因组名等搜索工具，GenBank数据库、BLAST序列比对搜索工具，PUBMED文献数据库，Taxonomy数据，COG蛋白家族库等。FTP可以下载它全部的数据库、BLAST的单机程序，以及各种工具程序。到目前为止，NCBI已成为世界级的生物信息资源中心，为生物医学和生命科学研究提供了大量数据和分

析工具的平台。

2. EBI(欧洲生物信息研究所)

EBI(http://www.ebi.ac.uk/)全称是 European Bioinformatics Institute。EBI 拥有超过 20 年生物信息学研究和服务经验，是全球收集和传播生物数据、提供免费生物信息服务的欧洲节点。该所管理维护着世界最全面的分子生物数据库，其中很多是生物学家熟悉的数据库，如 ENA(核酸序列数据库)、Ensembl(基因组)、ArrayExpress(基因表达数据)、UniProtKB 蛋白质序列、InterPro(蛋白质家族/域/蛋白指纹等)和 PDBe(大分子结构)。ENA 在原 EMBL-Bank 核酸序列数据库基础上发展起来，是欧洲最重要的核酸序列资源，与美国 NCBI 的 GenBank 和日本的 DDBJ 组成国际核酸序列数据库合作联盟(INSDC)。这三大数据库各自收录了世界上所报道的所有序列数据的一部分，并且每天实时更新交换各自的序列信息。EBI 的数据资源包括 IntAct(蛋白质相互作用)、Reactome(反应途径)、ChEBI(小分子)等。EBI 向全球提供免费的生物信息服务，发展和维护着多种用于浏览、检索、分析处理生物数据的工具服务。数据获取工具 SRS(序列检索系统)为用户提供了快速、便捷和友好的界面以搜索超过 400 个局域和公众数据库中大量不同种类的生命科学类数据。序列数据搜索包括 FASTA、NCBI BLAST 和 WU-BLAST 序列同源性和相似性对比工具。其他还包括蛋白质功能分析、进化树分析、大分子结构分析与多维显示等。

在应用数据库搜索序列时，有两种常见的序列文件格式，如图 1.9 和图 1.10 所示。

GenBank 序列文件中包含了一个基因的每个记录，出现在该图中的不同列中。属于平面文件格式，应用于文字处理计算机语言方面。域的名称显示在前面，完整的记录非常长，因此在这里只显示了顶部。

```
LOCUS      RATOBESE     539 bp ss-mRNA          ROD      23-SEP-1995
DEFINITION  Rat mRNA for obese.
ACCESSION   D49653
KEYWORDS    .
SOURCE      Rattus norvegicus (strain OLETF, LETO and Zucker, ) differentiated
            adipose cDNA to mRNA.
  ORGANISM  Rattus norvegicus
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Sarcopterygii; Mammalia; Eutheria; Rodentia;
            Sciurognathi; Myomorpha; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 539)
  AUTHORS   Murakami,T. and Shima,K.
  TITLE     Cloning of rat obese cDNA and its expression in obese rats
  JOURNAL   Biochem. Biophys. Res. Commun. 209, 944-952 (1995)
  STANDARD  full automatic
COMMENT     Submitted (10-Mar-1995) to DDBJ by:
            Takashi Murakami
            Department of Laboratory Medicine
            School of Medicine
            University of Tokushima
            Kuramotocho 3-chome
            Tokushima 770
            Japan
            Phone: +81-886-33-7184
            Fax:   +81-886-31-9495.
```

图 1.9　序列文件格式 GenBank

```
>gi|995614|gb|D49653|RATOBESE Rat mRNA for obese.
CCAAGAAGAAGAAGACCCCAGCGAGGAAAATGTGCTGGAGACCCCTGTGCCGGTTCCTGTGGCTTTGGTCC
TATCTGTCCTATGTTCAAGCTGTGCCTATCCACAAAGTCCAGGATGACACCAAAACCCTCATCAAGACCATTG
TCACCAGGATCAATGACATTTCACACACGCAGTCGGTATCCGCCAGGCAGAGGGTCACCGGTTTGGACTTCA
TTCCCGGGCTTCACCCCATTCTGAGTTTGTCCAAGATGGACCAGACCCTGGCAGTCTATCAACAGATCCTCAC
CAGCTTGCCTTCCCAAAACGTGCTGCAGATAGCTCATGACCTGGAGAACCTGCGAGACCTCCTCCATCTGCT
GGCCTTCTCCAAGAGCTGCTCCCTGCCGCAGACCCGTGGCCTGCAGAAGCCAGAGAGCCTGGATGGCGTCC
TGGAAGCCTCGCTCTACTCCACAGAGGTGGTGGCTCTGAGCAGGCTGCAGGGCTCTCTGCAGGACATTCTTC
AACAGTTGGACCTTAGCCCTGAATGCTGAGGTTTC
```

图 1.10　序列文件格式 FASTA

FASTA 序列文件中包含了 gi 号码、GenBank 检索号码、LOCUS 名称以及 GenBank 记录中的 DEFINATION 字段。第一行(>)表示一个新的序列文件的开始，为标记符。后面可以加上文字说明、gi 号码、GenBank 检索号码、LOCUS 名称等信息。第二行序列为 DNA 或蛋白质的标准符号。通常核苷酸符号大小写均可，而氨基酸则一般用大写字母。不过，有些程序对大小写有明确的要求，使用时需要注意。一般每行 60～80 个字母。

1.5.2　基准数据库

这类数据库中包含的数据一般是真实的、已知结构的蛋白质序列，根据手工比对和反复验证得到的标准结果，通常应用于测试或

衡量比较不同生物信息学软件的性能。例如以下的各数据库或软件：BAliBASE——测试蛋白质多序列比对的准确性；GAGE (genome assembly gold-standard evaluations)——高通量测序结果用于组装基因组，测试组装出来的正确率；CASP(critical assessment of protein structure prediction)/CAFASP——预测蛋白质结构；CAPRI(critical assessment of prediction of interactions)——预测蛋白质相互作用/结合；CAGI (critical assessment of genome interpretation)——预测基因组上的变异会对生物的表型产生什么影响；DREAMchallenges(http://dreamchallenges.org/)——多种不同生物信息学任务的比拼，如预测选择性剪接。

1. 基准比对数据库 BAliBASE

对于绝大多数多序列比对算法来说，很难从理论上给出所得结果与优化比对之间的偏差。当前多序列比对算法众多，且生物序列之间的进化关系也相当复杂，每种序列比对算法都有它相对适用的范围，并非对所有序列都有效。每一种新算法被提出时，作者都要选定一些数据集，与已存在的比对算法进行准确性的比较，这是评价一个算法优劣的方法。但由于每一种算法都有它自己的特性，因此由作者自己选定数据集与其他算法相比显然对其他算法是不公平的。

为了统一评判各种多序列比对算法的有效性，法国生物细胞分子基因组研究所(Institut de Génétique et de Biologie Moléculaire et Cellulaire，IGBMC)的 Thompson 等于 1999 年构建了由真实蛋白质序列组成的基准(benchmark)比对数据库 BAliBASE。该数据库中共存储了 1000 多条真实的蛋白质序列，构成了 142 组参考比对。这些参考比对是基于相应蛋白质的三维结构而确定的，其可靠的比对区域被标注为核心块。这些都已被多个程序证明且通过手工修正了的高质量的比对结果，并且具有良好的文档说明。这些参考比对又根据比对的特点，如序列的长度、相似性以及插/失空位数量

及位置等因素而被划分成五类，涵盖了多序列比对的绝大多数问题。由于 BAliBASE 不是为某一具体比对算法而精心设计的，因此它相对客观，是目前用于评价多序列比对算法有效性的一个通用平台。

2001 年 Bahr 等又在第一版的 BAliBASE 的基础上对原有的参考比对做了进一步修正，并新添加另外三类基准比对共计 1000 多条序列，推出了 BAliBASE2.01。2005 年 Thompson 等又改进了第二版，推出 BAliBASE3.0，在原有的基础上又增加了更多的序列。

一般情况下，常用于测试比对算法的还是 BAliBASE 中的前五类参考比对。

第一类参考比对是由少量的等长序列构成。任意两个序列间相同残基的百分比，即一致率(ID%)要在某一具体范围内，并且不存在大范围的插失，每一比对中都含有 3～7 条序列。

第二类参考比对是在一个蛋白质家族(序列间的亲缘关系较近>25%ID)序列中加入三条“孤儿”序列(“orphan”sequence，家族中亲缘关系较远<20%ID 的成员，但享有共同的折叠)。每一比对中至少含有 15 条近亲序列和 3 条“孤儿”序列。

第三类参考比对中，每一比对至少由 4 个不同家族的蛋白质构成，来自不同家族的任意两个序列相同残基的百分比<25%ID。

第四类参考比对中包含具有大量的 N/C 终端扩展的序列。

第五类参考比对中包含具有大量的内部插/失的序列。

更具体的 BAliBASE 每一类比对数目如表 1.4 所示。

BAliBASE 中提供了两个不同的评分分值 SPS 和 TCS 及程序，分别用于评价与 BAliBASE 中参考比对进行比较的一个测试比对算法的质量，应用该程序包可以直接计算 SPS 和 TCS 分值，如图 1.11 所示。SPS 和 TCS 分值越高，多序列比对算法越好。这两个评分标准将在本书第 4 章中详细介绍。

表 1.4　BAliBASE 数据库

类别	短序列 (<100 个残基)	中等序列 (200～300 个残基)	长序列 (> 400 个残基)
类 1：长度相似的等距离序列			
子类 1：(<25%的一致率)	7	8	8
子类 2：(20%～40%的一致率)	10	9	10
子类 3：(>35%的一致率)	10	10	8
类 2：带孤儿的家族	9	8	7
类 3：等距离的分歧家族	5	3	5
类 4：N/C 终端延展	12		
类 5：内部空位插入	12		

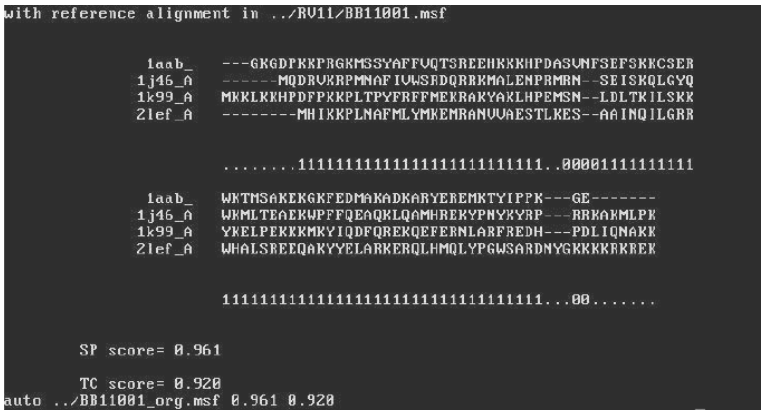

图 1.11　BAliBASE 自带 SPS 和 TCS 程序包 baliscore 运行界面

2. 其他蛋白质序列比对基准数据库

在应用基准数据库测试评估多序列比对算法性能时，除了

BAliBASE 数据库，还有其他的蛋白质结构数据库也被作为基准数据库，如 HOMSTRAD 同源结构比对数据库(http://www-cryst.bioc.cam.ac.uk/homstrad/)、SMART 注释了的蛋白质结构域序列(http://smart.embl-heidelberg.de/)、Pfam 蛋白质家族数据库、SCOP 由专家参与的蛋白质结构分类数据库、SABmark 序列比对基准(http://bioinformatics. vub.ac.be)、Prefab3.0、Rose version 1.3 等。

1.6　多序列比对常用工具

1.6.1　搜索工具

一般在多序列比对之前，首先要在数据库中搜索相对应的序列，这里主要介绍 EBI 的 FASTA 工具和 NCBI 的 BLAST 工具，它们是当前最常用的两大数据库搜索工具。

1. FASTA 工具

FASTA 是 FAST-ALL 的缩写，是由 Lipman 和 Pearson 于 1985 年提出的。其基本思路是：

(1) 识别与待查序列相匹配的很短的序列片段，称为 k-tuple。使用者可以改变 ktup 值，一般规定蛋白质序列的 ktup 默认值是 2，DNA 序列的 ktup 默认值是 6。

(2) 运算是寻找与最初识别的单词匹配的扩展，试图找到序列的无空位联配，该联配含有高密度的最初识别的单词匹配，然后再把这些联配加入到高分值的有空位的联配中。最后在识别了序列间的高分值联配后，通过动态规划联配全部序列高分区域，得出最终联配及其分值。

FASTA 可用于核酸和蛋白质序列的快速序列比对数据库搜索。其版本在不断更新升级，可以下载使用，也可以在线进行比对，最

新版本是 FASTA3 软件包，主要包括 FASTA、FASTF、FASTS、FASTX/Y、TFASTX/Y，如表 1.5 所示。

表 1.5　FASTA3 软件包相关内容

程序	查询序列类型	数据库类型
FASTA	DNA、蛋白质	DNA 蛋白质
FASTX、FASTY	DNA	蛋白质
TFASTA	蛋白质	DNA
TFASTX、TFASTY	蛋白质	DNA
FASTS、TFASTS	一系列多肽片段	蛋白质 DNA
FASTF、TFASTF	有序多肽混合物	蛋白质 DNA

随着各生物数据库中序列数量的快速增长，FASTA 工具的搜索速度越来越不能满足用户的要求，因此它逐步被一种速度更快的搜索工具 BLAST 所替代。

2. BLAST 工具

BLAST 是 Basic Local Alignment Search Tool 的缩写，是 Altschul 于 1990 年提出的。其基本思路是：

(1) 寻找打分比某一特定阈值(T)高且长度是 W 的单词。蛋白质序列的 W 默认值是 3，DNA 序列的 W 默认值是 11。使用者可以设置 W 和 T 值，但一般都使用默认值。

(2) 寻找与最初识别的单词匹配的扩展。BLAST 将个别单词匹配扩展，直到联配总分值从最高值跌落一段数量，产生无空位的联配。改进后的 BLAST 程序允许空位的插入。

BLAST是当前应用最广泛的序列相似性搜索工具，研究BLAST的最初目的是改善FASTA算法性能，通过寻找更小更好的热点，提高计算速度。为了进一步提高数据库搜索速度，BLAST增加了限制，即在序列的局部比对中不包括空缺字符。BLAST包含五个程序和若干个相应的数据库，分别针对不同的查询序列和要搜索的数据库类型，如表 1.6 所示。其中翻译的核算库指搜索比对时会把核酸数据按密码子所有可能的阅读框架转换成蛋白质序列。

表 1.6　BLAST 软件包相关内容

程序	查询序列类型	数据库类型	简述
BLASTP	蛋白质	蛋白质	适合寻找具有远源进化关系的匹配序列
BLASTN	DNA	DNA	适合寻找分值较高的匹配，不适合远源关系
BLASTX	DNA(翻译)	蛋白质	适合新DNA序列和EST序列的分析
TBLASTN	蛋白质	DNA(翻译)	适合寻找数据库中尚未标注的编码区
TBLASTX	DNA(翻译)	DNA(翻译)	适合分析EST序列

1.6.2　常用的在线多序列比对工具

通过数据库搜索工具收集待测序列后，接下来要进行多序列比对，有很多学者根据多序列比对的原理开发了非常方便好用的比对工具，如 MAFFT、CLUSTALW、T-COFFEE 等，应用这些比对工具能快速地得到较好的比对结果，成为当前多序列比对最常用的比对手段。在EBI官网中(http://www.ebi. ac.uk/Tools/msa/)提供了

这些常用比对工具对应的开源在线多序列比对工具，其中有 Clustal Omega、MAFFT、T-Coffee、MUSCLE 等工具，界面如图 1.12 所示。相对于线下工具，这些在线比对工具的版本更新及时，更为实用。

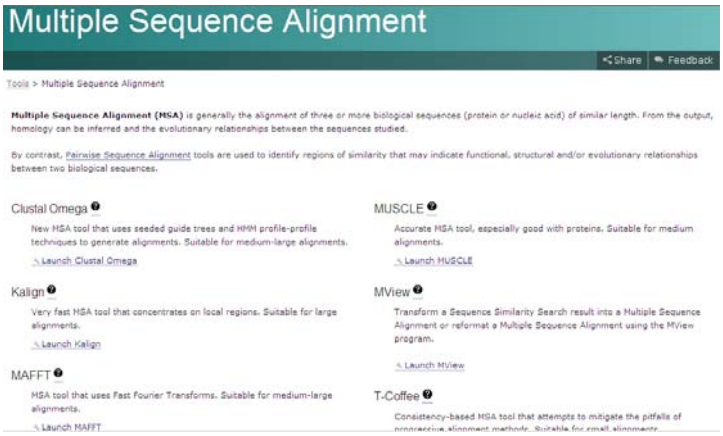

图 1.12　常用比对工具

图 1.13 所示为 MAFFT 工具进行在线序列比对的示例。

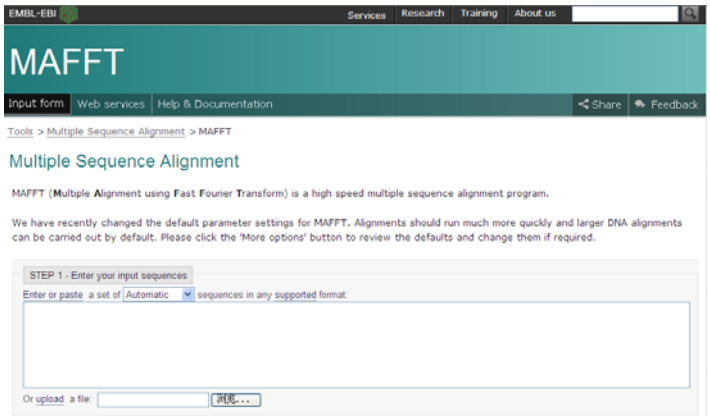

图 1.13　MAFFT 多序列比对工具

在 STEP1 中粘贴要比对的序列，或者单击“浏览”按钮直接选择要比对的序列，如图 1.14 所示。

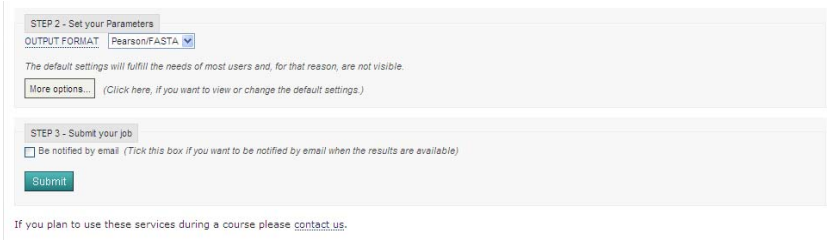

图 1.14　STEP1

在 STEP2 中选择要输出的格式。

在 STEP3 中单击 Submit 按钮，即出现序列比对的结果，如图 1.15 所示。

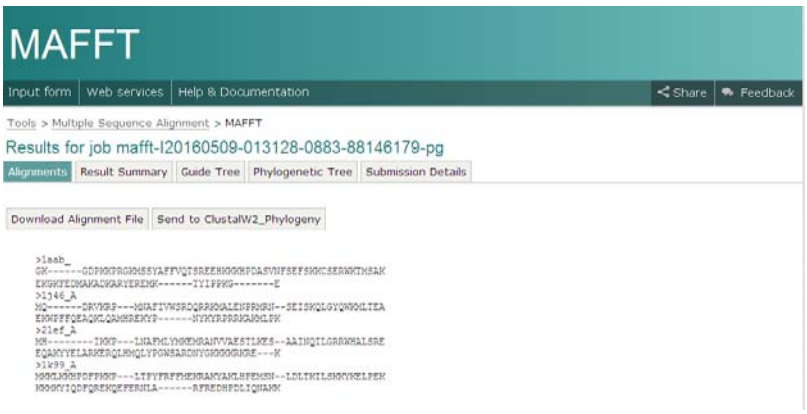

图 1.15　STEP3

另外，NCBI 也有在线蛋白多序列比对工具 COBALT(http://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi)，其界面如图 1.16 所示。

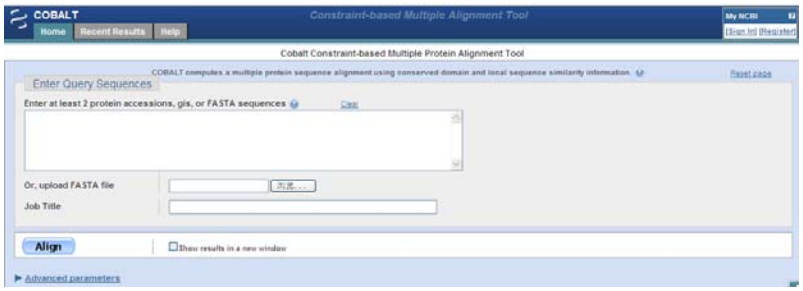

图 1.16　COBALT 多序列比对工具

1.7　多序列比对的应用

在 GenBank 数据库中记录了大量基因或基因组测量的数据，利用这些数据就可以进行多序列比对的计算与分析。下面介绍几个典型的分析与应用例子。

1. 线粒体基因组的比对分析与应用

线粒体在真核生物细胞内普遍存在，且不同生物体的线粒体基因组的长度非常接近。在 GenBank 数据库内保存了上千种线粒体的基因组数据，对线粒体基因组做 MSA 计算与分析是了解这些生物体进化过程的重要手段。

2. 关于流行病基因组的比对分析与应用

流行疾病病毒基因组的类型很多，对这些病毒基因组作比对分析可以了解这些病毒基因组在传播过程中的变异状况。对流行病病毒基因组的比对分析有：

(1) 关于 SARS 病毒的比对分析。在 GenBank 数据库中记录的 SARS 病毒组的 MSA 规模是 108×30kbp。通过 SARS 病毒基因组的 MSA 可以了解到从果子狸到人，以及人的早中晚期的基因突变状

况，尤其是在从果子狸到人，以及人在早中晚期传播时发生较大规模爆发时的基因突变特征。

(2) 关于 HIV-Ⅱ病毒的比对分析。HIV-Ⅱ病毒是一种艾滋病的病毒，它的病毒数据不仅年代长，而且分布地区广，测量所得到的数据还有潜伏期的长短等问题，因此对它们的分析要复杂得多。对 HIV-Ⅱ病毒有专门的数据库，对它们作比对分析的问题也很多，如传播的途径与方式问题；从境外到国内不同地区的传播过程的途径分析以及不同的传播方式(如性交、吸毒、血液等)的分析等；HIV-Ⅱ病毒的类型与测定时间的分析。对 HIV-Ⅱ病毒的分析还可以分解为对多种不同类型的 HIV-Ⅱ病毒，以及测定时病毒潜伏期时间长短的分析等。

(3) 其他类型的流行病，如流感、禽流感等。

1.8　其他说明

1.8.1　多序列比对算法存在的问题

多序列比对算法主要存在以下三大问题：

(1) 多序列的比对问题。序列比对问题目前所存在的问题就是优化双序列比对算法应用于多序列比对，该问题在计算生物学与生物信息学中仍被列为未解决的重大问题或非易计算问题，有的文献把多序列比对列为 NP 完全问题，计算复杂度为双指数问题，也就是计算复杂度为 $O(2^{m+n})$，其中 m 为比对序列的重数，n 为序列长度。因此目前由以上序列比对算法改良的多序列比对只能在小规模上进行，这与目前所出现的庞大数据是极不相称的，如何构造多序列比对的快速算法是计算机生物学与生物信息学中的重大问题。

(2) 比对结果的分析问题。同源序列比对的根本目标是确定它们的进化演变关系，在生物学界常常通过序列比对来构造进化树，

并由此来确定它们的进化关系。但是，比对与进化的关系到底如何，如何由大量序列的比对结果来构造进化树的逻辑过程是生物学家所关心的问题。一个典型的问题是，在现有的比对理论中，把寻找罚分最小(或得分最多)的比对结果看作序列突变与进化的结果，但是求罚分最小的比对结果与序列突变的结果并不完全一致。因此，如何由序列比对来确定序列的突变与进化关系，如何建立它们变化关系的数学模型与分析规则是序列比对理论中不可缺少的一个重要部分。

(3) 不同比对算法的效果分析问题。目前无论双序列还是多序列的比对都有大量算法的出现，另外，对这些比对算法结果也有许多度量性的指标进行评价与考核，如计算复杂度(时间复杂度与空间复杂度)、比对相似度、搜索相似度等，因此对这些不同的比对算法如何建立它们的考核体系尤为重要。目前，对生物信息学的考核还是以进行实际测试计算为最一般的考核方法，要对所设计的比对算法的各项度量指标做出理论上的说明是一个十分困难的问题，因为这涉及序列突变的总体或局部模型问题，这是一个十分复杂的问题，它不仅序列数据庞大，而且突变现象千变万化，所以不可能用一种或几种模型就能把它们概况说明。

1.8.2　多序列比对算法的运算指标

近几年内有许多多重序列的比对问题十分活跃，许多算法、软件包与比对结果大量出现，这些算法或软件包都是在次优解的意义下实现计算。因此，对于不同的多重序列比对算法的好坏需要比较其多种运算指标。多重序列比对算法的主要运算指标有：

(1) 比对规模。就是对比对序列 A 的长度与重数的要求，MSA 算法已基本上实现了无规模的限制。

(2) 运算速度。理想状态的运算速度就是实现 $O(mn)$ 的比对计算复杂度，其中 m 和 n 分别是多重序列的平均长度与重数，MSA 算

法可基本上实现该计算复杂度。

(3) 优化指标的讨论。优化指标是多重序列比对算法中的一个关键问题，建立多重序列比对的优化指标体系实际上是关系到如何理解多重序列比对的优化问题。

1.8.3　多序列比对算法的展望

近年来，随着人们对生物序列认识的逐渐深入，越来越多的蛋白质三维结构及其功能被人们所认识，新的生物序列的结构信息、功能信息、进化关系等也相应加入到序列比对模型中。多序列比对方法的研究现状表明，该领域的相关研究十分活跃，并且取得了巨大的进展，其方法也趋于成熟，但随着大量生物数据的不断加入，多序列比对仍然是进行生物序列分析的基础，在此领域仍有许多问题值得进一步的探索。

(1) 建立更能准确反映生物数据特性的多序列比对数学模型，使得比对结果更加精确。

(2) 改进现有的多序列比对算法，加快问题的求解速度。

(3) 结合渐进比对法和迭代比对法的优点，提出能够克服“局部最小化”问题的快速多序列比对算法。

(4) 比对算法的并行化。由于计算资源的限制和问题求解规模的逐步增大，因此需要实现算法的并行化。

(5) 算法性能分析。由于多序列比对问题的复杂性较高，算法性能分析与评价方法的研究对于算法的改进和优化具有非常重要的意义，因此需要研究不同条件和背景下的算法性能分析方法。

1.9　本章小结

为了详细描述生物多序列比对问题，本章从一些相关的概念

和知识引入，介绍了多序列比对的基本原理、方法、常用数据库等内容，最后介绍了多序列比对的常用工具和应用。

参 考 文 献

[1] 张春霆. 生物信息学的现状与展望[J]. 世界科技研究与发展，2000，22(6)：17-20.

[2] Chuong B, Do, Katoh K. Protein multiple sequence alignment[J]. Humana Press, 2008,484(12): 379-413.

[3] 邹权，郭茂祖，韩英鹏. 多序列比对算法的研究进展[J]. 生物信息学，2010(4)：311-315.

[4] Notredame C. Recent progress in multiple sequence alignment: a survey[J]. Pharmacogenomics, 2002, 3(1): 131-144.

[5] 张鹏帅，霍红卫. 多序列比对问题的粒子群优化算法求解[J]. 计算机工程与应用，2005，41(18)：84-87.

[6] Gusfield D. Algorithms on strings, trees and sequences: computer science and computational biology[M]. Cambridge university press, 1997.

[7] Thompson J D, Higgins D G, Gibson T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice[J]. Nucleic acids research, 1994, 22(22): 4673-4680.

[8] 程灏. 混合遗传算法在图着色及 MSA 问题中的应用[D]. 西安电子科技大学，2006.

[9] 林秋利. 基于进化算法和量子计算的多序列比对方法研究[D]. 西安电子科技大学，2006.

[10] Dan D B, Kececioglu J. Parameter advising for multiple sequence alignment[J]. Bmc Bioinformatics, 2015, 16(1): 516-518.

[11] Gondro C, Kinghorn B P. A simple genetic algorithm for multiple sequence alignment[J]. Genetics and Molecular Research, 2007, 6(4): 964-982.

[12] Wang L, Jiang T. On the complexity of multiple sequence alignment[J]. Journal of Computational Biology, 1994(1): 337-348.

[13] 张敏. 生物信息学中多序列比对等算法的研究[D]. 大连理工大学，2005.

[14] Notredame C, Higgins D G, Heringa J. T-COFFEE: a novel method for fast andaccurate multiple sequence alignments[J]. J.Mol.Evol., 2000, 302(1): 205-217.

[15] 韦树烽，刘羽，蒋财运. 基于 GPU 的遗传退火多序列比对并行研究[J]. 计算机工程与设计，2014，35(4)：1247-1252.

[16] Kaya M, Sarhan A, Alhajj R. Multiple sequence alignment with affine gap by using multi-objective genetic algorithm[J]. Computer methods and programs in biomedicine, 2014, 114(1): 38-49.

[17] Kwan M, Xiao N, Ding G. Assessing Activity Pattern Similarity with Multidimensional Sequence Alignment Based on a Multiobjective Optimization Evolutionary Algorithm[J]. Geographical Analysis, 2014, 46(3): 297–320.

[18] Notredame C, Higgins D G. SAGA: sequence alignment by genetic algorithm[J]. Nucleic acids research, 1996, 24(8): 1515-1524.

[19] 葛宏伟，梁艳春. 基于隐马尔可夫模型和免疫粒子群优化的多序列比对算法[J]. 计算机研究与发展，2006，43(8)：1330-1336.

[20] Chen W, Liao B, Zhu W, et al. An ant colony pairwise alignment based on the dot plots[J]. Journal of Computational Chemistry, 2009, 30(1): 93–97.

[21] Silva F J M, Sánchez Pérez J M, Gómez Pulido J A, et al. AlineaGA—a genetic algorithm with local search optimization for multiple sequence alignment[J]. Applied Intelligence, 2010, 32(2): 164-172.

[22] Silva F J M, Sánchez Pérez J M, Antonio J, et al. An evolutionary approach for performing multiple sequence alignment[C]. WCCI 2010 IEEE World Congress on Computational Intelligence, Barcelona, Spain, July. 2010: 18-23.

[23] Rodríguez M A V. Optimizing Multiple Sequence Alignment by Improving Mutation Operators of a Genetic Algorithm[J]. Intelligent Systems Design and Applications, 2009.ISDA '09.Ninth International Conference on, 2009: 1257-1262.

[24] Chakrabarti T, Saha S, Sinha D. DNA Multiple Sequence Alignment by a Hidden Markov Model and Fuzzy Levenshtein Distance based Genetic Algorithm[J]. International Journal of Computer Applications, 2013: 73.

[25] Naznin F, Sarker R, Essam D. Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment[J]. Evolutionary Computation, IEEE Transactions on, 2012, 16(5): 615-631.

[26] 张鑫源，胡晓敏，林盈. 遗传算法和粒子群优化算法的性能对比分析[J]. 计算机科学与探索，2014(1)：90-102.

[27] Zou Q, Shan X, Jiang Y. A Novel Center Star Multiple Sequence Alignment Algorithm Based on Affine Gap Penalty and K-Band[J]. Physics Procedia, 2012(33): 322-327.

[28] 杨锡南，孙啸. 生物信息学中基因数据可视化[J]. 计算机与应用化学，2001(18)：403-410.

[29] 贺向敏，周根宝. 基于遗传算法的多序列比对算法研究[D]. 内蒙古农业大学，2010.

[30] 张琎，张远. 基于 GC-GM 的多序列比对穷举遗传算法[J]. 计算机应用，2010，30(1)：146-149.

[31] 刘立芳，霍红卫，王宝树. PHGA-COFFEE：多序列比对问题的并行混合遗传算法求解[J]. 计算机学报，2006，29(5)：727-733.

[32] Fan H, Wu R, Liao B, et al. An Improved Genetic Algorithm

for Multiple Sequence Alignment[J]. Journal of Computational and Theoretical Nanoscience, 2012, 9(10): 1558-1564.

[33] Thomsen R, Boomsma W. Multiple sequence alignment using SAGA: investigating the effects of operator scheduling, population seeding, and crossover operators[M]. Applications of Evolutionary Computing. Springer Berlin Heidelberg, 2004: 113-122.

[34] Yanping Lv, Shaozi Li, Changle Zhou, etal. Improved Genetic Algorithm for Multiple Sequence Alignment Using Segment Profiles (GASP). Computer Science, 2006, 4093(2006), 388-395.

[35] 泽瓦勒贝 M，鲍姆 J O. 理解生物信息学[M]. 北京：科学出版社，2012.

[36] 陈铭. 生物信息学[M]. 北京：科学出版社，2012.

[37] Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput[J]. Nucleic Acids Research, 2004, 32(5): 1792-1797.

[38] 杨晶. 生物计算[M]. 北京：科学出版社，2010.

[39] 刘立芳. 生物信息学中的多序列比对与模体识别问题研究[D]. 西安电子科技大学，2006.

[40] 徐小俊. 群智能优化算法在多序列比对中的应用[D]. 陕西师范大学，2011.