第 1-3 章 两条序列联配算法及序列搜索

序列分析是生物信息学最主要的研究内容之一,它可以分为两个主要部分:一是序列组成分析(包括基因和基因组层次),二是序列之间的比较分析。两条序列或多条序列间的联配或比对,目的是对它们的序列相似性进行评估,找出这些序列中结构或功能相似性区域等。通过联配未知序列与已知序列(其功能或结构等已知)的相似程度,我们可以判断或推测未知序列的结构与功能。

第一节 序列联配基本概念

序列联配(sequence alignment)也叫序列对比,是生物信息学中的重要内容之一,许多生物信息学分析均涉及序列联配方法。如下两条 DNA 序列:

```
>seq1
TGCGGAGC
>seq2
TCGGAGC
```

我们简单地把它们联配如下,仅有两个碱基匹配:

```
TGCGGAGC
|    |
TCGGAGC
```

如果我们在一条序列中引入一个空位或空格(gap),即一条 DNA 序列在进化过程中经常发生的碱基删除事件,它们的联配结果会显著改善:

```
TGCGGAGC
|  ||||||
T-CGGAGC
```

序列联配我们可以定义如下:根据特定的计分规则,通过一定的算法对两条或多条 DNA 或蛋白质序列进行比较,找出它们之间最优匹配或最大相似度匹配。

序列联配的意义在于,通过序列比对可以获得一个序列相似性度比对值,通过这个值的统计学特征分析(见本章第四小节),可以获得两条序列相似度或同源性的一个统计学判断。如果达到显著水平,说明两条序列相似,进化上有亲缘关系。在序列分析中,大量问题涉及这样的相似度判断,如基因家族、系统发育等分析。

根据序列联配的目的不同,序列联配可以分为全局联配和局部联配两种方式。全局联配(global alignment)目的是对两条序列的全长进行比对,目标是基于它们的全长序列获得最优匹配结果。同样上例两条序列,以一定计分规则(如碱基匹配得 1 分,错配罚 1 分,一个空位罚 3 分),它们的全局最优联配结果如下:

TGCGGAGC
|　|||||
T-CGGAGC　　(7-3=4 分)

局部联配(local alignment)的目的是获得两条序列比对中得分最高的匹配片段。上例的局部最优联配结果为:

CGGAGC
||||||
CGGAGC　　(6 分)

上述案例的两条序列很短,略加比较就可以获得最佳匹配结果。而实际应用中,序列往往很长,可能的联配方式很多,这样就需要一个算法找到两条序列的最佳匹配结果(详见第三节)。

由此可见,进行序列匹配,需要一个计分规则/系统(生物信息学专业名称叫计分矩阵)和一个确定最优联配的算法,同时,还需要一个统计方法确定序列间的相似程度。以下各节就分别对计分矩阵、联配算法和统计方法进行讲解。

第二节　计分矩阵

计分矩阵(scoring matrix)是序列联配过程中使用的计分规则,是序列比对的重要组成部分,它给出序列联配中碱基或氨基酸匹配或错配值,故又称替换矩阵 (substitution matrix)。DNA 序列相对比较简单,只有 4 种碱基,而蛋白质序列有 20 种氨基酸,如何给出这些氨基酸匹配和错配的一个科学准确评价值,即准确反映它们的生物学特征,是生物信息学发展之初就面临的问题,也是最早被解决的序列联配关键问题。

一、计分矩阵的一般原理

构建计分矩阵,我们需要找到一个可以估计任何联配的某一统计数,使生物学关系最显著的联配统计数最大。

先看以下 2 条氨基酸序列的联配情况。如果我们将各残基按相同率处理,则 2 种联配方式(a 和 b)的得分是相等的(9 个残基中 5 个匹配):

(a) TTYGAPPWCS
　　TGYAPPPWS
　　*　***　*

(b) TTYGAPPWCS
　　TGYAPPPWS
　　*　*　***

但是联配(a)中是一些相对常见的残基(A、P、S 和 T)保持一致,而联配(b)则是有一些相对稀有残基(W-色氨酸和 Y-酪氨酸)相一致。我们需要一个更科学的赋分方法来反映匹配氨基酸间生物学和化学关系。实际联配中,C-C 匹配相对比 S-S 匹配更重要些,因为半胱氨酸(C)是具有非常特殊性质的相对稀有氨基酸,而丝氨酸(S)则相对常见或普通。同样,D-E错配值应取正值,因为这两个残基具有相同的化学性质,在两条联配的蛋白质序列中能起到相同的功用。但是,V-K 匹配结果则应被罚分,因为这两个残基毫无相似,不可能在两条序列中起到一样的作用。

用于 DNA 序列联配的替换矩阵相对比较直观。以下是一个常被使用的 DNA 替换

矩阵：

	A	C	G	T
A	0.9	−0.1	−0.1	−0.1
C	−0.1	0.9	−0.1	−0.1
G	−0.1	−0.1	0.9	−0.1
T	−0.1	−0.1	−0.1	0.9

矩阵中每个匹配的碱基对均计为0.9分，每个不匹配（错配）的碱基对被罚0.1分，这样，下面一个联配的得分应为5×0.9+2×(−0.1)=4.3：

```
GCGCCTC
GCGGGTC
***  **
```

用于蛋白质联配的替换矩阵相对复杂，因为没有一个矩阵可以适用各种情况。构建矩阵时，应考虑不同的蛋白质家族在进化过程中一种氨基酸突变成另一种氨基酸概率的差异，根据不同的蛋白质家族和预期的相似程度构建不同的替换矩阵。两个最著名的氨基酸替换矩阵分别是PAM和BLOSUM，它们分别是在1979年和1992年被提出的。

一个重要的概念必须明确。同源性（homology）和相似性（similarity）是不同的2个概念，不能混淆和混用。2条序列具有同源性，意味着这两条序列存在进化方面的关系，它们从一条共同的祖先序列进化而来；而相似性，只是表明两条序列间具有一定的相似程度。

序列联配计分中另一个重要问题是空位问题。空位处理是针对序列进化过程中可能发生的插入和缺失而设计的。插入和缺失可能只涉及1个或多个碱基或残基，也可能是整个功能域（domain），所以，在进行空位罚值设计时必须反映这些情况。

一般用两个参数应用于空位罚值（gap penalties）设定，一个与空位设置（gap opening）有关，另一个与空位扩展（gap extension）有关。任一空位的出现均处以空位设置罚值，而任一空位的扩大则处于空位扩展罚值。对于一个空位长度为 k 的罚值 w_K 可用下式表示：

$$w_k = a + bk$$

其中 a 是空位设置罚值，b 为空位扩展罚值。这两个参数值设置的变化会对联配产生明显影响（表1-3.1）。

表1-3.1 空位设置和空位扩展罚值对联配的影响

空位设置罚值（a）	空位扩展罚值（b）	联配效果
大	大	极少插入或缺失，适用于非常相关蛋白质间的联配；
大	小	少量大块插入，用于整个功能域可能插入的情况
小	大	大量小块插入，适用于亲缘关系较远的蛋白质同源性分析

如何设定罚值并无明确的理论可循，大的空位设置罚值配以很小的空位扩展罚值，被普遍证实是最佳的设定思路。经过多年的试验，一个合适的空位罚值方式已经被确定下来，大多数联配程序均对特定的替换矩阵设定了空位罚值的缺略值（default）。如果使用者希望使用不同的替换矩阵，则可以根据特殊问题设置合适的空位罚值标准。

二、氨基酸替换矩阵

1. PAM 替换矩阵

已故 Dayhoff 是蛋白质列序比较的先驱,她和她的同事们通过对蛋白质进化模式的研究,建立了一组被广泛应用的氨基酸替换矩阵,这些矩阵常被称为 Dayhoff 矩阵、MDM(Mutation Data Matrix)或 PAM(Percent Accepted Mutation)矩阵。

由于蛋白质最有可能是自然选择的目标,可以认为蛋白质序列的分析比 DNA 分析更具有生物学意义。蛋白质分析完全避免了几个三联体可能编码同一氨基酸的遗传密码简并问题。各种氨基酸间特性不一样,在进化过程中一种氨基酸被另一种氨基酸替换的概率大小也不一样。例如氨基酸可分成中性疏水(G、A、V、L、I、F、P、M)、中性亲水(S、T、Y、W、N、E、C)、碱性(K、R、H)和酸性(D、E)氨基酸等。在比较许多具有相似特性蛋白质序列的基础上,Dayhoff 等(1979)确定了进化过程中一种氨基酸被另一种氨基酸替换的经验数据。他们收集了大量蛋白质家族序列,通过比较,她们共观测到 1572 次替换"事件"。以此为基础,她们建立了表 1-3.2 的"可观测或可接受点突变矩阵 A"(accepted point mutation matrix)(由于舍入误差使表中的数值相加不完全等于 1 572)。氨基酸 i 被氨基酸 j 替换的经验次数(记作 A_{ij})可从上表中找到。矩阵 A 可被称为原始 PAM 矩阵。

表 1-3.2　氨基酸替换次数表(Dayhoff 等,1979)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
R	30																		
N	109	17																	
D	154	0	532																
C	33	10	0	0															
Q	93	120	50	76	0														
E	266	0	94	831	0	422													
G	579	10	156	162	10	30	112												
H	21	103	226	43	10	243	23	10											
I	66	30	36	13	17	8	35	0	3										
L	95	17	37	0	0	75	15	17	40	253									
K	57	477	322	85	0	147	104	60	23	43	39								
M	29	17	0	0	0	20	7	7	0	57	207	90							
F	20	7	7	0	0	0	0	17	20	90	167	0	17						
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7					
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269				
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696			
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0		
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17

注:总计观测到 1572 次替换;表中次数均已乘 10;祖先序列不明时,次数以平分处理

由矩阵 A 可以进一步获得“突变概率矩阵 M”(mutation probability matrix)。矩阵 M 的元素 M_{ij}表示经过一定的进化时期氨基酸 j 被氨基酸 i 所替换的经验频率。Dayhoff 等进而把可观测突变百分率(percent accepted mutation 或 point accepted mutation per 100 residues),即 PAM 作为一种时间度量单位。假设同一位点不会发生二次以上的突变,则 1PAM 等于 100 个氨基酸多肽链中预期发生一次替换所需的时间。Schwarts 和 Dayhoff(1979)发现将突变概率矩阵 M 250 次方处理获得的 250PAM 矩阵,对于研究远缘蛋白质之间进化关系是一个合适的时间单位。Dayhoff 等(1979)进一步定义了一个相对概率矩阵 R(relatedness odds matrix),表 1-3.3 中各元素已经对数处理,为对数概率矩阵(log-odds matrix),并将最有可能发生相互替换的氨基酸归类排列。

表 1-3.3 PAM250 的对数概率矩阵(Dayhoff 等,1979)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

*表中数值均乘以 10

1PAM 相当于所有的氨基酸平均有 1%发生了变化,经过 100PAM 的进化,并非每个氨基酸的残基均发生变化:有一些可能突变多次,甚至又变成原来的氨基酸,而另一些氨基酸可能根本没有发生过变化。总体上说,利用大于 100PAM 的时间间隔可能区分进化关系较远的同源性蛋白质。应该注意,PAM 与进化时间之间没有大致对应关系,因为不同的蛋白质家族的进化速率是不同的。当 2 条序列进行相似性比较时,事先不知道怎样的进化时间

(PAM)是恰当的。对于相近的序列,比较容易选择,即使不太合适的矩阵也无妨。在很多年里,PAM250(矩阵后面数字表示一种进化的距离,数字越大,进化距离越远)是应用最广的替换矩阵,因为该矩阵是唯一由 Dayhoff 最初发表的矩阵。后来一些学者利用大量新出现的蛋白质序列数据更新 Dayhoff 矩阵的频率数值,由此构建新的 PAM 矩阵,但它们与最初的 PAM 矩阵没有太大的差异。

2. BLOSUM 替换矩阵

另外一种构建矩阵的方法是由 Henikoff 等于 1992 年提出的,建成的矩阵称为 BLOSUM(blocks substitution matrix)。他们直接利用多序列联配(multiple alignment)分析亲缘关系较远的蛋白质,而不是用近缘的序列。这方法的优点是符合实际观测结果,不足之处是它不能和进化挂起钩来。大量的试验表明,BLOSUM 矩阵总体比 PAM 矩阵更适合于生物学关系的分析和局部相似性搜索。

假设 f_{ij} 为序列联配中氨基酸 i 和 j 对(忽略顺序),则 i,j 对氨基酸所占比例为:

$$q_{ij}=f_{ij}/\sum_{i,j} f_{ij}$$

在完全独立的状况下,该比例的期望值为:

$$e_{ij}=\begin{cases}p_i^2 & i=j\\ 2p_ip_j & i\neq j\end{cases}$$

$$p_i=q_{ii}+\frac{1}{2}\sum_{j\neq i}q_{ij}$$

则 BLOSUM 矩阵元素(i,j)定义为:

$$s_{ij}=2\log_2(q_{ij}/e_{ij})$$

蛋白序列的高度保守区(highly conserved regions)或称为模块(block)数据被用于构建 BLOSUM 矩阵。BLOSUM 矩阵后的数字表示用于构建矩阵模块的最小相似比例,例如 BLOSUM62 为用于构建矩阵的模块序列数据中,序列片段的各联配点上至少 62%是相同的。矩阵后的数字越大,则表示关系越近。

三、位置特异性计分矩阵(PSSM)

PSSM(position-specific scoring matrix)是由一个简单对数变换而来的矩阵,它给出不同来源的一小段保守序列(基序)各个特定位置氨基酸的频率。PSSM 可以用于一条序列的保守序列的搜索。一条序列中,与 PSSM 最相似的位置即为 PSSM 代表的基序位置。PSSM 在数据库搜索,特别是保守短序列(功能域)搜索方面有很好的应用。例如 PSI-BLAST,DELTA-BLAST 等均使用 PSSM 进行搜索等(详见第四节)。

下面举例说明 PSSM 构建过程:

例如如下一个保守序列联配结果:由 5 条序列,每条 6 个碱基构成,即从左到右 6 个位置上均有 5 个碱基构成。

AGGCTT
AAGCTA
AAACTT
TAACTA
AGACTT

我们如何将其转换成一个 PSSM?

步骤 1:联配 6 列位置上四种碱基数量统计;合计(即背景)统计四种碱基

	#1	#2	#3	#4	#5	#6	合计
A	4	3	3	0	0	2	12
G	0	2	2	0	0	0	4
C	0	0	0	5	0	0	5
T	1	0	0	0	5	3	9

步骤 2:各个位置四种碱基频率计算;合计四种碱基频率计算

	#1	#2	#3	#4	#5	#6	背景频率
A	0.8	0.6	0.6	0	0	0.4	0.40
G	0	0.4	0.4	0	0	0	0.13
C	0	0	0	1.0	0	0	0.17
T	0.2	0	0	0	1.0	0.6	0.30

步骤 3:标准化(去除背景碱基的影响),即各个位置碱基频率除以背景相应碱基频率

	#1	#2	#3	#4	#5	#6	背景频率
A	2.0	1.5	1.5	0	0	1.0	0.40
G	0	0.3	0.3	0	0	0	0.13
C	0	0	0	5.9	0	0	0.17
T	0.7	0	0	0	3.3	2.0	0.30

步骤 4:进行以 2 为底对数 log2 数据转换,完成 PSSM 构建

	#1	#2	#3	#4	#5	#6
A	1.0	0.6	0.6	−3.3	−3.3	1.0
G	−3.3	−1.7	−1.7	−3.3	−3.3	−3.3
C	−3.3	−3.3	−3.3	2.6	−3.3	−3.3
T	0.7	−3.3	−3.3	−3.3	1.7	1.0

PSSM 可以直接用于序列搜索,确定未知序列中与 PSSM 序列相似的保守区段。对于一条未知序列,可以用 PSSM 对该序列进行扫描(即从左到右以 PSSM 宽度为窗口宽度,逐个碱基位置进行比对),这样就可以获得未知序列中各个位置片段与 PSSM 相似性概率。例如某一未知序列:

未知序列:……A G A C T A ……

	#1	#2	#3	#4	#5	#6
A	1.0	0.6	0.6	−3.3	−3.3	1.0
G	−3.3	−1.7	−1.7	−3.3	−3.3	−3.3
C	−3.3	−3.3	−3.3	2.6	−3.3	−3.3
T	0.7	−3.3	−3.3	−3.3	1.7	1.0

利用上述 PSSM(6 个碱基窗口)对该条未知序列进行逐个碱基位点扫描。在上述未知序列的某一个特定位点,基于该 PSSM 的几率对数值(log odds score)为(1.0−1.7+0.6+2.6+1.7+1.0)= 5.2。

“5.2”意味着 37/1 比率支持未知序列中该 6 个碱基连续片段与 PSSM 并非随机匹配。也就是说,该未知序列的确包含与该保守区域高度同源的序列片段。

以上有关替换矩阵的讨论仅仅提及蛋白质序列的比较,相关的原则同样适用于 DNA 序

列的比较。如前所述,DNA 替换矩阵非常简单,所有四个碱基的匹配与不匹配的替换数值均设为相同,不同的只有匹配与否(匹配 0.9 分和错配罚分 0.1)。一个较复杂的模型可以把转换(两种嘧啶或两种嘌呤间的突变)频率设为高于颠换(嘧啶与嘌呤间的突变)频率。

第二节　两条序列联配算法

一、Needleman-Wunsch 算法

Needleman-Wunsch 算法是一种全局联配算法,它从整体上分析两个序列的关系,即考虑序列总长的整体比较,用类似于使整体相似 (global similarity)最大化的方式,对序列进行联配。两个不等长度序列的联配分析必需考虑在一个序列中一些碱基的删除即在另一序列做空位(gap)处理。Needleman 和 Wunsch(1970)最初提出的算法是寻求使两条序列间的距离最小,或最短距离,它使用的是一个动态规划(dynamic programming)的方法。该算法可以用于核酸和蛋白质序列,是生物信息学最经典算法之一。在给定计分规则(替换矩阵和空位罚值)情况下,它们总是能给出具有最高(优)联配值的联配结果。但是,这个联配结果并不一定具有生物学意义,因为它可能达不到生物学意义上的显著水平。

如果将两条联配的序列沿双向表的上轴和左侧轴放置,两条序列的所有可能的联配方式都将在它们所形成的方形图中(例举见图 1-3.1a)。图中标出了一条序列所有碱基与另外一条序列碱基所有可能的联配方式(碱基匹配、不匹配和删除即空位),这样所有可能的联配方式都在这个方形图中。从最上角出发,到右下角结束,任何一个联配方式均可以画出一条联配路径,或反过来,任何一条路径也对应一种联配方式(如图中标注的一个路径及其对应的联配结果)。这样一来,我们确定任意 2 条序列最优联配结果的问题就转化为寻找最优路径问题了。所有可能的路径(联配方式)如果都在这个方形图中,那么我们如何找到最短路径?

对于任一联配位点,即图中的任一单元格,仅有三种可能的方式延伸联配过来(图 1-3.1b):(1)碱基匹配或不匹配,即每一序列均加上一个碱基(x 路径),并给其增加一个规定的距离权重(匹配加分,错配罚分);(2)在一个序列中增加一个碱基而在另一序列中增加一个空位或反之亦然(y 和 z 路径)。这三种延伸方式的权重值分别加上到达上一个位点的累计得分(x,y,z),就可以得到三种可能联配方式的得分,然后得分最高(H)的路径作为到达本位点的最佳路径。引入一个空位时也将增加一个规定的距离权重(空位罚分)。因此,图中的一个单元可以从(最多)三个相邻的单元达到。为了获得最优路线,我们必须保证从一开始就每步最优,即把到达单元格距离最小的方向作为序列延伸的方向。将这些方向记录下来,并在计算了所有的单元之后,沿着记录的方向就有一条路径可从方形图右最下角(两个序列的末端)追踪到左最上角 (两个序列的起点)。由此所产生的路径将给出具有最短距离的序列联配(即最优连配结果)。如两条路径获得等距离(相同得分),意味着存在两条最短路径。

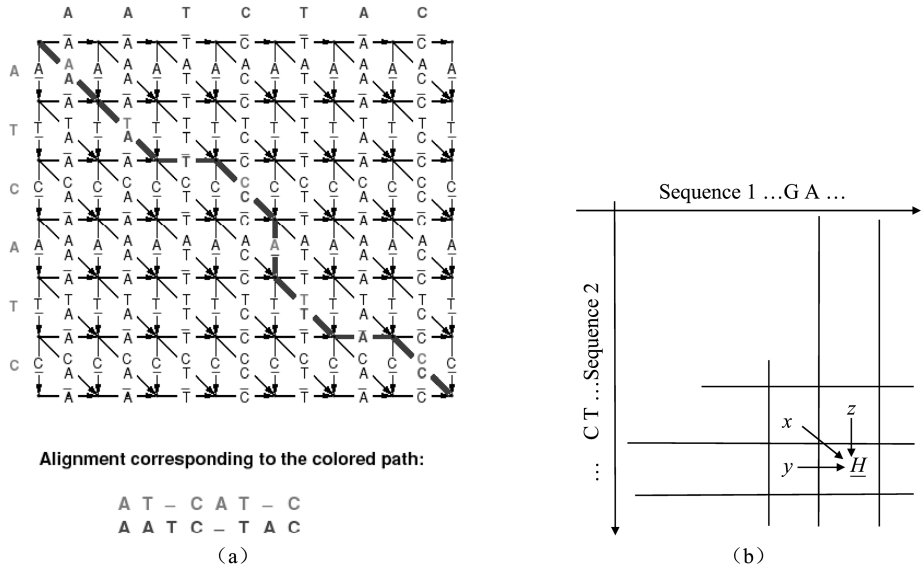


图 1-3.1 两条序列联配方式与路径

以两个短序列为例,将上述过程说明如下:

>seq1

CTGTATC

>seq2

CTATAATCCC

设定计分规则为:碱基匹配不罚分(0),碱基错配时距离权重(罚分)为-1,引入一个空位时距离权重为-3

两条序列分列于图两侧(图 1-3.2):初始值设为 0,然后依次从左上角向右下角计算最佳路径。根据图 1-3.1b,每个联配位点均可以计算其 x,y,z,H 值:例如本例中,第一位点(浅色框)为:$(x,y,z,H)=(0,-3,-3,0)$,意思是进行碱基联配路径(x)得 0 分(不罚分),而在其中任何一条序列中加入一个空位(y,z)均罚 3 分,最后到达这个位点的最高得分是进行碱基联配(即最佳路径),H 得 0 分(初始位点值=0+碱基匹配=0),而其他两个路径分别得分-6(上一个位点得分=-3+该路径得分=-3)。再看一个位点(深色框):$(x,y,z,H)=(-1,-3,-3,-3)$,其最佳路径是在序列 1 中插入一个空位(z),其得分为-3(上一个位点=0+(-3)),而序列 2 中插入空位的路径(y)达本位点最后得分-9(上一个位点=-6+(-3)),碱基联配路径(x)得分-4(上一个位点=-3+(-1)),后两个路径值均不及第一个 z 路径,因此 $H=-3$。

联配过程中可以用箭头标出最佳路径。如此依次类推,就可以最终确定这两条序列最后一个联配位点的得分。反推得到这个得分的路径,就可以得到其联配方式,而该联配方式就是这两条序列的最佳联配结果。

本例中,沿箭头所指方向在表中从右最下角向左最上角追踪,可以发现 6 条路径可以从初始位点抵达最后一个位点,说明存在 6 种最佳联配方式:

CTATAATCCC　　CTATAATCCC　　CTATAATCCC
CTGTA-TC--　　CTGTA-T-C-　　CTGTA-T--C

CTATAATCCC　　CTATAATCCC　　CTATAATCCC
CTGT-ATC--　　CTGT-AT-C-　　CTGT-AT--C

在上述 6 种联配中,罚分均为-10,即在较短序列中有 6 个匹配碱基、1 个错配碱基和 3 个空位。

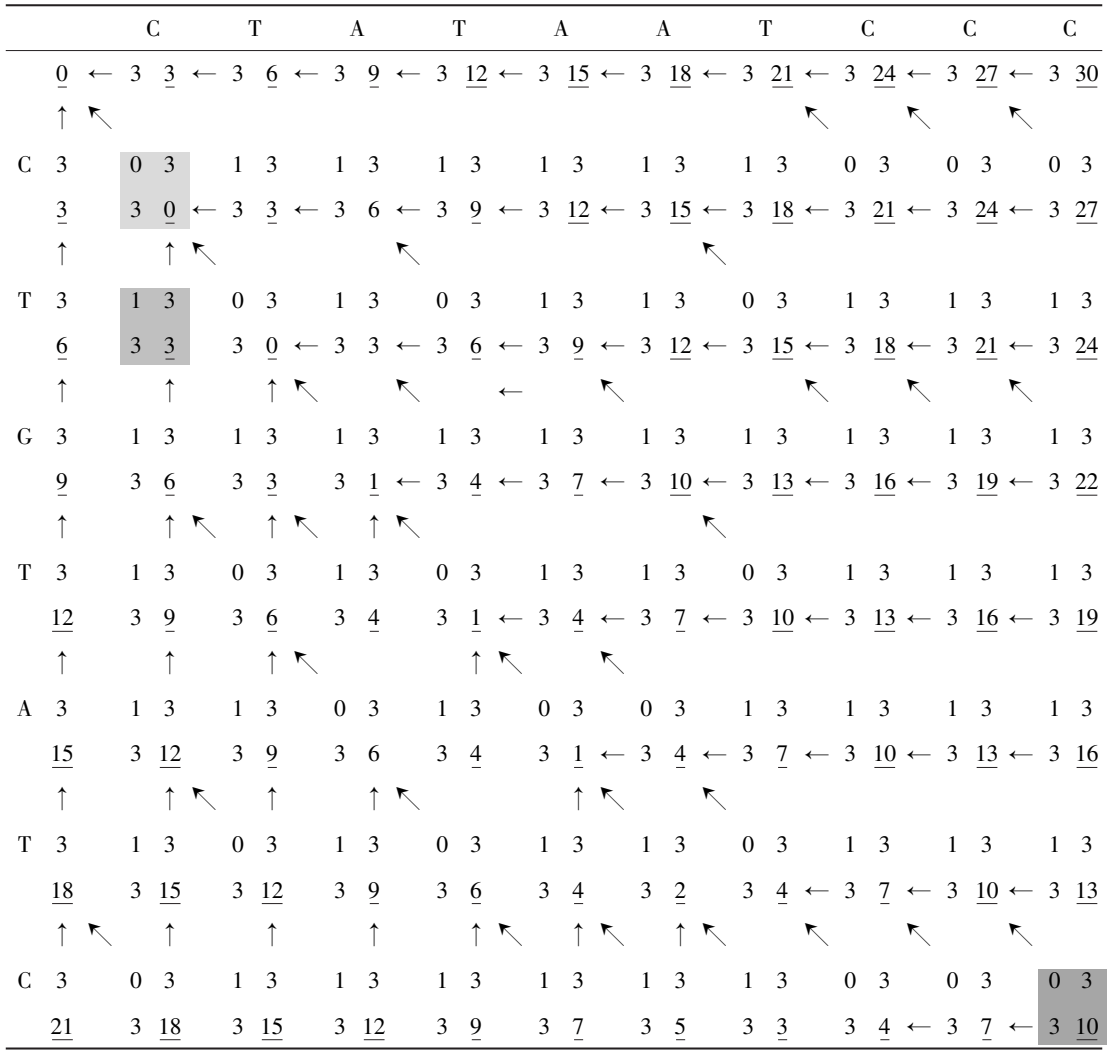


图 1-3.2　Needleman-Wusch 算法实例

计分方式:设定碱基错配罚 1 分,单个碱基缺失或插入时罚 3 分,碱基匹配不罚分即得 0 分。图中箭头为每个位点最佳联配方向(来源)。每个单元格(3 个特别标出举例)内有四个数字分别为 x,y,z,H 值(H 值下划线)(图 1-3.1)。两条序列的全局最优联配 H 值=10(即罚 10 分),其联配路径见反向箭头。

该算法可以用代数形式来描述。设具有碱基 a_i 和 b_j 的两个序列 a 和 b,这两个序列间距离为 $d(a,b)$。通过评价序列 a 中前 i 个位置和序列 b 前 j 位置的距离 $d(a^i,b^j)$,递归地得到距离 $d(a,b)$。如果 a 和 b 的长度为 m 和 n,则其期望距离为 $d(a^m,b^n)$。上表中引入的第

1 行 1 列单元的距离为 0(相当于空序列或初始位点),在单元(i,j)内,使到达该单元距离增加的三种可能事件为:

1. 从单元(i-1,j)向(i,j)的垂直移动,相当于在 b 序列中插入一个空位使相似序列延伸。换言之,b 序列由 a 序列中 a_i 的缺失所产生,这一事件的权重记作 $w\_(a_i)$。

2. 从单元(i-1,j-1)向(i,j)的对角线移动,相当于增加碱基 a_i 和 b_j 使相似序列延伸。换言之,b 序列由 a 序列中的 a_i 被 b_j 取代所产生,这一事件的权重记为 $w\_(a_i,b_j)$。

3. 从单元(i,j-1)向(i,j)的水平移动,相当于在序列 b 中插入一个空位使相似序列延伸。换言之,b 序列由 b_j 插入 a 序列所产生,这一事件的权重记为 $w_+(b_j)$。

因此,单元(i,j)的距离 $d(a^i,b^j)$ 可看成三个相邻单元的距离加上相应权重后的最小者,即

$$d(a^i,b^j)=\min\begin{cases}d(a^{i-1},b^j)+w\_(a_i)\\ d(a^{i-1},b^{j-1})+w(a_i,b_j)\\ d(a^i,b^{j-1})+w_+(b_i)\end{cases}$$

且初始条件为

$$d(a^0,b^0)=0$$

$$d(a^0,b^j)=\sum_{k=1}^{j}w_+(b_k)$$

$$d(a^i,b^0)=\sum_{k=1}^{i}w\_(a_k)$$

在图 1-3.2 的实例中

$$w\_(a_i)=-3 \qquad (\text{对于每一个 } i)$$

$$w(a_i,b_j)=\begin{cases}0 & (i=j)\\ -1 & (i\neq j)\end{cases}$$

$$w_+(b_j)=-3 \qquad (\text{对于每一个 } j)$$

当联配两个序列时,通过计算其随机重排序列的联配距离(如 1 000 次),可以得到这两个序列间的最小距离估计(距离分布)。如果实际得到的联配距离小于 95%的重排序列距离(1 000 个距离值),则表明实际获得的联配距离达到了 5%的显著水平,是不可能由机误造成的。

二、Smith-Waterman 算法

Smith-Wateman 算法是在 Needleman-Wunsch 算法基础上发展而来的,它是一种局部联配算法。由于亲缘关系较远的蛋白质序列可能只有一些相互独立的保守片段,所以进行局部相似性分析有时可能比整体相似性分析更合理。Smith 和 Waterman(1981)提出了一种查找具有最高相似性片段的算法。对于序列 A=(a_1,a_2,…,a_m)和 B=(b_1,b_2,…,b_n),H_{ij}被定义为以 a_i 和 b_j 碱基对结束的片段(亚序列)的相似性值。与 Needle-Wunsch 算法一样,Smith-Waterman 算法也要利用递推关系来确定 H 值,H 的初始值为:

$$H_{i0}=0,\quad 0\leqslant i\leqslant n,\quad H_{0j}=0,\quad 0\leqslant j\leqslant m$$

相似性计算中包括 2 个统计量:碱基对(或氨基酸等序列因子对) a_i,b_j 的相似性值 $S(a_i,b_j)$和空位权重(罚分)$w_k=v+uk$(k 为空位长度)。Smith-Waterman 算法可以给出两条序

列的最大相似性值。以 a_i,b_j 碱基对结束的片段可以由以 a_{i-1} 和 b_{j-1} 结束片段增加碱基(因子)来获得,或者 a_i 可以删除 k 长度的碱基片段,b_j 可删除 l 长度碱基片段。具体算法如下:

$$P_{ij}=\max(H_{i-1,j}-w_1,P_{i-1,j}-u)$$

$$Q_{ij}=\max(H_{i,j-1}-w_1,P_{i,j-1}-u)$$

$$\text{则 } H_{ij}=\max\begin{cases}H_{i-1,j-1}+S(a_i,b_j)\\P_{ij}=\max_{1\leqslant k\leqslant i}(H_{i-k,j}-w_k)\\Q_{ij}=\max_{1\leqslant l\leqslant j}(H_{i,j-l}-w_l)\\0\end{cases},(1\leqslant i\leqslant m,1\leqslant j\leqslant n)$$

其中 $P_{0,0}=P_{0,j}=Q_{0,0}=Q_{i,0}=0$

该算法可以确保具有最大 H_{ij} 值的序列片段是最优联配结果或相似性最好的。从(a_i,b_j)为起点,向后追踪 H_{ij} 矩阵,直到到达 0 值。对于具有最大相似性片段以外部分的差异性,不会影响到该片段的 H 值。

举例说明这一算法。我们同样以 Needleman-Wunsch 算法中的两条短序列为例。将两条序列(CTGTATC 和 CTATAATCCC)排于表 1-3.4 的两侧,相应的 H_{ij},P_{ij} 和 Q_{ij} 值分别列入表中。本例的权重等根据 Smith 和 Waterman(1981)的例子设定为:

$$S(a_i,b_j)=\begin{cases}1 & a_i=b_j\\-1/3 & a_i\neq b_j\end{cases}$$

$$w_k=-(1+k/3)\quad(1\leqslant k)$$

对于 4 个碱基具有相同频率的随机长序列,$S(a_i,b_j)$ 值的平均值为零。w_k 值应至少不小于匹配与不匹配权重的差值。

表 1-3.4 的最大 H_{ij} 为 4.33(8 行与 7 列相交处),说明该位点是局部可以达到最高序列相似度的联配位点,反推其到达路径,就得到一条最优路径(星号 * 表示),其对应具有最大相似性的片段联配方式为:

CTGTA-TC

CTATAATC

表 1-3.4　Smith-Waterman 算法例举

			$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$
				C	T	G	T	A	T	G
$i=0$		H_{ij}	0	0	0	0	0	0	0	0
		P_{ij}	0	0	0	0	0	0	0	0
		Q_{ij}	0	0	0	0	0	0	0	0
$i=1$	C	H_{ij}	0	$\underline{1.00^*}$	0	0	0	0	0	1.00
		P_{ij}	0	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33
		Q_{ij}	0	-0.33	-0.33	-0.67	-1.00	-1.33	-1.33	-1.33
$i=2$	T	H_{ij}	0	0	$\underline{2.00^*}$	0.67	1.00	0	1.00	0
		P_{ij}	0	-0.33	-0.67	-0.67	-0.67	-0.67	-0.67	-0.33
		Q_{ij}	0	-0.33	-0.67	0.67	0.33	0	-0.33	-0.33

续表

			$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$
				C	T	G	T	A	T	G
$i=3$	A	H_{ij}	0	0	0.67	1.67*	0.33	2.00	0.67	0.67
		P_{ij}	0	−0.67	0.67	−0.67	−0.33	−1.00	−0.33	−0.67
		Q_{ij}	0	−0.33	−0.67	−0.67	0.33	0	0.67	0.33
$i=4$	T	H_{ij}	0	0	1.00	0.33	2.67*	1.33	3.00	1.67
		P_{ij}	0	−1.00	0.33	0.33	−0.67	0.67	−0.67	−0.67
		Q_{ij}	0	−0.33	−0.67	−0.33	−0.67	1.33	1.00	1.67
$i=5$	A	H_{ij}	0	0	0	0.67	1.33	3.67*	2.33	2.67
		P_{ij}	0	−1.33	0	0	1.33	0	1.67	0.33
		Q_{ij}	0	−0.33	−0.67	−1.00	−0.67	0.00	2.33	2.00
$i=6$	A	H_{ij}	0	0	0	0	1.00	2.33*	3.33	2.00
		P_{ij}	0	−1.33	−0.33	−0.33	1.00	2.33	1.33	1.33
		Q_{ij}	0	−0.33	−0.67	−1.00	−1.33	−0.33	1.00	2.00
$i=7$	T	H_{ij}	0	0	1.00	0	1.00	2.00	3.33*	3.00
		P_{ij}	0	−1.33	−0.67	−0.67	0.67	2.00	2.00	1.00
		Q_{ij}	0	−0.33	−0.67	−0.33	−0.67	−0.33	0.67	2.00
$i=8$	C	H_{ij}	0	1.00	0	0.67	0.33	1.67	2.00	4.33*
		P_{ij}	0	1.33	−0.33	−1.00	0.33	1.67	2.00	1.67
		Q_{ij}	0	−0.33	−0.33	−0.67	−0.67	1.00	0.33	0.67
i=9	C	H_{ij}	0	1.00	0.67	0	0.33	1.33	1.67	3.00
		P_{ij}	0	−0.33	−0.67	−0.67	0	1.33	1.67	3.00
		Q_{ij}	0	−0.33	−0.33	−0.67	−1.00	−1.00	0	0.33
$i=10$	C	H_{ij}	0	1.00	0.67	0.33	0	1.00	1.33	2.67
		P_{ij}	0	−0.33	−0.67	−1.00	−0.33	1.00	1.33	2.67
		Q_{ij}	0	−0.33	−0.33	−0.67	−1.00	−1.33	−0.33	0

第四节 BLAST 算法及数据库搜索

当你面对大量两条序列间的比对时,运算时间变得非常重要。例如数据库序列搜索就是这样一个问题,未知序列或递交序列(query sequence)需要跟数据库(如 GenBank)中大量已有序列进行比对,确定与递交序列的相似序列,这时就需要在短时间内(如 1 分钟)返回搜索比对结果。直接利用上述两种算法,将需要大量运算时间,难以达到时间要求,必须提出新算法来解决这一问题。Altschul 等人 1990 年提出的用于数据库搜索的 BLAST (Basic Local Alignment Search Tool)算法和 1988 年 Pearson 和 Lipman 提出的 FASTA(Fast All)算法

很好地解决了这一问题。这两个算法(尤其 BLAST)是目前应用最广泛的方法,美国生物信息技术中心(NCBI)和欧洲生物信息学研究所(EBI)等分别应用它们来进行 DNA 和蛋白质序列相似性搜索。这两个算法都是一种基于局部联配搜索工具,给出的是递交序列与数据库序列的最佳局部联配结果。两者算法相似,本节以 BLAST 为例进行讲解。

一、BLAST 算法

BLAST 算法同样是利用动态规划算法,与 Smith-Waterman 算法类似,其不同之处是引入了所谓“字”或“字符串”(Word 或 *K*-tuple,*K*-mer 等)的检索技术。所有序列其实都是由若干字符串组成,例如我们以 3 个碱基长度的字符串为例,下列 DNA 序列包括了 6 个字长为 3 的字符串,其中 2 个字符串(GCG 和 CGG)各出现了 2 次:

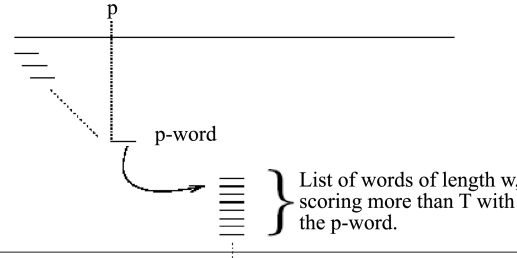
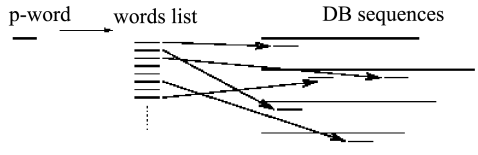
>seq1

TGCGGAGCGG

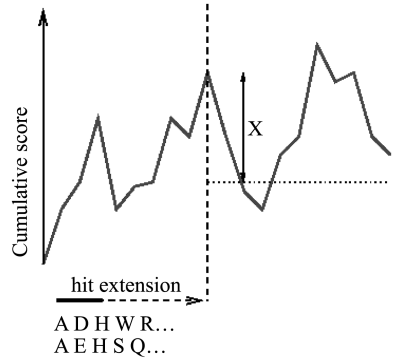
其包含的字符串:TGC,GCG,CGG,GGA,GAG,AGC

为了降低比对时间,BLAST 算法中一个重要手段,是建立序列数据库的“字”检索系统,即将数据库中所有序列所包含的不同长度字符串进行扫描,并建立索引。这样,数据库中存贮的序列包含哪些特定长度字符串就已知了,如常见的 11 个碱基长度 DNA 和 3 个氨基酸长度蛋白质字符串等。当对递交序列进行数据库搜索比对时,首先对递交序列进行扫描,确定其包含的所有特定长度字符串,然后仅对包含相同字符串的数据库序列(一般仅占整个数据库序列很小的比例)进行进一步比对。这样就节省了大量无关序列比对的耗时。对于包含相同字符串的数据库序列,进行进一步序列比对时,基于匹配上的字符串,分别向两端延伸,序列延伸规则基于动态规划算法。

下面用图解方式进行具体说明:

第一步:扫描递交序列的每个位点,发现特定位点(p)起始、字长为 w、匹配值超过 T 的所有字并列表	First step: For each position p of the query,find the list or words of length w scoring more than T when paired with the word starting at p: 
第二步:根据上述字列表,确定目标数据库中所有包含与列表中字完全一致的序列	Second step: For each words list,identify all exact matches with DB sequences: 

续表

第三步:对于每个匹配的字得到的联配(所谓“hit”),向两端以动态规划算法向外延伸。延伸过程中联配值 S 不断变动,当联配过程中联配值 S 降低超过某一临界值 X,延伸结束。这样我们可以获得所谓高得分联配对(HSP, high scoring segment pair)。然后列出所有超过某一设定临界联配值或 E 值的 HSP。E 值是一个统计参数(详见下节),表明随机情况下,可以获得等于或超过联配 S 值的 HSP 数量。该值越小,说明在统计学意义上递交序列与得到的 hit 序列相似性越高。	Third step: For each word match (“hit”), extend ungapped alignment in both directions.Stop when S decreases by more than X from the highest value reached by S.  HSP=High Scoring Segment Pair MSP=Maximal Segment Pair Reports all HSPs having score S above a threshold, or equivalently, having E-value below a threshold. E-value=the number of HSPs having score S (or higher) expected to occur only by chance. Apply sum-statistics to evaluate the significance of a combination of HSPs involving the same DB sequence. 

上述对于每个匹配字得到的联配,以动态规划算法向两端延伸,延伸方式是以一个特定数值(X)为限定,如果超过该值就终止联配。这种动态规划联配方式是一种数值限定类型(score-limited DP)(图 1-3.3)。当然我们也可以用其他条件进行联配限定,如不允许插入空格(ungapped DP),或插入空格数量进行限定(banded DP)。如果没有限定,就是我们熟知的全局联配方式,所谓允许空格的全联配。这些限定 DP 往往针对特殊问题或目的进行选用,以提高搜索速度和高效获得特定靶序列。

二、利用 BLAST 进行数据库序列搜索

采用 BLAST 的基本算法目前形成了若干不同的工具,分别用于特定序列数据库和特定目的的序列搜索。以 NCBI 提供的在线序列数据库搜索工具 BLAST(2014 年 12 月)为例(图 1-3.4),BLASTN 是对核苷酸递交序列库搜索核苷酸序列数据库,BLASTP 是在蛋白质序列库中搜索蛋白质序列,TBLASTN 则可以在核酸序列库中搜索蛋白质序列,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,它要将所递交的核酸序列按所有 6 种读框翻译,然后再用它搜索蛋白质序列库。同时,特定目的或方式搜索工具不断被开发出来(图 1-3.5),例如 Altschul 等人(1997)提出了一个寻找蛋白质家

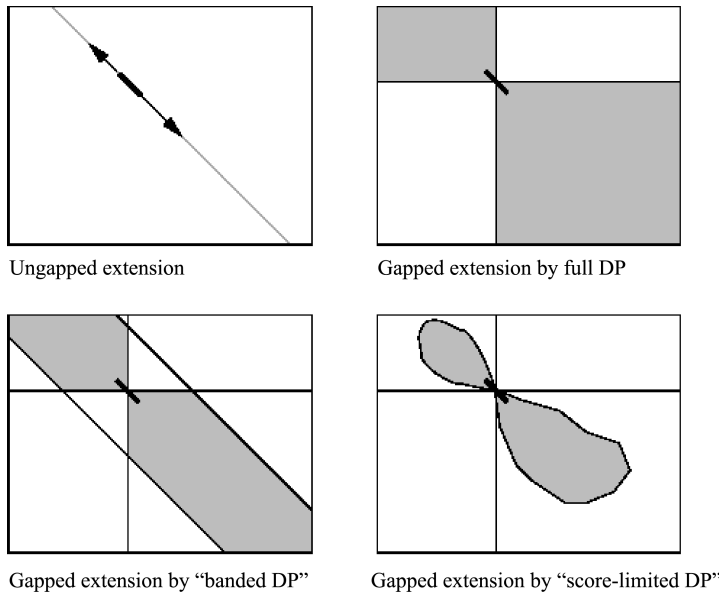


图 1-3.3　序列动态规划联配方式。可以利用空格(是否允许)、空格数量和联配值等进行限定联配过程

族保守序列新算法——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,每一轮都能提高检索的敏感性。

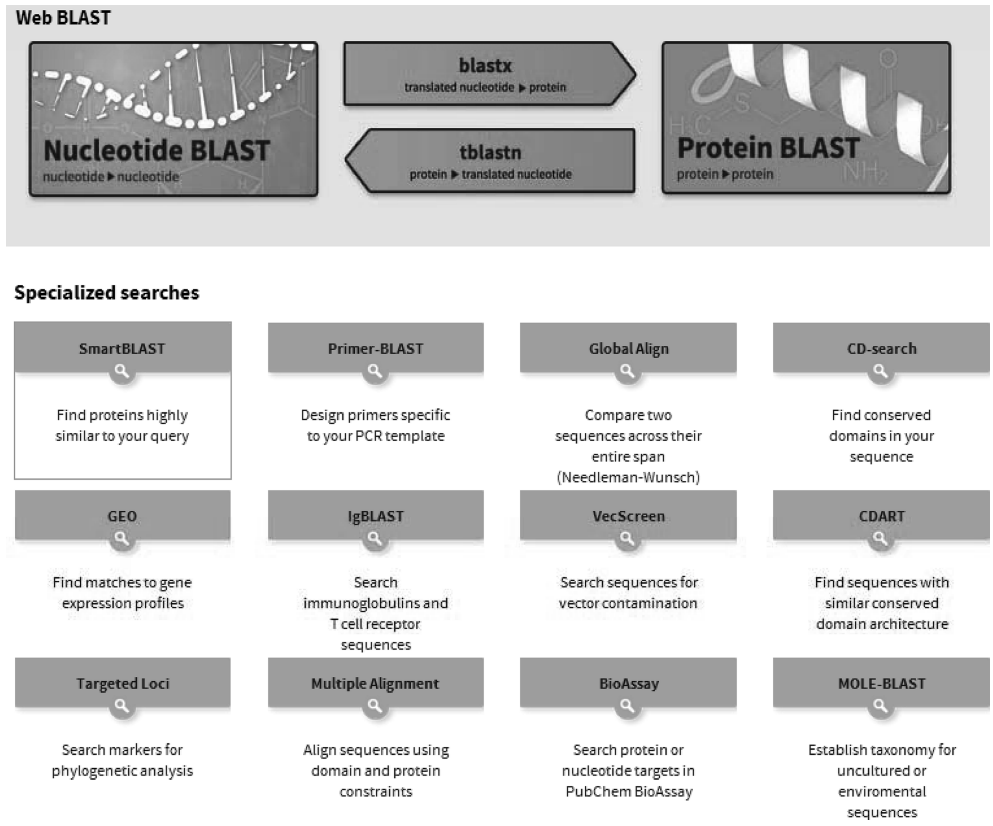


图 1-3.4　美国国家生物信息中心(NCBI)提供的在线 BLAST 序列搜索工具(2014 年 12 月)

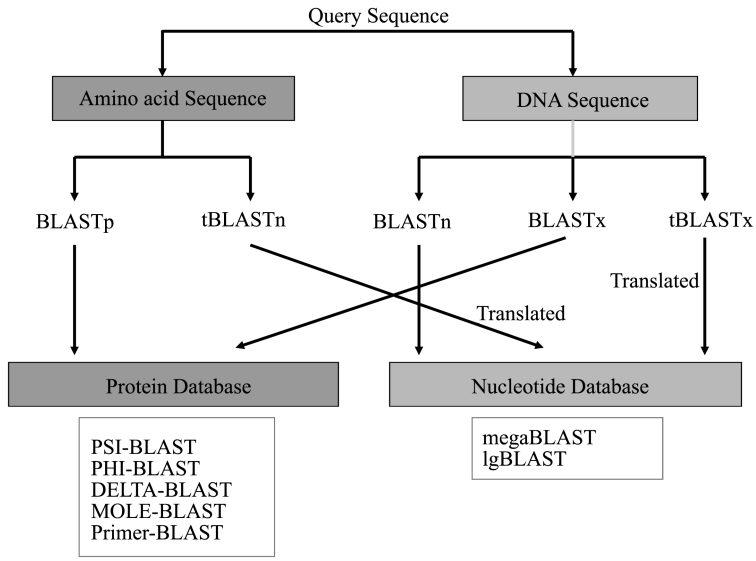
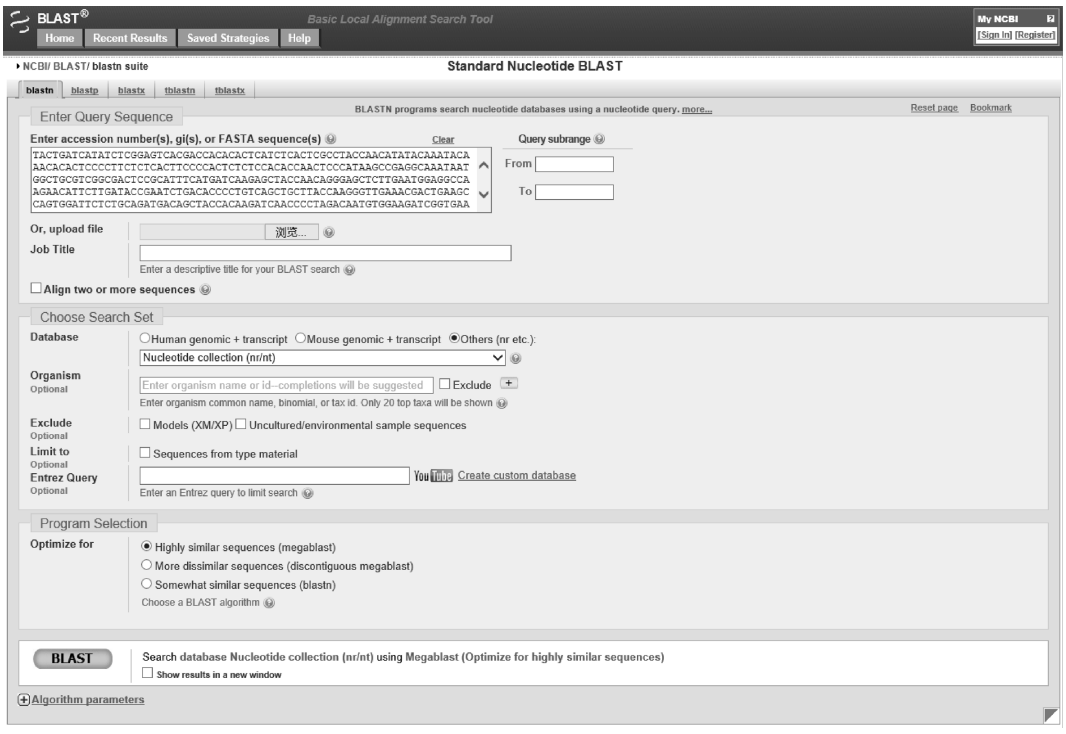


图 1-3.5 数据库 BLAST 搜索工具

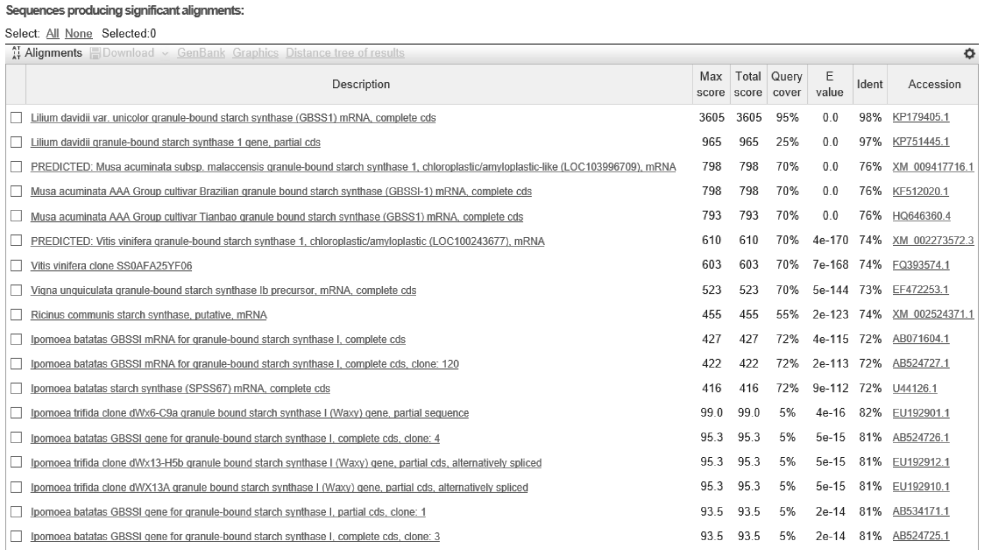
基于局部动态规划算法,除了标准 BLAST 工具(如 BLASTP 等),不断发展了新的用于蛋白质和核苷酸序列数据的 BLAST 工具(下面框中 PSI-BLAST 等)

利用 BALST 工具搜索数据库案例:

对一条未知 DNA 序列,利用 BLASTN 工具搜索核苷酸数据库:



获得如下搜索结果:



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Lilium davidii var. unicolor granule-bound starch synthase (GBSS1) mRNA, complete cds	3605	3605	95%	0.0	98%	KP179405.1
Lilium davidii granule-bound starch synthase 1 gene, partial cds	965	965	25%	0.0	97%	KP751445.1
PREDICTED: Musa acuminata subsp. malaccensis granule-bound starch synthase 1, chloroplastic/amyloplastic-like (LOC103996709), mRNA	798	798	70%	0.0	76%	XM_009417716.1
Musa acuminata AAA Group cultivar Brazilian granule bound starch synthase (GBSSI-1) mRNA, complete cds	798	798	70%	0.0	76%	KF512020.1
Musa acuminata AAA Group cultivar Tianbao granule bound starch synthase (GBSS1) mRNA, complete cds	793	793	70%	0.0	76%	HQ646360.4
PREDICTED: Vitis vinifera granule-bound starch synthase 1, chloroplastic/amyloplastic (LOC100243677), mRNA	610	610	70%	4e-170	74%	XM_002273572.3
Vitis vinifera clone SS0AFA25YF06	603	603	70%	7e-168	74%	FQ393574.1
Vigna unguiculata granule-bound starch synthase Ib precursor, mRNA, complete cds	523	523	70%	5e-144	73%	EF472253.1
Ricinus communis starch synthase, putative, mRNA	455	455	55%	2e-123	74%	XM_002524371.1
Ipomoea batatas GBSSI mRNA for granule-bound starch synthase I, complete cds	427	427	72%	4e-115	72%	AB071604.1
Ipomoea batatas GBSSI mRNA for granule-bound starch synthase I, complete cds, clone: 120	422	422	72%	2e-113	72%	AB524727.1
Ipomoea batatas starch synthase (SPSS67) mRNA, complete cds	416	416	72%	9e-112	72%	U44126.1
Ipomoea trifida clone dWx6-C9a granule bound starch synthase I (Waxy) gene, partial sequence	99.0	99.0	5%	4e-16	82%	EU192901.1
Ipomoea batatas GBSSI gene for granule-bound starch synthase I, complete cds, clone: 4	95.3	95.3	5%	5e-15	81%	AB524726.1
Ipomoea trifida clone dWx13-H5b granule bound starch synthase I (Waxy) gene, partial cds, alternatively spliced	95.3	95.3	5%	5e-15	81%	EU192912.1
Ipomoea trifida clone dWX13A granule bound starch synthase I (Waxy) gene, partial cds, alternatively spliced	95.3	95.3	5%	5e-15	81%	EU192910.1
Ipomoea batatas GBSSI gene for granule-bound starch synthase I, partial cds, clone: 1	93.5	93.5	5%	2e-14	81%	AB534171.1
Ipomoea batatas GBSSI gene for granule-bound starch synthase I, complete cds, clone: 3	93.5	93.5	5%	2e-14	81%	AB524725.1

上述结果中可见未知 DNA 序列与搜索获得的数据库序列联配得分(score)和统计测验结果(E-value)等。其中一条数据库记录(KP751445)与未知序列具体联配结果如下：

Lilium davidii granule-bound starch synthase 1 gene, partial cds

Sequence ID: gb|KP751445.1| **Length:** 567 **Number of Matches:** 1

Range 1: 1 to 567 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
965 bits(522)	0.0	552/567(97%)	0/567(0%)	Plus/Plus

```
Query  1024  GGGAGAAAAATAAACTGGATGAGGGCTGGAATTTTAGAATCCGACGCCGTTGTAACTGTG  1083
             ||||| ||||| || ||||||| |||||||||||||||| ||||||||||||||||||||
Sbjct  1     GGGAGGAAAATCAATTGGATGAAGGCTGGAATTTTAGAAGCCGACGCCGTTGTAACTGTG  60

Query  1084  AGCCCATACTATGCTAAAGAGCTCGTCTCTGGAGAAGATAAAGGTGTTGAGTTGGACAAA  1143
              || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    AGTCCATACTATGCTAAAGAGCTCGTCTCTGGAGAAGATAAAGGTGTTGAGTTGGACAAA  120

Query  1144  GATATAACCATGATTGGCATCAAAGGGATTGTGAATGGGATGGATATTAATTTTTGGAAT  1203
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GATATAACCATGATTGGCATCAAAGGGATTGTGAATGGGATGGATATTAATTTTTGGAAT  180

Query  1204  CCATTGACAGACAAGTATATCACTGCCAATTATGATGCGACAACGGTAATGGAGGCAAAG  1263
              |||||||||||||||||||||||||||||||||||||||||||||||| ||||||  |||
Sbjct  181   CCATTGACAGACAAGTATATCACTGCCAATTATGATGCGACAACGGTAACGGAGGCGAAG  240

Query  1264  CGTGTCAATAAGCAAGCACTACAAGCAGAAGTTGGCTTGCCTGTAGACCCAGACATTCCA  1323
              |||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  241   CGTGTCAATAAGCAAGCACTACAAGCAGAAGTTGGCTTGCCTGTTGACCCAGACATTCCA  300

Query  1324  GTGATAGTCTTCGTAGGAAGGCTAGAGGAGCAGAAAGGCTCAGACATTCTCGCTGCAGCA  1383
              |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  301   GTGATAGTCTTCGTAGGAAGGCTAGAGGAGCAGAAAGGCTCAGACATTCTCACTGCAGCA  360

Query  1384  ATTCCAGATTTCATTGATGAGAATGTGCAGATAATAATTCTCGGAACCGGCAAGAAAATC  1443
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   ATTCCAGATTTCATTGATGAGAATGTGCAGATAATAATTCTCGGAACCGGCAAGAAAATC  420

Query  1444  TTTGAAAAACAGGTCGAAGAAATAGAAGAAAAGTACCCGGACAAGGCGAGAGGAATTGCG  1503
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   TTTGAAAAACAGGTCGAAGAAATAGAAGAAAAGTACCCGGACAAGGCGAGAGGAATTGCG  480

Query  1504  AAATTCAATATTCCCTTAGCTCATATGATGATGGCTGGAGGTGATCTTATCATAGTTCCT  1563
              |||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   AAATTCAACATTCCCTTAGCTCATATGATGATGGCTGGAGGTGATCTTATCATAGTTCCT  540

Query  1564  AGTAGATTTGAGCCGTGTGGGCTTATT  1590
              |||||||||||| ||||| || || |||
Sbjct  541   AGTAGATTTGAACCGTGCGGTCTCATT  567
```

该搜索相关参数列表：

Search Parameters	
Program	blastn
Word size	28
Expect value	10
Hitlist size	100
Match/Mismatch scores	1,-2
Gapcosts	0,0
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Feb 8, 2016 1:49 AM
Number of letters	111,527,859,682
Number of sequences	34,665,943
Entrez query	none

Karlin-Altschul statistics		
Lambda	1.33271	1.28
K	0.620991	0.46
H	1.12409	0.85

Results Statistics	
Length adjustment	35
Effective length of query	2158
Effective length of database	110314551677
Effective search space	238058802518966
Effective search space used	238058802518966

该参数列表中，可以看到我们用的搜索参数：默认搜索字长（Word size）为 28nt，结果输出的默认联配值（Expect value）10（调低该值，可以减少低相似度序列的输出），序列联配计分系统（Match/Mismatch scores/Gapcosts）为匹配 1 分，错配罚 2 分，不考虑空位罚分（下面 BLASTx 为新开空位罚 11 分，空位每延伸一个罚 1 分）；数据库中 3 466.6 万条序列进行了比对；统计测验参数（Karlin-Altschul statistics）k，λ 和 H 取值情况（详见下节）；数据库搜索的有效序列数据统计结果等列在最后。

同样，可以对该未知 DNA 序列用 BLASTX 工具进行蛋白质序列数据库搜索：

该工具首先按照 6 个 ORF 阅读框对未知 DNA 序列进行翻译，然后比对蛋白质数据库序列数据。搜索结果如下：

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
granule-bound starch synthase [Lilium davidii var. unicolor]	1129	1129	82%	0.0	98%	AJG44453.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Phoenix dactylifera]	831	831	82%	0.0	71%	XP_008775302.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Elaeis guineensis]	821	821	82%	0.0	69%	XP_010940833.1
PREDICTED: granule-bound starch synthase 1b, chloroplastic/amyloplastic-like [Elaeis guineensis]	812	812	79%	0.0	74%	XP_010917976.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Musa acuminata subsp. malaccensis]	801	801	82%	0.0	69%	XP_009415991.1
granule bound starch synthase [Musa acuminata AAA Group]	796	796	82%	0.0	69%	ADZ30929.4
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Vitis vinifera]	796	796	77%	0.0	72%	XP_002273608.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Nelumbo nucifera]	794	794	79%	0.0	71%	XP_010252174.1
UDP-Glycosyltransferase superfamily protein isoform 1 [Theobroma cacao]	793	793	82%	0.0	68%	XP_007039341.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Pyrus x bretschneideri]	792	792	77%	0.0	72%	XP_009366600.1
granule-bound starch synthase 1, chloroplastic/amyloplastic [Nelumbo nucifera]	791	791	79%	0.0	71%	NP_001289785.1
granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Malus domestica]	788	788	77%	0.0	72%	NP_001280836.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Malus domestica]	786	786	77%	0.0	72%	XP_008376222.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Citrus sinensis]	783	783	77%	0.0	72%	XP_006491364.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Musa acuminata subsp. malaccensis]	783	783	82%	0.0	68%	XP_009393091.1
granule-bound starch synthase [Codonopsis pilosula]	783	783	77%	0.0	71%	AJA91185.1
hypothetical protein CICLE_v10019346mg [Citrus clementina]	783	783	77%	0.0	72%	XP_006444732.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic isoform X1 [Jatropha curcas]	782	782	78%	0.0	69%	XP_012086630.1
hypothetical protein PRUPE_ppa002955mg [Prunus persica]	780	780	77%	0.0	72%	XP_007218864.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like isoform X1 [Gossypium raimondii]	780	780	77%	0.0	70%	XP_012439861.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like isoform X2 [Gossypium raimondii]	778	778	78%	0.0	70%	XP_012439862.1
hypothetical protein CISIN_1g007224mg [Citrus sinensis]	778	778	77%	0.0	72%	KDO86605.1
unnamed protein product [Vitis vinifera]	778	778	77%	0.0	72%	CBI34608.3

granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Malus domestica]

Sequence ID: ref|NP_001280836.1| **Length:** 615 **Number of Matches:** 1

See 2 more title(s)

Range 1: 43 to 615 GenPept Graphics　Next Match　Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
788 bits(2035)	0.0	Compositional matrix adjust.	414/573(72%)	474/573(82%)	5/573(0%)	+1

```
Query  244   YQGLKRLKPVDSLQMTATTRSTPRQC-GRSVNC----GGAISCSTGMNLVYVGTETGPHS  408
             + GL+ L  VD L++      S  RQ  G++VN       G I C +GMNLV++GTE GP S
Sbjct  43    HNGLRALNSVDELRVRIMANSVARQTRGKTVNSTRKTSGVIVCGSGMNLVFLGTEVGPWS  102

Query  409   KTgglgdvlgglPPAMAARGHRVMVVTPRYDQYKDAWDTGVVAEFKVGDKIETVRYFHLY  588
             KTGGLGDVLGGLPPAMAA GHRVM ++PRYDQYKDAWDT V  E KVGDK ETVR+FH Y
Sbjct  103   KTGGLGDVLGGLPPAMAANGHRVMTISPRYDQYKDAWDTEVTVELKVGDKTETVRFFHCY  162

Query  589   KRGVDRVFIDHPWFLEKVWGKTGGKLYGPVTGTDYDDNQLRFSLLCLAALEAPRVLNLNN  768
             KRGVDRVF+DHP FLEKVWGKT  K+YGPV G D+ DNQLRFSLLC AAL APRVLNLN+
Sbjct  163   KRGVDRVFVDHPLFLEKVWGKTASKIYGPVAGVDFKDNQLRFSLLCQAALVAPRVLNLNS  222

Query  769   SEYFSGPYGEDVVFIANDWHTGPLSCYLKSMYQAVGIYKSAKVAFCIHNIAYQGRFPFAD  948
             S+YFSGPYGE+VVFIANDWHT  L CYLK++Y+  GIYK+AKVAFCIHNIAYQGRF FAD
Sbjct  223   SKYFSGPYGEEVVFIANDWHTALLPCYLKAIYKPKGIYKTAKVAFCIHNIAYQGRFAFAD  282

Query  949   FSLLNLPdkfkssfdffdGYLKPVKGRKINWMRAGILESDAVVTVSPYYAKELVSGEDKG  1128
             F+LLNLP++FKSSFDF DGY KPVKGRKINWM+AGILESD V+TVSPYYA+ELVS  +KG
Sbjct  283   FALLNLPNEFKSSFDFIDGYNKPVKGRKINWMKAGILESDKVLTVSPYYAEELVSSVEKG  342

Query  1129  VELDKDITMIGIKGIVNGMDINFWNPLTDKYITANYDATTVMEAKRVNKQALQAEVGLpv  1308
             VELD  +    I+GIVNGMD+  WNP+TDKY T  YDA+TV +AK + K+ALQAEVGLPV
Sbjct  343   VELDNILRKSRIQGIVNGMDVQEWNPVTDKYTTVKYDASTVADAKPLLKEALQAEVGLPV  402

Query  1309  dpdipvivFVGRLEEQKGSDILAAAIPDFIDENVQIIILGTGKKIFEKQVEEIEEKYPDK  1488
             D DIPVI F+GRLEEQKGSDIL  AIP FI ENVQII+LGTGKK  EKQ+E++E +YPDK
Sbjct  403   DRDIPVIGFIGRLEEQKGSDILIEAIPHFIKENVQIIVLGTGKKPMEKQLEQLETEYPDK  462

Query  1489  ARGIAKFNIPLAHMMMAGGDLIIVPSRFEPCGLIQLEGMQYGMPVICSTTGGLVDTVKEG  1668
             ARGIAKFN+PLAHM+ AG D ++VPSRFEPCGLIQL  M+YG   I ++TGGLVDTVKEG
Sbjct  463   ARGIAKFNVPLAHMITAGADFMLVPSRFEPCGLIQLHAMRYGTVPIVASTGGLVDTVKEG  522

Query  1669  FTGFHMGAFTVNCEAIDPvdvvatvktvkkalkvYGTPAFSEMVQNCMAQDLSWKGPAKK  1848
             FTGFHMGAF V CE +DPVDV A   TV +AL  YGTPAF+E++ NCMAQDLSWKGPAKK
Sbjct  523   FTGFHMGAFNVECEVVDPVDVQAIATTVTRALGSYGTPAFTEIISNCMAQDLSWKGPAKK  582

Query  1849  WEELLLGLGVHGSQPGIDGEEIAPMSKENVATP  1947
             WEE+LL LGV  S+ GI+GEEIAP++KENVATP
Sbjct  583   WEEVLLSLGVANSELGIEGEEIAPLAKENVATP  615
```

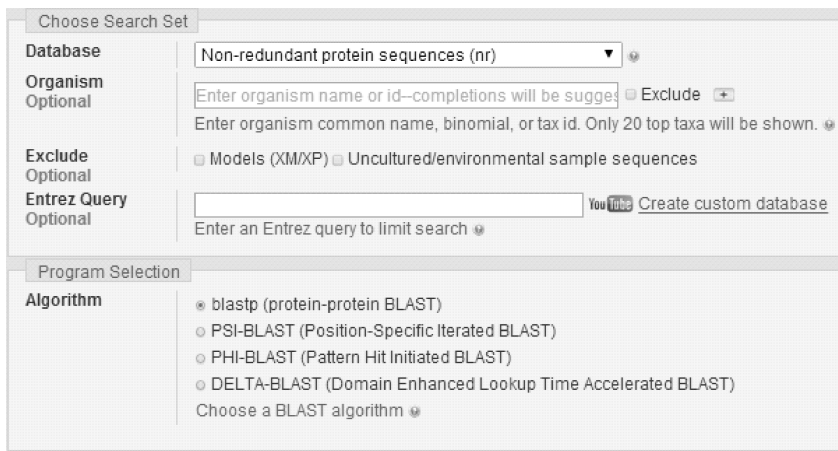
Search Parameters	
Program	blastx
Word size	6
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Low Complexity Filter	Yes
Filter string	L;
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	Feb 8, 2016 1:46 AM
Number of letters	29,838,499,437
Number of sequences	81,622,391
Entrez query	none

Karlin-Altschul statistics		
Lambda	0.317606	0.267
K	0.133956	0.041
H	0.401215	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

可见 BLASTX 与 BLASTN 分别在蛋白质和核苷酸序列水平上进行搜索,所使用的搜索参数完全不同。

前文提及 PSSM,该计分矩阵在生物信息学领域应用非常广泛,一个最常见的应用就是 BLAST 搜索中的应用。在蛋白质序列搜索的几个算法中,PSI-BLAST 和 DELTA-BLAST 均使用了 PSSM(见下图:NCBI 的 BLAST 主页)。这两种算法均利用了一种循环搜索策略,即利用初次 BLASTP 搜索结果,临时构建一个新的计分矩阵(即 PSSM),然后利用该 PSSM 作为下次搜索的计分矩阵,搜索获得结果后再构建新的 PSSM,如此循环往复,直到搜索结果不再变化为止。这样的搜索算法可以最大限度地找到与递交序列具有同源性的序列(有时序列相似性会很低)。



三、序列相似性的统计推断

BLAST 搜索返回的结果中,提供了递交序列与数据库中序列比对结果的得分(score)和一个统计测验结果(E-value)。到目前为止,对局部联配的统计学问题已基本搞清楚,特别是那些不含有空位的局部联配更是如此。我们不妨首先考虑不含有空位的局部联配问题,BLAST 最初的搜索程序便是以此为基础的。

无空位局部联配涉及的是等长度的一对序列片段,两个片段的各部分彼此比较。Smith-Waterman 算法方法可以找到所有最高比值片段对(HSP),即这些片段对的比值(S)不会因片段的延伸而进一步升高。为了分析上述分值随机性产生的几率大小,需要建立一个随机序列模型。对于蛋白质而言,最简单的序列模型可通过从一条序列中随机地选取氨基酸残基(当然这一条序列中各种残基的频率必须一定)。另外,一对随机氨基酸的联配期望值必需为负值,否则不论联配片段是否相关,都会得到高比值,统计理论也将派不上用场。

就像独立随机变量之和总是倾向于正态分布(normal distribution)一样,独立随机变量的最大值倾向于极值分布(extreme value distribution)。在进行两条序列最佳局部联配结果统计测验时,主要涉及的是后一种情况。

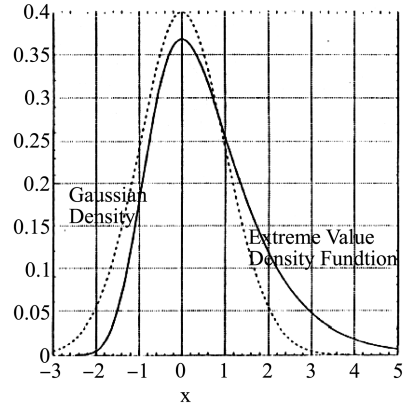


图 1-3.6　概率密度函数正态分布(虚线)和极值分布(实线)比较

对于两条序列,在一定的序列长度 m 和 n 限定下,HSP 的统计值可由 2 个参数(k 和 λ)确定。最简单的形式,即不小于比分值为 S 的 HSP 个数,可由下列公式估计其期望值:

$$E = kmn\mathrm{e}^{-\lambda s} \qquad (1\text{-}3.1)$$

我们称该期望值为比值 S 的 E 值(E-value)。

上述公式非常灵敏。在给定比值的情况下,将两条比较序列长度加倍,则 HSP 数(即 E 值)也将加倍,同样,S 值为 2X 的某个 HSP,其长度必是 S 值为 X 的 HSP 的两倍,所以 E 值将随着 S 值的增大急剧减少。参数 k 和 λ 可分别被简单地视为搜索空间(search space size,即数据库序列数据量)和计分系统的特征数。

最初获得的比值(S)在没有计分系统,或统计量 k 和 λ 的辅助下,没有什么意义。单独的比值就如同没有单位(米或者光年)的距离。可将比值按下式标准化:

$$S' = \frac{\lambda s - \ln k}{\ln 2} \qquad (1\text{-}3.2)$$

获得 S' 值就如同得到了具有标准单位的数值。

E 值因此可简化为:

$$E = mn2^{-S'} \qquad (1\text{-}3.3)$$

二进制值(bit score)使所用的计分系统赋予了统计学意义,除了可以确定搜索空间外,同样可以计算相应的显著水平。

具有大于或等于某一比值 S 的随机 HSP 数可由泊松分布确定。由此可以计算出搜索到某一比值大于或等于 S 的 HSP 的概率为

$$e^{-E}\frac{E^X}{X!} \tag{1-3.4}$$

式中 E 由(1-3.1)式确定。

作为一个特例,搜索不到比值≥S 的 HSP 概率为 e^{-E},所以至少发现一个 HSP 比值≥S 的概率为

$$P=1-e^{-E}=1-\exp(-kmne^{-\lambda x}) \tag{1-3.5}$$

这是与比值 S 相关的 P 值(概率值)。例如,在可能搜索到 3 个 HSP 比值≥S 的情况下,至少发现一个 HSP 的机率为 0.95[可由(1-3.5)式算得]。BLAST 程序中使用了 E 值而非 P 值,这主要是从直观和便于理解的角度考虑。比如 E 值等于 5 和 10,总比 P 值等于 0.993 和 0.999 95 更直观。但是当 E<0.01 时,P 值与 E 值趋于相同。

E 值计算公式[公式(1-3.1)]可以直接应用于 2 个蛋白质序列长度分别为 m 和 n 的比较,但是对于某一序列长度为 m 的蛋白序列,如何在那些长短不一的数据库序列中找到与之匹配良好的序列呢?一种思路是把数据库中的所有蛋白序列与递交序列的关系都视为相同重要,也就是说对于 E 值均较低的短或长序列,它们是等同重要的。FASTA 程序便是采用这一策略。另一种思路是把长序列视为比短序列更重要,因为长序列往往包括更多的特异功能域(domain)。如果对序列长度上进行相关优先处理,则在计算数据库序列长度为 n 的 E 值时,将乘以 N/n,其中 N 为数据库中序列的总长度。根据公式(1-3.1),E 值的计算可简单地把整个数据库序列视为长度为 N 的单条序列。BLAST 程序采用了这一策略。FASTA 策略中 E 值的计算还需再乘上数据库的序列条数。如果考虑到核酸数据库的序列长度变化更大,则在 DNA 序列相似性搜索时,BLAST 的策略可能会是合理的选择。

一些数据库搜索程序(例如 FASTA 或其它基于 Smith-Waterman 算法的程序)在进行序列搜索时,会对数据库中的每条序列进行联配并给出联配值,这些值大部分与递交序列无关,但它们被用于了 k 和 λ 参数的估计。这一方法避免了因使用真实序列(real sequence)导致随机序列模型的偏向性,但同时产生了使用相关序列估计参数的难题。BLAST 仅通过部分而不是全部无关序列计算最适联配值,这赢得了搜索速度。因此,对于某一选定的替换矩阵和空位罚值,必须进行 k 和 λ 参数的预先估计,估计中使用真实序列,而非通过随机序列模型产生的模拟序列。

根据统计理论,上述统计方法只适用于不含有空位的局部联配(非空位联配)。但是,许多计算试验和分析结果充分证明,上述统计方法同样适用于包括空位的联配结果。对于非空位联配,可用基于替换矩阵和比较序列的残基频率的办法,估计统计参数;对于空位联配,参数的估计则必须根据大量无关序列的比较。

以上统计学方法对于短序列来说有些偏差。这些统计方法的基础理论是一个渐近理论,该理论假设局部联配可以适用于任何规模的联配。但是,一个高比值联配必须有一定的长度,不能从接近二条序列末端的地方开始。这种边际效应(edge effect)可以通过计算序列的"有效长度"(effective length)来修正。BLAST 程序中包含了这一修正过程。对于长于 200 残基的序列可以不进行边际效应的修正。

局部联配的结果与所选用的替换矩阵紧密相关。没有任何一个计分系统(即替换矩阵+罚分办法)可以适用于所有研究目标,对于局部联配的计分基础理论的正确理解可以极大促进序列分析准确性(相关内容详见本章第二节)。

如前所述,BLAST 算法中空位罚值应用 2 个参数,一个与空位设置(gap opening)有关,另一个与空位扩展(gap extension)有关。任一空位的出现均处以空位设置罚值,而任一空位的扩大必须加入空位扩展罚值。经过多年的试验,一个合适的空位罚值已经被确定下来。大多数联配程序均对特定的替换矩阵设定了空位罚值的默认值(default),如果使用者希望使用不同的替换矩阵,则原来的空位罚值设定不一定合适。如何设定罚值并无明确的理论可遁,较大的空位设置罚值配以很小的空位扩展罚值(如 11/1)被普遍证实是最佳的设定思路。

在上节 BLAST 搜索结果中,有一个统计量 H。它用于估计 BLAST 搜索所用计分矩阵相对信息量或熵值(具体见第 1-4 章具体介绍),H 值越大,说明所用替换矩阵或计分系统的特异性越高,区分相关与不相关序列的能力就越强(Mount,2004)。

习　题

1. 目前计分矩阵主要有哪些?比较它们的异同。

2. 请利用动态规划 Needleman-Wunsch 算法对下列两条蛋白质序列进行全局联配,获得最优联配结果:

P1=AGWGAHEA

P2=PAWHEAEAG

计分系统:计分矩阵 BLOSUM50,空位罚 8 分。

BLOSUM50 (部分)

	A	E	G	H	P	W
A	5	-1	0	-2	-1	-3
E		6	-3	0	-1	-3
G			8	-2	-2	-3
H				10	-2	-3
P					10	-4
W						15

3. 数据库搜索同源性或相似性较远序列时,为什么用蛋白质序列搜索比 DNA 序列要好?

4. 请解释 BLAST 搜索结果中“score”和“E value”含义。

5. BLASTN 和 BLASTX 两者搜索方式有何不同?