

第三节 数据库搜索——BLAST 和 FASTA 应用

一. 数据之海与一叶轻舟

《科学》(*science*)杂志在 2001 年 2 月 16 日的人类基因组专刊上发表了一篇题为“生物信息学：努力在数据的海洋里畅游”的文章，文章写到：“我们身处急速上涨的数据海洋中...我们如何避免没顶之灾？”一条可靠的办法可能是赶紧找到“一叶轻舟”，而且在轻舟上装上先进的电子设备，诸如卫星定位系统、卫星信息传输系统等等.....BLAST 和 FASTA 便是这样的一条“轻舟”，它们往来穿梭，速度奇快。

比较和确定某一数据库中的序列与某一给定序列的相似性是生物信息学中最频繁使用和最有价值的操作。本质上这与两条序列的比较没有什么两样，只是要重复成千上万次。但是要严格地进行一次比较必定需要一定的耗时，所以必需考虑在一个合理的时间内完成搜索比较操作。目前有二个最为常用的程序服务于未知序列的数据库相似性搜索，即 BLAST 和 FASTA。FASTA 使用的是 Wilbur-Lipman 算法的改进算法，进行整体联配，重点查找那些可能达到匹配显著的联配。虽然 FASTA 不会错过那些匹配极好的序列，但有时会漏过一些匹配程度不高但达显著水平的序列。BLAST(Basic Local Alignment Search Tool, 基本局部联配搜索工具)是基于匹配短序列片段，用一种强有力的统计模型来确定未知序列与数据库序列的最佳局部联配。

大多数研究目前都通过国际互联网 Internet 应用 NCBI 研制的 BLAST 程序 (Basic Local Alignment Search Tool) 来进行 DNA 和蛋白质序列相似性搜索。用一组 BLAST 程序联配可以快速进行核酸和蛋白质序列库的相似性检索。采用 BLAST 的基本算法编成了若干各不同的程序，分别使用特定的序列库和用于特定类型的输入序列。BLASTN 是在核苷酸序列库搜索核苷酸序列。BLASTP 是在蛋白质序列库中搜索氨基酸序列。TBLASTN 则可以在核酸序列库中搜索氨基酸序列，此时序列库在搜索之前要按所有 6 种读框即时翻译。与此相反的一项分析则由 BLASTX 来完成，它要将所输入的核酸序列按所有 6 种读框翻译，然后再以之搜索蛋白质序列库。近期 Altschul S.F. 等人 (1997) 提出了一个通过寻找蛋白质家族保守序列来提高算法敏感性的 PSI-BLAST (Position-Specific Iterated BLAST) 算法，并开发了相应的软件。PSI-BLAST 可以对数据库进行多轮循环检索，每一轮的检索速度都大约是 BLAST 的两倍，但每一轮都能提高检索的敏感性。它是目前 BLAST 程序家族中敏感性最高的成员。

表 3.15 数据库相似性搜索程序 BLAST 和 FASTA 程序清单

程序 (Program)	待检序列类 型 (Probe type)	数据库类型	说 明 (Comment)
BLASTP	p	p	在蛋白质序列库中比对待检蛋白质序列
BLASTN	n	n	在核酸序列库中比对待检核酸序列
BLASTX	n	p	在蛋白质序列库中比对待检核酸序列 (用所有 6 种读框翻译)
TBLASTN	p	n	在核酸序列库(用 6 种读框即时翻译)中 比对待检蛋白质序列
TBLASTX	n	n	在核酸序列库(用 6 种读框即时翻译)中 比对待检核酸序列(同样用所有 6 种读 框翻译)
FASTA3	p	p	在某一蛋白质序列库中搜索蛋白质相 似序列
	n	n	在某一核酸序列库中搜索核酸相似序 列
TFASTA3	p	n	在核酸序列库(已被即时翻译)中比对 待检蛋白质序列
FASTX3	n	p	在蛋白质序列库中比对待检核酸序列 (用 6 种读框翻译)
TFASTX3	p	n	在核酸序列库中比对待检蛋白质序列
SSEARCH	P/n	P/n	使用 Smith-Waterman 算法联对比对

注：n：核酸序列或核酸序列库；p：蛋白质序列或蛋白质序列库

如果目的序列中有蛋白质编码区，则用翻译的蛋白质序列来搜索蛋白质序列库要比用 DNA 序列搜索核酸序列库更有价值。由于蛋白质序列的进化要比 DNA 序列慢一些，在蛋白质序列水平上的远缘关系在 DNA 水平上可能被错过。如果无法确定编码区，则可利用 BLASTX 按所有 6 种读框来翻译 DNA 序列，然后用它搜索蛋白质序列库。由于蛋白质序列库仅包含已鉴定的蛋白质，所以必须采用 TBLASTN 程序在现有的 GenBank、EMBL 或 DDBJ DNA 序列库中检索新确定的氨基酸或翻译过的 DNA 序列。这种检索有时可以找到一些显著相似的 DNA 序列，而原本并不知道这些序列可编码蛋白质。

BLAST 的一项重要特性就是所报告的匹配序列的统计学显著性评分。这一统计学显著性评分是用 Karlin-Altschul 算法决定的，所算出的 Poisson 概率表明所得到的序列相似性随机出现的可能性。

另一个常用的核酸和蛋白质序列库搜索程序是 FASTA，即 FASTN 和 FASTP 程序的新版本。FASTA 首先在序列库中进行快速的初检，找出与待检序列高度相似的序列。这一快速检索局限于待检序列和序列库序列之间较短的完全相同序列区段上。

FASTA 首先要建立一个其长度由 K-tuple (ktup) 值决定的所有可能的总表或字典。这一程序中使用的字长参数(或 K-tuple)表示所用的初始相配序列长度。

K-tuple 的大小可以变化并将间接影响搜索的速度和敏感度。然后程序要对待检序列和序列库中的所有序列进行处理,找出字典中长度与 K-trple 相等的所有序列段的位置。比较两个序列的字典要比比较两个序列本身快得多,可以有效地找出小段相似区。一旦通过初始的快速检索找到一批评分最高的序列,就可以仅对这些高分序列进行第二轮比较。第二轮的序列比对是采用 Needleman-Wunsch 算法(1970)进行空位联配计算,得出分析的最后结论。如果 FASTA 运行后找到较好的相似序列,有时采用较小的 K-tuple 值或换一个评分矩阵重新检索分析,也许会有帮助。

在终端计算机上 FASTA 检索的一个简便方法是使用电子邮件服务。有好几个机构都可以通过电子邮件自动地接受 FASTA 计算机检索要求,用电子邮件中所提供的序列,对多个序列库进行搜索,然后又通过电子邮件将结果送回。图 7.7.5 就是一人要求 FASTA 服务的电子邮件示范。BLAST 检索同样可以通过电子邮件或 Internet 服务进行。

```
TITLE A test search of the EMBL other Mammalian DNA sequences
LIB EMAM
WORD 4
LIST 100
ALIGN 20
SEQ
tgcttggtgaggagccataggacgagagcttctggtgaaagtgtgtttcttgaatcagcaccaccatg
gacagcaaa
END
```

图 3.6 送往 EBI FASTA 电子邮递服务中心(电子邮件地址: fasta@ebi.ac.uk) 的一份邮件的内容。这份邮件要求用该序列对 EMBL 序列库中的其它哺乳动物序列进行检索。送回的答案中包括 100 条最匹配序列和头 20 条最匹配序列的联配结果。

不论是 FASTA 还是 TFASTA 都提供一项评分,以评价用前述 PAM250 矩阵生成的每一对联配序列的匹配程度。但 FASTA 并不像 BLAST 程序那样给出一项显著值。无论采用 FASTA 或 BLAST,推断相似性是否具有生物学意义都取决于研究者。要作出决断,必须充分考虑蛋白质已知的或推断的功能,与已知活性位点或模序的相似程度等等。

因为 BLAST 和 FASTA 采用不同的算法,同时用这两种搜索引擎重新检索某一特定序列往往是可取的。如果用其中一种找不到显著相似序列,不妨试一试另一程序。如果 BLAST 和 FASTA 均找不到显著匹配的序列,还可以选择第 3 条比较费时的搜索策略。一些网站允许用户使用基于 Smith-Waterman 算法的搜索程序,如 BLITZ。BLITZ(www.ebi.ac.uk/searchs/blitz.html)被设计在大型并行计算机上运行,因此使检索更灵敏。虽然运行这样的程序比较费时,但它们有时会发现一些被 BLAST 和 FASTA 错过的勉强达到显著的联配。

由于数据库相似性搜索是生物信息学最为重要的组成部分,所以很多网站都提供了 BLAST 和 FASTA 搜索服务。在选择何种数据源时,有很多标准可以应用。并非所有的 BLAST 和 FASTA 均提供相同的服务,你所搜索的数据库各不相同,这就如同我们有多种替换矩阵一样。另外,一些网站还为熟练使用者提供了特别服务。总之,在一些非冗余序列数据库中搜索均是被允许的。这类数据库至少包

括在 SWISS-PROT 和 PIR(蛋白质)或 EMBL 和 GenBank(核酸)的所有记录,这往往是最佳选择。但不要滥用这些资源,例如,如果你正在构建序列重叠群(contig),则只需进行最终组合序列的 BLAST 或 FASTA 搜索即可,而不必对每个序列片段均进行搜索。同样,为了查找克隆载体的污染序列而进行整个非冗余数据库的 BLAST 运行,也不是一个有效办法。

二. BLAST : 核苷酸数据库搜索

BLAST 包含有 5 个子程序,它是目前运行速度最快的检索搜索程序。最初的程序版本(Version1.4)不允许设置空位(gap),这对运行速度的提高有好处。正如前文所述,空位直接关系到搜索结果,所以目前的 BLAST 版本(Version2.0)均能进行空位联配。BLAST 的快速得益于它的统计算法:BLAST 使用的是快速局部而不是缓慢、整体的联配策略。BLAST 不追求整条序列的匹配。

1. BLAST 实战操作(1)

如果这是你初次使用 BLAST,那不妨先按以下要求操作一次,先有个感性认识,然后再进一步了解和认识其细节:

在 Internet 中进入 EXPASY BLAST 主页

Basic BLAST	Advanced BLAST
-------------	----------------

Basic BLAST

Usage: Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported [formats](#) into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...
 Make sure that the format button (next to the sequence field) shows the correct format .
 See also our [BLAST database description](#).

Please select the program: [Program](#)

Please select the database:

DNA databases

Protein databases

Gapped alignment on/off [Select matrix](#)

BLAST filter on/off [Select format](#)

Graphic output on/off [Query title \(option\)](#)

Paste your sequence here:
(or ID or accession number)

[required for tblast\[nx\] programs ->](#) [E-mail address](#)

HTML

假如有一条人类基因序列,对这条序列我们一无所知。序列的提供者想对这

条序列进行常规分析来鉴定它。你可以这样进行：

复制该序列：

```
AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTGAGGTTAGGTAA
TATGGTTTGGTATCCCTGTAGTTAAAAGTTTTTGTCTTATTTTAGAATACTGTGAT
CTATTTCTTTAGTATTAATTTTTCTTCTGTTTTCTCATCTAGGGAACCCCAAGA
GCATCCAATAGAAGCTGTGCAATTATGTAATAATTTCAACTGTCTTCTCAAATA
AAGAAGTATGGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAGAAGCACAGA
ATATTTTTTATATTTCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAAT
TTCCTTTTTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGC
AAGATGTGAATTATTGACACTGAACTTAATAACTGTGTACTGTTCCGAAGGGGTTT
CTCAAATTTTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTTAAGTTCTTA
TGAGGAGGGGAGGGTAAATAAACCACTGTGCGTCTTGGTGTAATTTGAAGATTGCC
CCATCTAGACTAGCAATCTCTTCATTATTCTCTGCTATATATAAAAACGGTGCTGTG
AGGGAGGGGAAAAGCATTTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTA
ATAAGCAATCAAGGTTATAATTTTTTTTTAAAATAGAAATTTTGTAGAAGGCAATA
TTAACCTAATCACCATGTAAGCACTCTGGATGATGGATTCCACAAAACCTGGTTTT
ATGGTTACTTCTTCTCTTAGATTCTTAATTCATGAGGAGGGTGGGGGAGGGAGGTG
GAGGGAGGGAAGGGTTTCTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTG
TAACTTGCTTAAATTTCTTATGTGACATTAACAAATAAAAAAGCTCTTTTAATATTA
GATAA
```

进入 EXPASY(EMBLnet)BLAST 服务器主页。如果因故不能进入该服务器，也可使用其它网站的 BLAST 服务，但以下仅以 EXPASY BLAST 服务器为例；

选择相关程序：BLASTN。该程序是在核酸数据库中进行相似核酸序列的搜索；

选择数据库：EMBL without ESTs(DNA)。这是 EMBL 的主要核酸数据库；

缺省替换矩阵选项：在 BLASTN 中不必应用矩阵；

选择序列输入格式：Plain TEXT。以文本格式发送核酸序列；

按如下选定： Gapped Alignment ON

BLAST filter: ON

Graphic Output: ON

粘贴未知序列到输入框(Paste your sequence here:)内；

按下运行按钮：Run BLAST；

等待，并检查运行结果。

2. BLAST：结果报告

BLAST 的结果报告可能显得零乱，但是最主要的部分非常容易抓住。在报告的上部是有关程序的描述(如 BLASTN)、程序的版本和相关信息，接下来是你输入的未知序列，如果部分序列片段在过滤时未通过，则可看到一串 N 序列片段。再下来的几行提供了你搜索的数据库信息，包括该数据库最新更新时间。最后部分是在“searching”和“done”之间一系列(共 50 个)点(、)，如果是星号(*)，则表示程序在搜索该数据库时发生了障碍，少于 50 个点表明程序未能搜索整个数据库。这些因素必须予以考虑，你可能考虑重新运行一次。

在“Searching...Done”行下，你将看到一幅图。图最上面红色一条线代表未知的待搜索序列。在该线上有一个刻度，刻度下的数字为序列长度。其它不同颜色的线分别代表数据库中与之相似性显著的序列。可以看到，在本搜索进行的时候，数据库中只有一条与未知序列长度相仿的序列被列出，而其它找出的序列

均很短。由该图得出这样的结论：数据库中只有一条序列与你的未知序列有高度的相似性。

结果报告的再下面是一行一行的达到联配显著的序列描述。其中第一行(代表上图中与未知序列等长度的序列)如下：

**emb|L37747|HSLAM11 [Homo sapiens]Homo sapiens lamin B1 gene,
ex... 416 e-114**

在这行描述中，E值(E-value)很重要，它是一行中最近面的一个数字(e-114)。E-114 可以表示为 1×10^{-114} 或者说就是非常之趋近于零。这个数值表示你仅仅因为随机性造成获得这一联配结果的可能次数。这一数值越接近零，发生这一事件的可能性越小。从搜索的角度看，E值越小，联配结果越显著。

我们知道我们列举的序列来自人类，所以在以上结果中只有第一和第二行的序列是我们想要的。其它序列的E值较大，说明这些匹配结果很有可能是随机产生的，而且绝大部分序列来自其它生物。

注意！“Lamin”基因的序列很特别。当你搜索你自己的序列时，可能会得到1个以上匹配极好的序列，但是，统计上最显著的(E值最小)并不总是你所要找的序列。应注意短的重复序列和模序家族(motif family)，它们可能不被统计联配算法(如BLAST)看中。只把显著性当作一种导向，结合你的分子生物学知识和序列来源，你的人为判断能力在数据库搜索时总是有用的。

我们在报告中可以进一步看到实际的联配情况。它们的排序与上面各行的排序是一致的。一个短序列联配的例子：

```
Query:   1 ggccccaccacgccgctcag 20
          |||
Sbjct: 701 ggccccaccacgccgctgag 720
```

我们看到，未知序列与目标序列间几乎100%匹配。一条竖线(|)连接两个碱基，表明它们是相同的。在未知序列中的第18个碱基C与数据库找到的匹配序列的第18个碱基G不相同，它们之间是空的，没有竖线。在其它的一些联配中，可以看到很大片的空缺。序列间不匹配除了缺少同源性外，还可能存在其它一些原因，如测序错误、未知序列的点突变等等。

我们这次的检索结果明白无误地告诉我们，第一条序列与我们的未知序列是一样的。回到报告中对序列的行描述部分，可以点击EMBL的身份号(HSLAM11)并查阅EMBL的数据库记录，记录中包括了该序列的相关信息，例如它所编码的蛋白质序列等。

3. BLAST 选项

我们回到EXPASY BLAST服务器主页，点击“Advanced BLAST”按钮，将出现一个有很多选项的页面。对于大多数搜索，最佳选项设置往往已被设为缺省状况，但是你可以方便地改变这些设置进行一些必要的搜索研究。我们需要准确地理解这些选项的真正含义。

“WORDLENGTH” (字长)选项：

BLAST程序是通过比对未知序列与数据库序列中的短序列来发现最佳匹配序列的。最初进行“扫描”(scanning)就是确定匹配片段。序列的匹配程序由短

序列(定义为“word”,即字)的联配得分总和来决定。联配时,“字”的每个碱基均被计分:如果碱基对完全相同(如A与A),得某一正值;如果碱基对不很匹配(W与A或T),则得某一略小的正值;如果两个碱基不匹配,则得一负值。总的合计得分便决定了序列间的相似程度。

得分高的匹配序列被称为高比值片段对(high-scoring segment pairs, HSP)。BLAST程序在两个方向扩展HSP,直至序列结束或联配已变为不显著。替换矩阵在扫描(scanning)和扩展过程被应用。最后在BLAST报告中被列出的序列都是所有得分最高的序列。

以上述及的初始字长便是由W(WORDLENGTH)值设定。BLAST只对字长为W的“字”进行扩展联配。BLAST的字长缺省值为11,即BLASTN将扫描数据库,直到发现那些与未知序列的11个连续碱基完全匹配的11个连续碱基长度片段为止。然后这些片段(即字)被扩展。11个碱基的字长已能有效地排除中等分叉的同源性和几乎所有随机产生的显著联配。

“Filter”(过滤器)选项:

BLAST2.0版本已有序列过滤器功能。过滤器将锁定诸如组成低复杂(low compositional complexity)序列区(如Alu序列),用一系列N(NNNNNN)替代这些程序。N代表任意碱基(IUB-code)。只有未知待检序列被过滤替代,而数据库的序列将不被过滤。

过滤对绝大多数序列都是有益的,“Filter”项的缺省选项为ON。例如,多A碱基的尾部和脯氨酸富积的序列,会得到人为的高联配得分而误导分析。这是因为这类序列数量极大,遍布整个基因组,直至整个数据库。

“Matrix”(矩阵)选项:

如前所述,联配的显著性是由返回的比对分值决定的,该分值反映的是所得到的联配随机产生的概率有多大。矩阵被用于鉴别数据库中的序列,同时又用来预测匹配的显著性大小。一般应接受运行程序推荐的矩阵。BLAST系列程序主要使用两种类型矩阵(PAM和BLOSUM)。要准确地选择矩阵,必须了解矩阵和矩阵的具体计分方式。这方面的知识可参阅本章第二节“替换矩阵”部分。

注意!直接比较使用不同替换矩阵而获得的联配得分是没有意义的。同时,你可以为BLAST、TBLASTN或TBLASTX选择不同的矩阵,例如PAM30、PAM70、BLOSUM80、BLOSUM62等等,但是BLASTN不需要这些矩阵,搜索时,不必选定。

“EXPECT”选项:

你可能会想为搜索设定一个期望值阈值(EXPECT),例如缺省值设为10。这一设置则表示联配结果中将有10个匹配序列是由随机产生,如果联配的统计显著性值(E值)小于该值(10),则该联配将被检出,换句话说,比较低的阈值将使搜索的匹配要求更严格,结果报告中随机产生的匹配序列减少。

“Score Value”(分值)选项:

在“WORDLENGTH”选项中曾论及碱基对匹配程度的赋分问题,其赋分的标准可由分值选项的M和N两个参数设置。M参数为匹配碱基的赋值,必需为一正整数;N参数为不匹配碱基的赋值,必需为一负整数。

M/N的比率决定了你所接受的进化分歧程度(degree of divergence),M和N的缺省值为5和-4。该比率(1.25)相当于在100个残基中约有47可以观测到的核酸点突变(PAM)。PAM是被用来预测分子序列从祖先序列进化而来的程度。如果你调整M和N使比率提高,则PAM矩阵也应选择大些(指PAM矩阵后的数字),以适应相应的较大的分歧程度。

输入框选项

你也许已注意到,在序列的输入框内可以键入 EMBL 的身份号 (ID) 或 GenBank 的记录号 (accession number)。这样的输入选择将仅返回数据库中的某一序列资料 (最新版本), 该序列与键入的记录号相对应。在不少情况下需要类似检索, 例如核对 PCR 产物。其它一些选项情况可参阅 BLAST 的在线使用手册。

4. BLAST 实战操作(2)

写出以下几个问题的答案, 然后与随后的答案比较一下:

复制以下序列, 运行 BLAST 程序搜索, 鉴别该序列。除必需改变的设置, 使用缺省设置。提示: 你可能需要选择某一数据库和 BLAST 程序。

```
GTCCGGCCTGGGCGACAGAGCAAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA  
AAAAAAAAAA
```

该序列取自 GenBank 的一个记录 (记录号为 S56967)。使用 BLAST 服务器找到该记录。

仔细察看以上序列, 你会发现未能鉴别出该序列并没有什么奇怪。该序列是一条 Alu 序列! 所以有如此多的匹配序列。

再复制该序列并使用 BLASTX 程序。该程序是将待检序列翻译成蛋白质序列 (6 种读框), 然后在蛋白质序列库 SWISS-PROT 中进行联配搜索。

选择 BLASTX、SWISS-PROT 等选项并运行后, 检索结果与上次结果略有不同。分析得到的结果 (BLAST2), 你可以知道该序列为 Alu 序列, 你可能需要测定该基因的其它不同的片段或一条更长的扩展片段, 以便能真正鉴别它。

复制以下序列并运行 BLAST 搜索, 检查检索结果。

```
GAATTCTAATCTCCCTCTCAACCCTACAGTCACCCATTTGGTATATTAAGATGTGT  
TGTCTACTGTCTAGTATCCCTCAAGTAGTGTGTCAGGAATTAGTCATTTAAATAGTCTG  
CAAGCCAGGAGTGGTGGCTCATGTCT
```

你将能从检索结果中确定该序列编码的是人 β -血球蛋白 (beta heamoglobin)。在写作本书时, 检索结果中前 2 条序列不仅匹配程度很好 (100% 和 99% 同源), 而且它们与以上序列长度也一致。其它的匹配序列都很短。这很清楚地说明这是个 β -血球蛋白基因, 但是第二条序列中有一个 C 碱基与第一条不相同, 这提醒你在最后确定前应该再检索一下你的测序结果是否正确。

问题 中的序列非常特异, 如果同该序列的前 15 个碱基去搜索是什么样的结果?

```
GAATTCTAATCTCCCTCTCAACC
```

没有发现任何线索! 这很奇怪, 因为我们刚刚进行了检索。这一情况告诉我们这样一个事实: 检索结果中没有匹配的序列 (“NO Hits”) 并不一定是数据库中没有这些序列, 而是可能因为检索标准设置的问题。

再复制这 15 个碱基序列, 回到 BLAST 主页。点击 “Advanced BLAST” (高级 BLAST) 按钮, 使用 EMBL 数据库 “nr” 亚类, E 值调整到 100, 关掉 (off) “XBLAST - repsim filter” 过滤器, 然后运行。

在写作本书时，结果中只有 2 个匹配序列，即以上搜索到的 2 条序列。由此可以确定该序列极有可能是 α -血球蛋白基因。

三. BLAST : 蛋白质数据库搜索

蛋白质数据库搜索是应该掌握的最重要的生物信息学技能，因为该搜索的灵敏性大约是核酸数据库搜索的 2-5 倍。蛋白质数据库搜索灵敏性好的原因包括：DNA 密码只有 4 个，在每个位置上的密码只有 4 种可能，而蛋白质有 20 种可能；遗传编码的多样性， n 个三联体密码编码一种氨基酸；虽然某一蛋白质序列与你的未知序列相同，但是你不能得到一个明确匹配的 DNA 序列。另外，蛋白质序列的相似性比 DNA 序列更保守。

蛋白质的直系同源性(orthologue)检索已越来越成为分子生物学的重要组成部分。目前一种酵母(*Sacharomyces cerevisiae*)和一种线虫(*Caenorhabditis elegans*)等的基因组序列已完成，同源性分析已在有效地进行中。如果人类的某一特异蛋白与以上的某一同源族相匹配，则可以确定该蛋白的可能功能，这将节省大量研究时间和经费。这一方面研究已有很好的例子(可参阅 Chervitz SA, et al. Comparison of complete protein sets of worm and yeast: orthology and divergence. *Science*. 1998, 282:2022-2028)。

两个主要蛋白质数据库(PIR 和 SWISS-PROT)的记录没有象三个主要核酸数据库一样相互交换。两个数据库各有优缺点，你必须考虑选择合适的数据库进行搜索。

下面我们进行 SWISS-PROT 数据库的进行实战搜索操作：

在以下的操作中，你将结合以上有关 BLAST 的知识，学会如何从相关数据库中获取信息。

选择序列联配程序：你可能选择 BLAST 或 FASTA 服务器，本例选 BLAST。

复制以下人类蛋白质序列：

```
MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNQAISLIFSLKEEVGALAKVLR
LFEENDVNLTHIESRPSRLKKDEYEFFTHLDRSLPALTNIKILRHDIGATVHE
LSRDKKKDTVPWFPRTIQELDRFANQILSYGAELDADHPGFKDPVYRARRK
QFADIAYNYRHGQPIPRVEYMEEEEKKTWGTVFKTLKSLYKTHACYEYNHIFP
LLEKYCGFHEDNIPQLEDVSQLQTCTGFRLRPVAGLLSSRDFLGGLAFRVF
HCTQYIRHGSKPMTPEPDICHELLGHVPLFSDRSFAQFSQEIGLASLGAPD
EYIEKLATIWFTVEFGLCKQGDSIKAYGAGLLSSFGEQYCLSEKPKLLPLEL
EKTAIQNYTVTEFQPLYVAESFNDAKEKVRNFAATIPRPFSVRYDPYTQRIE
VLDNTQQLKILADSINSEIGILCSALQKIK
```

粘贴以上序列到输入框内并调整相关选项。你不必进行高级 BLAST 检索，但你必须选择数据库和程序。本例选 SWISS-PROT 数据库。

运行 BLAST。使用提供的链接功能阅读检索结果报告。获得的报告可以用 NICE-PROT 阅读，其界面更友好和完整。

回答以下问题。你也许需要点击与 SWISS-PROT 报告链接的相关数据库信息。

回答这些问题需要一定的时间，但它可以使你明白你能得到哪些信息并如何

得到它们。

问 题	答 案
该记录的 SWISS-PROT 名称是什么？	PH4H_Human
SWISS-PROT 最初的记录号是多少？	P00439
该蛋白的最普通名？	Phenylalanine-4-Hydroxylase
该基因名称？	PAH
哪一年该催化功能区的晶体结构被确定？作者是谁？	1997, Erlandsen
该酶发挥功能是否需要协因子(co-factor)?如果是, 是哪个因子？	是, ferrous ion
与该酶缺失直接有关的最普通疾病是什么？	Phenylketonuria(PKU)
该基因的细胞遗传学位点？(例如 13p10.1)	12q24.1
PAHdb 是什么？	PAH 突变体数据库
该蛋白质有多少氨基酸残基？	452
该蛋白质的分子量是多少？	51.862kDa
如何得到该蛋白质的三维图象？	进入 PDB 数据库

获得的以上答案的正确操作：选择 BLASTP 和 SWISS-PROT，结果显示只有人类的 PAH1 序列与未知序列完全(100%)相同，由此可以确定未知序列。点击 PAH1 记录号，进入 SWISS-PROT 数据库，查阅该记录信息。在此处可用 NICE-PROT(点击)阅读。观察三维图象时，可在“Cross-references”(交叉文献)下点击 PDB 数据库链接按钮。

四. FASTA：另一种搜索策略

1. FASTA算法¹

FASTA 的原型是 David Lipman 和 William Pearson(1985)提出的用于蛋白质同源比较的 FASTP。FASTA 提高了 FASTP 的灵感性但速度并没有损失多少(pearson and Lipman, 1988)。它可以用来进行 DNA 对 DNA，DNA 对蛋白质(将 DNA 按 6 个读框“翻译”成氨基酸序列，再与蛋白质比较)和蛋白质对蛋白质的同源比较。下面以两条氨基酸序列的比较为例介绍算法的基本思路。

算法可以分为 4 步：

第一步：

FASTA 首先找出进行比较的两条序列所有长度为 K-tuple 的连续的一致序列片段。例如以下两条蛋白质序列：

序 列	位 置						
	1	2	3	4	5	6	7
1	F	L	W	R	T	W	S
2	T	W	K	T	W	T	

设 K-tuple = 2，则序列 2 中有两个符合条件的片段(用下划线表示)，相对于序列 1 的偏移(offset)分别是 4 和 1 [对于一对开始位置为 (x_1, x_2) 的一致片段，偏移

¹本部分内容主要取自 F. 奥斯伯, R. E. 金斯顿等. 精编分子生物学实验指南, 北京: 科学出版社, 1998

定义为 x_1-x_2 。在上例中有两对 (x_1, x_2) ，即 $(5, 1)$ 和 $(5, 4)$ 。这种片段的一致性可以表示为对角线图，两条序列中的一对一致片段在图中表示为一段对角线。(图 3.5)。

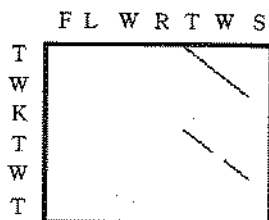


图 3.5 序列 FLWRTW 和 STWKTWT 比较形成的对角线图

本例是两条非常短的氨基酸序列，在实际比较长的蛋白质序列或 DNA 序列时，对角线图如图 3.6A 所示。

对于图中每一条完整的对角线(即同一偏移)上的一致片段，如果片段间距小于用户界定的界限，则将片段连接起来作为一条一致片段。对这些片段进行计分，每一对对应的元素，一致的加分，不一致的扣分。完成了所有一致片段的计分后，选出 10 条分值最高的片段进入下一轮计算，如图 3.6B。

第 2 步：

FASTA 将这 10 对片段重新计分。这轮计分允许保守突变，对蛋白质来讲，就是使用 PAM250 等替换矩阵。简单地说，替换矩阵就是对应于 20×20 种氨基酸替换(比如 R 替换成 P)的计分规则所构成的 20×20 的矩阵。这种矩阵是从蛋白质进化实例中总结出来的经验矩阵，它给予进化上相对保守的氨基酸替换比非保守的替换更高的分值。在重新计算分值后，在每一条这样的片段中找出分值最高的子片段，作为“初始区域”(initial region)进入下一步。在 initial region 中，最高的分值计为 $initl$ 。

第 3 步：

在这一步中，FASTA 选出分值高于用户确定的界限且相互之间不重叠的初始区域，并尝试将这些初始区域连接起来。当然，由于连接而出现的缺失和插入情况要作相应的扣分。FASTA 在这一步才考虑插入和缺失的情况，最终找出能够得到的最高分值的初始区域或连接起来的数个初始区域。这一步计算出的最高分计为 $initn$ 。见图 3.6C。

第 4 步：

以 $initl$ 片段或($initn$ 的片段)为中心，向前后延伸一定的长度。在这样一个区域中(见图 3.6D 中虚线间的区域)应用 Smith-Waterman 算法进行重新对齐，最终的得分计为 opt 。

在实际操作中，用户可以在需要达到的灵感性程度和所需时间之间进行权衡(一般来说，要达到更高的敏感性总是需要更长的运算时间)，决定采用 $initn$ 还是 opt 作为两条序列相似程度的分值。研究表明：使用 $initn$ 与使用 opt 相比，前者损失的敏感性并不太大，但运算速度却快得多(Pearson WR, 1991)。

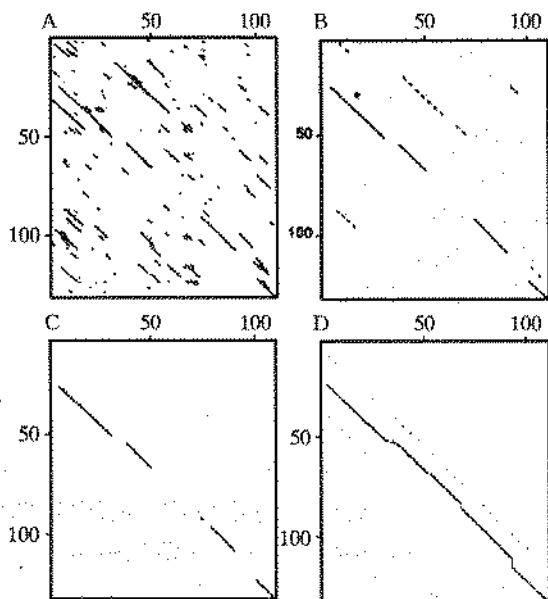


图 3.6 FASTA 的 4 步 (A-D) 算法图示

2. FASTA 选项

只有 10 个碱基长度的短序列可以用 FASTA 检索。其检索速度主要由 KTUP 值决定，该值用于限定字长。在 BLAST 中，“字”(word)表示用于联配的短序列片段，高比值的字(HSP)被选定并进行扩展。在 FASTA 中，字不被计分，联配只有在完全匹配的情况下再继续下去。

“Matrix”(矩阵)选项：

FASTA 和 BLAST 在最初扫描和扩展(FASTA 只使用扩展)阶段主要的不同之处是 FAST 允许多义密码子(IUB ambiguity codes)包括其中。大多数 FASTA 服务器提供了 BLOSUM 或 PAM 系列替换矩阵。FASTA 的缺省推荐矩阵是 BLOSUM50。如果 BLOSUM50 不合适，BLOSUM62 是另一个可行的选择。如果你在进行突变性质的进化分析时，不妨试试 PAM。

Smith-Waterman 算法：

你可以选择比以上讨论的算法更严格的算法进行联配。SSEARCH3 程序使用了 Smith-Waterman 算法，适用于高精度的检索，但运行速度非常慢，但是在你有类似搜索需要时，它无疑是值得应用的。往往要求键入一个 E-mail 地址，以便将搜索结果通过 E-mail 发送给你。

“KTUP”选项：

KTUP 值(字长)可在 1-6 整数之间选择。KTUP=2 的选项设置将是 KTUP=1 的设置搜索速度快 5 倍，因为 KUTP=1 时，服务器将对每个碱基进行联配，而 KUTP=2 时，则以 2 个碱基进行联配。有些服务器限制 KUTP 设置必须在 3-6 之间。缺省设置为 KUTP=6。

“GAOPEN”和“GAPTEXT”(空位设置与空位扩展)选项：

与 BLAST 一样，空位设置和空位扩展的罚值必须为负值，它们的缺省设置分别为 -16 和 -4。注意：FASTX 和 TFSTX 对移码(frameshift)也可以设置罚值。

“STRAND”(转向)选项：

正常情况下，STRAND 设置为“upper”，另外选项为“bottom”。如果选

“bottom”，则 FASTA 将对未知待检序列转向后再进行搜索。

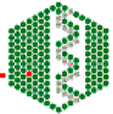
3. FASTA 的实战操作及其结果报告

如果你已经进行过 BLAST 的实战操作并读懂了其结果报告，则 FASTA 就不在话下了。此处仅给出了一条例举序列及其 FASTA 结果报告，不妨自己试试。

进入 EMBL 的 EBI 网站 FASTA3 服务器，并复制下列未知 DNA 序列：

```
CCAGATCCTGGACAGAGGACAATGGCTTCCATGCAATTGGGCAGATGTGTGAGGCACCTGTGGTGACC
```

EMBL
European Bioinformatics Institute



Fasta3 [Help](#) [Tools](#) [EBI Home](#) [Run Fasta3](#) [RESET FORM](#)

YOUR EMAIL	SEARCH TITLE	RESULTS	DNA STRAND	MATRIX
	Sequence	interactive	none	BLOSUM50
GAP PENALTIES	SCORES & ALIGNMENTS	KTUP/HISTOGRAM	PROGRAM	DATABASES
OPEN -12 RESIDUE -2	SCORES 50 ALIGNMENTS 50	KTUP 2 HIST no	fasta3 fastx3 fasty3 fast3 facts3	Protein swall swiss-prot swiss-new sptrembl

Enter or Paste a DNA/RNA Sequence in any format:

```
CCAGATCCTGGACAGAGGACAATGGCTTCCATGCAATTGGGCAGATGTGTGAGGCACCTG
TGGTGACC
```

首先选择 fasta3 程序、EHUM 数据库和 bottom (STRAND) 选项，运行 fasta3。可以得到如下结果：

```
FASTA (3.39 May 2001) function [optimized, +5/-4 matrix (5:-4)] ktup: 6
join: 45, opt: 30, gap-pen: -16/-4, width: 16
Scan time: 54.270
The best scores are:                                     opt bits E(145773)
EM_HUM: AF015262 AF015262 Homo sapiens Down Syn (79920) [r] 125 36 0.83
EM_HUM: HS229043 AJ229043 Homo sapiens 959 kb c (48446) [r] 125 36 0.96

>>EM_HUM: AF015262 AF015262 Homo sapiens Down Syndrome cr (79920 nt)
rev-comp initn: 74 init1: 74 opt: 125 Z-score: 120.7 bits: 36.3 E(): 0.83
67.164% identity (68.182% ungapped) in 67 nt overlap (68-2:209228-209293)

          60      50      40
EMBOSeq-   GGTCAACCACAGGTGCTCACACATCTGCC
          :::: :::: : : :: :::: :: :::: ::::
EM_HUM GCACCAACCGTGTCCAGGCTCTCTCAGGTGGTCTCCATAACTACCCCACTCACCTGCC
      209200   209210   209220   209230   209240   209250

          30      20      10
EMBOSeq-   AATTGCATGGAAGCCATTGTCCTCTGTCCAGGATCTGG
          :: : : ::::: : : : : : : : ::::
```

有两个基因序列 (AF015262 和 HS229043) 报告，但它们与未知序列的碱基相同率均不到 70%。

重新进行 fasta 选项设置：选 fastx 程序和 SWISS-PROT 数据库，并重新运

行 fasta3 , 得到以下结果 :

```
FASTX (3.39 May 2001) function [optimized, BL50 matrix (15:-5:-1)] ktup: 2
  join: 36, opt: 30, gap-pen: -12/ -2 shift: -20, width: 16
  Scan time: 2.150
The best scores are:
SW:BRC1\_HUMAN\_P38398 BREAST CANCER TYPE 1 SUSC (1863) [f] 149 55 2.6e-07
SW:BRC1\_CANFA\_Q95153 BREAST CANCER TYPE 1 SUSC (1878) [f] 143 53 1e-06
SW:NODL\_RHIME\_P28266 NODULATION PROTEIN L (EC (183) [f] 70 29 2.2

>>SW:BRC1\_HUMAN\_P38398 BREAST CANCER TYPE 1 SUSCEPTIBILI (1863 aa)
  initn: 148 init1: 148 opt: 149 Z-score: 253.3 bits: 55.4 E(): 2.6e-07
Smith-Waterman score: 149; 95.238% identity (95.238% ungapped) in 21 aa overlap

          30          60
EMBOSS SWTEDNGFHAIGQMCEAPVVT
          .....
SW:BRC AWTEDNGFHAIGQMCEAPVVT
          1820          1830
```

得到两个匹配良好的基因序列 (P38398 和 Q95153), 碱基相同率均在 90%以上。应如何确定未知序列和解释以上两次搜索结果 ?

第四节 寡核苷酸设计

有关序列分析的内容非常丰富, 本节只对引物设计进行简单讨论, 而其它一些内容(如 ORF 的查找)在下章中论述。

一. 寡核苷酸设计

聚合酶链式反应(PCR)技术的广泛应用, 刺激了多种辅助设计和用于PCR的寡核苷酸引物程序的兴起。一些程序可通过Internet免费索取, 例如Primer、OSP、PGEN、Amplify等(见附录)。一般而言, 这些程序通过检索已知的重复序列元件, 然后再分析假定引物的长度和GC含量从而优化 T_m 值, 实现PCR引物的辅助设计。

1. 引物设计

许多软件可以根据相应的标准为你的序列设计引物。如果你熟悉 PCR, 你将理解软件中的有关选项; 如果不熟悉, 相关软件中均会有使用手册备查。以下是应用 Primer3 程序(见 EMBnet 挪威站点)进行的一次引物设计, 注意设计结果中一些有用的信息(如 G-C 组成比率、建议的退火温度等) :

例举序列 :

```
GACTGTGGCTGCTGGCGTTGAGGGAAACCTGCCTGTACGTGAGGCCCTAAAAAGCCA
GAGACCTCACTCCCGGGGAGCCAGCATGTCCACTGCGGTCTGGAAAACCCAGGCTT
GGGCAGGAACTCTCTGACTTTGGACAGGAAACAAGCTATATTGAAGACAACTGCAA
TCAAATGGTGCCATATCACTGATCTTCTCACTCAAAGAAGAAGTTGGTGCATTGGC
CAAAGTATTGCGCTTATTTGAGGAGAATGATGTAAACCTGACCCACATTGAATCTAG
ACCTTCTCGTTTAAAGAAAGATGAGTATGAATTTTTACCCATTTGGATAAACGTAG
CCTGCCTGCTCTGACAAACATCATCAAGATCTTGAGGCATGACATTTGGTGCCTGT
CCATGAGCTTTCACGAGATAA
```

结果 :

```

No mispriming library specified
Using 1-based sequence positions
OLIGO      start  len   tm     gc%   any   3' seq
LEFT PRIMER  112   20   59.98  55.00  3.00  3.00  CTTGGGCAGGAACTCTCTG
RIGHT PRIMER 364   20   59.99  50.00  3.00  3.00  GATGTTTGTTCAGAGCAGGCA
SEQUENCE SIZE: 420
INCLUDED REGION SIZE: 420

PRODUCT SIZE: 253, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00

  1  GACTGTGGCTGCTGGCGTTGAGGGAAACCTGCCTGTACGTGAGGCCCTAAAAAGCCAGAG

 61  ACCTCACTCCCGGGGAGCCAGCATGTCCACTGCGGTCTCTGGAAAACCCAGGCTTGGGCAG
      >>>>>>>>

121  GAAACTCTCTGACTTTGGACAGGAAACAAGCTATATTGAAGACAACCTGCAATCAAAATGG
      >>>>>>>>>>

181  TGCCATATCACTGATCTTCTCACTCAAAGAAGAAGTTGGTG-CATTGGCCAAAATATTGCG

241  CTTATTTGAGGAGAATGATGTA AACCTGACCCACATTGAATCTAGACCTTCTCGTTTAAA

301  GAAAGATGAGTATGAATTTTTTCACCATTGGATAAACGTAGCCTG-CCTGCTCTGACAAA
      <<<<<<<<<<<<<<<<<<<<<<

361  CATCATCAAGATCTTGAGGCATGACATTGCTGCCACTGTCCATGAGCTTTTACGAGATAA
      <<<<<

KEYS (in order of precedence):
>>>>> left primer
<<<<<< right primer

```

用于测序或 PCR 的引物，需要选定可特异识别靶区的适当序列，然后检查该序列，以杜绝寡核苷酸形成稳定二级结构的可能。序列中的反向重复查找可通过找寻重复序列或 RNA 折叠的程序来进行。如果查出可能的茎区结构，引物序列可以向前或向后移动几个核苷酸，以期尽量削弱所预测形成的二级结构。寡核苷酸序列还应与适当的载体及插入 DNA 两条链上的序列进行比较。显而易见，测序引物应仅与靶 DNA 的一个区段相配对。若引物与靶 DNA 序列的非目标区很相似，即使只有一个位置不完全配对，这种情况一般也要避免。对于用于扩增基因组 DNA 的 PCR 引物，引物序列应与 GenBank 序列库中的序列进行比较，以检查是否有显著相似的配对区，如果寡核苷酸序列出现在任何已知 DNA 序列中，或者有更严重的情况，也就是寡核苷酸序列出现在任何已知的重复序列元件中，那么引物序列就必须改变。

2. 用于检测相关基因的简并探针

一旦找出保守的蛋白质序列，并可以设计简并寡核苷酸作为杂交探针来筛选文库，找出蛋白质家族中的其他成员。设计这一用途的寡核苷酸，必须先将保守的蛋白质序列翻译成简并 DNA 序列。大多数软件包都可提供这一功能，其输出结果是采用 IUPCC 简并核苷酸代码表示的 DNA 序列。随后就可以合成对应于这一翻译蛋白质序列的简并寡核苷酸。多数 DNA 合成仪都可以合成除了末端核苷酸以外的序列内部任何位置上带有一种以上核苷酸的寡核苷酸。