第 1-4 章 多序列联配算法及功能域分析

第一节 多序列联配概念及其算法

一、多序列联配概念

许多生物学研究都涉及多条序列甚至几十上百条序列的比较,因此多序列联配是生物信息学一个重要课题。通过多序列联配结果,我们可以确定这些序列的亲缘关系,通过序列保守性判断功能域或功能位点等等。多序列联配同样包括全局和局部两种联配方式。

上节我们说明了两条序列的联配问题,通过两条序列联配算法(Needleman-Wunsch 算法)和一定的计分系统,我们总是可以获得一个最优联配结果。但是,当我们将三条及以上的序列放在一起联配时,情况就不一样了,问题变得异常复杂。以三条序列为例:如上一章所述,两条序列联配所有可能的联配方式(即路径)均在由两条序列构成的平面内(图 1-3.1),那么三条序列所有可能的联配方式(路径)是在三条序列构成的立体空间内(图1-4.1),从起始到终点可能的路径数量几何方式增长,从中找出最优路径就困难许多。如果三条以上序列进行联配,可能的联配方式就更加巨大,目前还没有一种有效算法能很快获得其最优联配结果。目前实用的多序列联配方法均采取一种所谓启发式方法(heuristic technique),算法往往能给出一个很好的联配结果,但还不能保证给出的一定是最优联配结果。

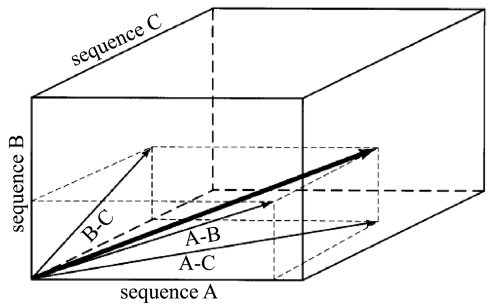


图 1-4.1 三条序列全局联配路径空间示意图(引自 Mount,2004)

目前实用性多序列联配算法可分为几类,如渐进式全局联配(progressive global alignment)、迭代(iterative method)和基于统计模型的方法等。

二、多序列全局联配算法

多序列全局联配算法目前主要是以 Clustal 算法为代表的渐进式全局联配方法。渐进式全局联配算法是 20 世纪 80 年代发展起来的(Waterman 和 Perlwitz, 1984; Feng 和 Doolittle, 1987),其中以软件工具 Clustal 算法最为成功。下面就重点介绍 Clustal 算法。

Clustal 算法是由 Feng 和 Doolittle(1987)等人发展的,后来不断完善和程序化(Thompson 等,1994a; Higgins 等,1996),ClustalW 是目前最新算法和程序版本,目前许多数据库和生物信息学网站均提供该算法在线服务。

Clustal 算法作为渐进式全局联配方法的代表,其基本思路还是利用动态规划算法:首先判断各条序列间差异度大小,然后将最相近的两条序列首先进行序列联配,采取动态规划算

法获得其最优联配结果,然后逐步增加次相近的单条序列或序列联配(作为一条序列看待)。换句话说,由于两条序列的最优联配结果可以很容易地获得,多序列联配便可以在连续使用两条序列联配算法(如 Needleman-Wunsch 算法)基础上,通过先建"树"的思路来进行逐一多序列联配,所以这一方法同样是一种动态规划方法。

多序列联配大致过程如下:

①对所有序列进行两两联配分析,N 条序列应有 N×(N-1)/2 对;

②基于两两联配的结果(如碱基替换率)进行聚类分析,产生联配等级或次序。该等级可用分叉树(binary tree)形式或简单的排序来表示;

③根据以上联配次序,首先从所有联配中相似性最好的两条序列开始,然后是剩余序列中相似性最好的两条序列或一条序列进行联配……依次类推,直至多序列联配结束。一旦两条序列的联配被列入,则序列的位置就被固定下来。例如,对于序列 A、B、C、D,如果 A 与 C、B 与 D 分别是两两联配的最佳联配结果,则 A、B、C、D 四条序列的联配,则通过 A-C 和 B-D 两个联配结果(作为一条序列看待)来进一步联配。

下面以一个实际例子说明其具体算法(Baxenavis 和 Oullette, 2001):

对来自 7 个不同生物的 7 条同源序列(HAHU、HBHU、HAHO、HBHO、MYWHP、PILHB 和 LGHB)进行多序列联配(图 1-4.2)。首先进行两两比对,获得任何两条序列之间的替换率,形成 7×7 的矩阵;根据该两两联配的结果,可以获知替换率最低(相似度最高)到最高的比较结果,构建系统发生树,然后根据系统发生树确定各序列联配次序。首先从相似度最高的两条序列进行联配。树中 HBHU/HBHO 分在一支,替换率最低,首先利用动态规划算法进行它们的两序列联配,获得最优联配结果;依次,HAHU/HAHO 替换率次之,再进行它们两

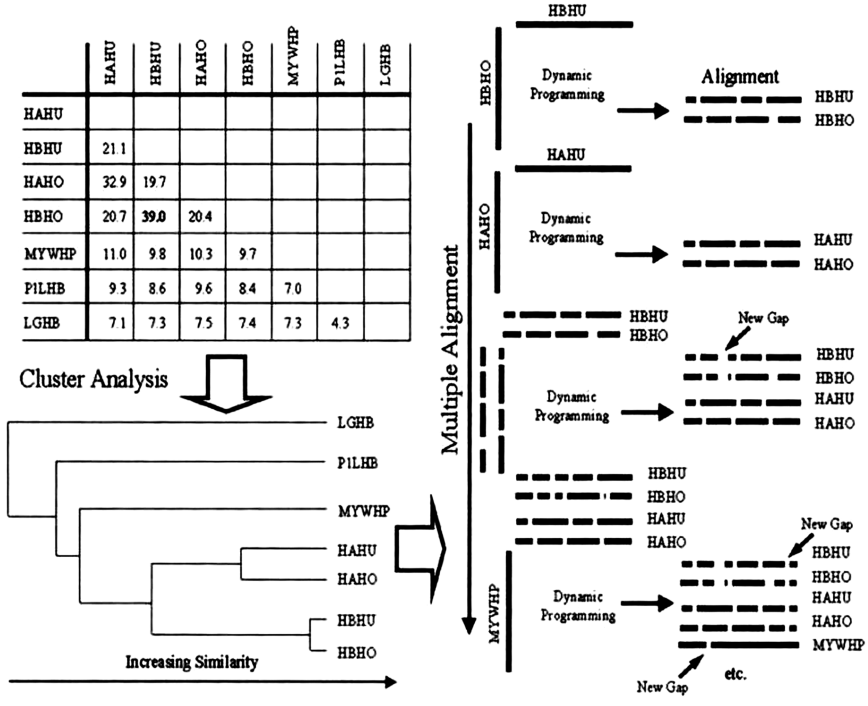


图 1-4.2　多序列联配算法案例(引自 Baxenavis 和 Oullette, 2001)

条序列的联配。根据系统发生树,HBHU/HBHO 和 HAHU/HAHO 次相近,需要对它们进行联配。接下去是该算法的关键:将 HAHU/HAHO 和 HAHU/HAHO 联配结果分别作为一条独立序列,进行两条序列联配! 联配还是采用动态规划算法进行。如此联配,计分方式需进行一定的调整。如一个联配结果作为一条序列看待,联配时加入空格需一起加入。该联配结果进一步与树中最邻近的物种(MYWHP)进行两序列方式联配,直至所有序列都完成联配。由此可见,该算法实际是把多序列联配问题转化为两序列联配问题,而两序列联配已有成熟方法(动态规划方法)。

应该指出,目前还没有一个快速获得最优多序列联配的有效算法,多序列联配程序给出的结果,往往可以通过人为的修正而得到改进。

三、多序列局部联配算法

具有相同功能的基因往往在序列上存在局部相似性或保守性,这些保守性跟相应功能和选择压等有关(详见下节)。图 1-4.3 列举了一个多序列局部保守性 4。生物信息学的一个重要任务是找到这些保守序列。

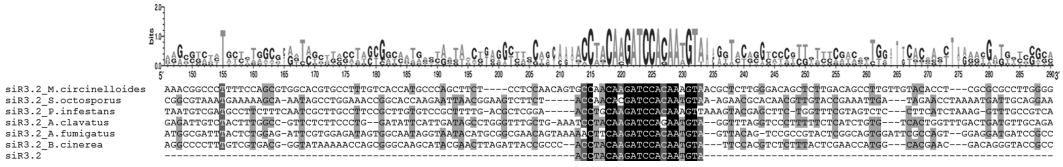


图 1-4.3 多序列局部保守性例举

7 条序列中有一段序列高度保守(黑色区域);图最上一栏为相对信息量图,表示各列信息量大小

多序列局部联配的目的是找出多条序列共同保守的区域。进行多序列局部联配一般可以采取 2 个策略,一个进行上述多序列全局联配,基于全局联配结果,获得局部保守序列的联配结果,这里就不做累述了;另外一个策略是不基于全局联配找出保守区域。该策略目前有两者比较常用的方法,一是简单的哈希(hash)方法,二是基于统计的模式识别方法。

哈希表(Hash table,也叫散列表),是根据关键码值(key value)而直接进行访问的数据结构,利用哈希方法可以实现快速查找。给定表 M,存在函数 f(key),对任意给定的关键字码 key,代入函数后若能得到包含该关键字的记录在表中的地址,则称表 M 为哈希表,函数 f(key)为哈希函数。哈希方法是计算机领域经常使用的算法。

以两条蛋白质序列举例(表 1-4.1)说明利用哈希表进行联配。

首先基于一个统一地址给两条蛋白质序列建立索引,然后每条序列上各个氨基酸在两条序列中的索引位置相减,观察到两条蛋白质序列中,c,s,p 三个氨基酸对应的数值相同,将它们的位置对应起来,这样就找到了一种可能的联配方式。

基于统计的模式识别方法,包括最大期望(EM)、吉布森抽样(Gibbs Sampling)和 HMM 等。以下重点介绍 EM 方法。

EM(Expectation Maximization)算法是于 1977 年被提出,它是进行参数极大似然估计的一种方法。该算法可以基于非完整数据集对参数进行最大似然估计,是一种非常简单实用的学习算法。这种方法可以广泛地应用于处理缺损数据和带有噪声等所谓的不完全数据(incomplete data)。它是一种迭代算法,用于含有隐变量(hidden variable)的概率参数模型的最大似然估计或极大后验概率估计。用一个比喻来说明该算法:比如说食堂的大师傅炒了

表 1-4.1　利用哈希表进行两条蛋白质序列局部联配

位置	1 2 3 4 5 6 7 8 9 10 11		
protein1	n c s p t a		
位置	1 2 3 4 5 6 7 8 9 10 11		
protein2	a c s p r k		
氨基酸	所在位置		位差(offset)
amino acid	protein 1	protein 2	pro.1-pro.2
a	6	6	0
c	2	7	-5
k	-	11	
n	1	-	
p	4	9	-5
r	-	10	
s	3	8	-5
t	5	-	

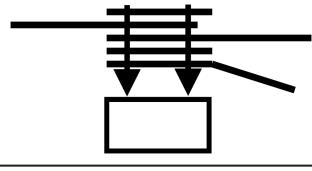
氨基酸 c,s 和 p 在两条蛋白质序列中具有相同的位差,一个可能的联配结果可以马上得出:

protein 1　n c s p t a
　　| | |
protein 2　a c s p r k

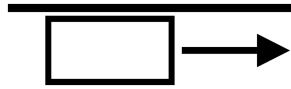
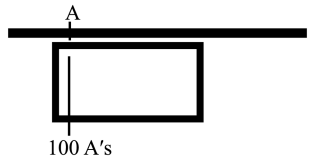
一份菜,要等分成两份给两个人吃,显然没有必要拿来天平精确的去称量,最简单的办法是先随意的把菜分到两个碗中,然后观察是否一样多,把比较多的那一份取出一点放到另一个碗中,这个过程一直迭代地执行下去,直到大家看不出两个碗中的菜有什么分量上的不同为止。EM 算法就是这样,假设我们要估计 A 和 B 两个参数,在开始状态下二者都是未知的,但如果知道了 A 的信息就可以得到 B 的信息,反过来知道了 B 也就得到了 A。可以考虑首先赋予 A 某种初值,以此得到 B 的估计值,然后从 B 的当前值出发,重新估计 A 的取值,这个过程一直持续到收敛为止。

EM 算法通过两个步骤交替进行计算:第一步是计算期望(E),利用对隐藏变量的现有估计值,计算其最大似然估计值;第二步是最大化(M),最大化是在 E 步上求得的最大似然值来计算参数的值。M 步骤上找到的参数估计值被用于下一个 E 步骤计算中,这个过程不断交替进行。

EM 方法在进行多序列保守区块识别分为 9 个步骤完成,其中步骤 1-7 为计算期望(E),步骤 8 为最大化(M)。具体如下:

步骤 1	对多条序列随机排列形成一个多序列联配结果 
步骤 2	对上述多序列联配结果选择一个联配宽度
步骤 3	基于该多序列联配结果和宽度构建其初始 PSSM 矩阵 

续表

步骤 4	利用该 PSSM 对每条序列进行逐位点扫描 
步骤 5	计算每条序列各个位点与 PSSM 的匹配几率(odds): 例如...,100/1,1/25,33/1,1/3,...
步骤 6	合计一条所有匹配几率值(例如 5 000),计算该序列各个位点与上述 PSSM 匹配概率: ...,100/5 000 (即 0.02),0.04/5 000,33/5 000,1/15 000,...
步骤 7	对所有序列进行上述计算
步骤 8	基于上述步骤获得的所有序列 PSSM 匹配概率值更新该 PSSM: 例如上述某一序列某一位点 PSSM 匹配概率为 0.1,而该位点碱基为 A,对应使用的 PSSM 第一列有 100 个 A,则更新 PSSM 第一列 A 的数量为 100.1 个 A;该位点下一个碱基(即 PSSM 第二列位置)依次如此更新。 
步骤 9	基于更新后的 PSSM 重复步骤 4-7,即再扫描所有序列,计算它们的 PSSM 匹配概率。此过程重复 100 次以上,直到 PSSM 各列的碱基频率不再变化为止。

以下举例说明该算法(Mount,2001):

由 100nt 构成的 10 条 DNA 序列,生化和遗传证据表明,它们共同具有一段 20nt 长度的蛋白质结合位点。如何利用 EM 方法在 10 条序列中找到它们保守的 20nt 结合位点?

根据上述算法,首先进行步骤 1-7(期望):随机排列 10 条序列构成一个 20nt 长度的多序列联配结果,然后统计各列和背景(所有序列)碱基构成的频率,得到初始 PSSM(如表 1-4.2)。EM 算法首先随机在 10 条 DNA 序列上确定一个 20nt 区段,组成一个初始的结合位点联配结果。然后统计该 20nt 多序列联配在每个位点(列)上各个碱基个数,然后将它转换为频率。例如:在这 10 条序列的第一个位点(列)总共有 4 个 G,则第一个位点(列)G 的频率为 4/10=0.4。按照这个方法计算每一个在 20nt 蛋白结合位点(列)碱基的频率。对于上述 20nt 区段以外的序列,同样计算每个碱基出现的频率,作为背景碱基频率。四个碱基的背景频率定义为它在背景中出现的次数除以背景碱基总数。例如:G 在 800 个背景中出现了 224 次,则 G 的背景频率为 224/800=0.28。由此可以构建基于随机排列获得的 10 条序列联配结果(20nt 宽度),即 4×20 的初始 PSSM 矩阵(见表 1-4.2)。

表 1-4.2　EM 算法查找多序列蛋白质结合位点案例——初始 PSSM 矩阵的构建

碱基	背景频率	位点(列)1	位点(列)2	…	位点(列)20
G	0.27	0.4	0.1	…	0.2
C	0.25	0.4	0.1	…	0.2
A	0.25	0.2	0.1	…	0.4
T	0.23	0.2	0.7	…	0.2
合计	1.00	1.0	1.0	…	1.0

然后利用初始 PSSM 对序列中每个位点开始扫描,每次扫描窗口长度为 20nt。设序列 1(seq1)中以位置 1(site1)为起始的头两个碱基是 A 和 T,则序列 1 中位置 1 的概率 $P_{site1,seq1}$ = 0.2(A 在位置 1)× 0.7(T 在位置 2)× Ps(下面 18 个位置) × 0.25(A 在 20nt 蛋白结合位点以外的第一位置)× 0.23(T 在第二个边侧位置)× Ps 下面 78 个边侧位置。相似地,可以计算 $P_{site2,seq1}$ 到 $P_{site78,seq1}$。然后综合比较推测结合位点出现在序列中位置的概率。

对于序列 1,最佳定位的概率定义为结合位点在位置 k 的概率,除以所有可能结合位点位置概率的总和:

$$P_{sitek,seq1}/(P_{site1,seq1}+P_{site2,seq1}+\cdots+P_{site78,seq1})$$

对于每条序列计算最佳位点概率(方法同上),用各序列的上述最佳位点概率,生成每个位点碱基数量期望值的新表格,位点概率作为度量。例如:设 20nt 蛋白结合位点出现在 100nt 序列的第一个位点的最佳定位概率 P(序列 1 中位点)= 0.01,出现在第二个位点的最佳定位概率 P(序列 1 中位点 2)= 0.02。在上面的例子中,位置 1 的第一个碱基是 A,位置 2 的第一个碱基是 T。然后 0.01 A 和 0.02 T 作为起始结合位点将加到位点列 1 的累积对应碱基频率表中,对序列 1 的其它 76 个可能结合位点位置重复这一计算过程。相似地,可以从序列 1 的 78 个可能起始结合位点位置 2(20 个 nt 的第二位置)计算位点列 2 的期望值新表。

最后期望值最大化。对所有序列 PSSM 匹配概率值更新 PSSM。然后利用更新后的 PSSM 重复上述过程,直到 PSSM 碱基频率不再变化。

第二节　蛋白质序列功能域分析与模型

一、功能域概念

蛋白质功能域(domain)的概念最早由 Wetlaufer(1973)研究蛋白质结构时提出。蛋白质功能域一般是指一条蛋白质序列中一段保守的区域,该区域能够独立行使功能、进化等。在蛋白质结构中,功能域是指一个蛋白质结构的一部分,它能形成一个紧密的三级结构,能独立折叠且结构稳定,同样具有独立功能和进化等特征。许多蛋白质序列包含若干结构功能域(一般一条蛋白质序列包含 3 个功能域)。在分子进化上,不同功能域可以作为一个单元被重组,产生新的蛋白质序列,行使不同的功能,因此,一个功能域可能在许多不同蛋白质序列中存在。功能域长度不一,可以从 25 到 500 氨基酸不等。

由此可见,功能域可以从序列和结构两个水平上来定义和研究。结构功能域,特别是在二级结构水平上已开展了大量研究,并建立了相应的功能域数据库,如 CATH 和 SCOP 等

(第 1-7 章将详细介绍);在序列水平上,功能域的研究是生物信息学的一个传统研究领域,大量研究者已开展几十年的深入研究,并建立了相应的功能域数据库。本章将重点介绍序列水平的功能域。

说到蛋白质功能域,就不得不说基序(motif)。在序列水平上,基序是指一小段连续的氨基酸或核苷酸序列,它是构成功能域的功能单元。在蛋白质结构水平上,基序可以通过三维结构组成的氨基酸短序列,而这些氨基酸并不一定相邻。基序的概念由 Doolittle 首先提出(Doolittle,1981)。该概念是生物信息学领域一个重要进展(见绪论部分),它是建立序列保守性与序列功能关系的重要基础。一般一个蛋白质功能域由若干个基序串联构成(见图 1-4.4 蛋白质功能域列举)。

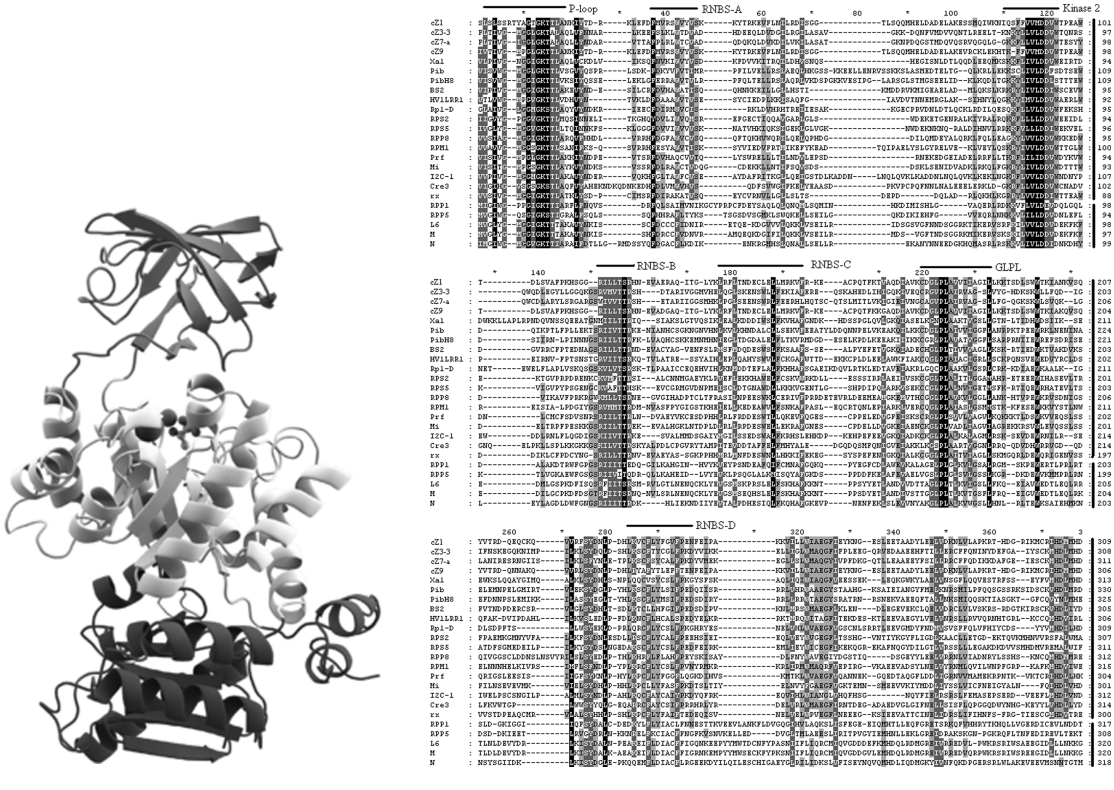


图 1-4.4 蛋白质功能域列举

A. 蛋白质结构水平上的功能域:丙酮酸激酶(Pruvatekinase,PDB 数据库:1PKN)的三个功能域(不同颜色);B.蛋白质序列水平上的功能域:植物 NBS 类抗性基因的 NBS 功能域(引自 Tian 等,2004)。图中标出了该功能域的若干基序(如 Kinase2、GLPL 等)。

由于功能域直接与基因蛋白质功能相关,功能域的查找和应用吸引了大量生物信息学家进行研究,并将发现的基因功能域收集起来,构建所谓蛋白质功能域数据库(表 1-4.3),目前这些数据库在基因功能预测等方面发挥重要作用,特别是 PROSITE 和 Pfam 等数据库。

表 1-4.3　主要国际蛋白质功能域数据库及其它们使用的模型方法

数据名称	模型方法	InterPro 联盟
PROSITE	Pattern/Profile	是
ProDom	Aligned motif (PSI-BLAST) (Pfam B)	是
PRINTS	Aligned motif, OWL	是
Pfam	HMM (Hidden Markov Model)	是
SMART	HMM	是
TIGRfam	HMM	是
DOMO	Aligned motif	否
BLOCKS	Aligned motif (PSI-BLAST)	否
CDD(CDART)	PSI-BLAST(PSSM) of Pfam and SMART	否

二、功能域模型

功能域和基序通过多序列联配等途径可以获得它们的联配结果(如图 1-4.4 NBS 功能域)。在分子生物学领域,大量功能基因被克隆,大量功能域被发现;同时,基于序列分析,也可以发现大量基因共同保守的区段,这些是未知功能的候选功能域。同时,通过同源克隆的途径,已知功能基因的同源基因不断增加,这些包含已知功能域的同源序列会增加功能域的总体遗传多态性。随着功能域数量和序列数据的增加,一个问题随之而来:除了多序列联配结果,是否有更好的方式可以描述这些功能域并在实际功能预测中进行应用? 生物信息学家提出了多种模型来描述功能域,包括一致序列、模式、概型和 HMM 模型等(表1-4.3),其中概型和 HMM 模型在生物信息学领域应用最为广泛。

一致序列(consensus sequence)和模式 pattern 相对比较简单。一致序列是多序列联配结果中每一列出现最多的碱基或氨基酸(或使用兼并码)构成的序列,它是一条单一序列,而正则表达式则是把每一列出现的碱基或氨基酸都列出,形成一个正则表达式(regular expression)(图 1-4.6)。这两种功能域描述方式或模型在实际功能域数据库中很少应用(除了 PROSITE 数据库利用模式),大量使用的是概型(Profile)和 HMM 模型(Profile HMM)。

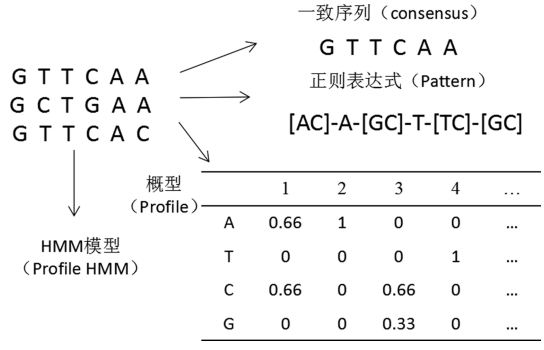


图 1-4.6　多序列联配结果(功能域)的四种描述方法/模型

概型是一个类似 PSSM 的矩阵,但它可以包含匹配、错配、插入和缺失等情况(详见第 1-3 章 PSSM 构建方法)。该矩阵提供了多序列联配(功能域)中每一列出现各种氨基酸(或空格)的概率(经过对数转换并取整数)。所以,概型的矩阵一般为 23 列,其中每种氨基酸一列,同时,该功能域序列的一条一致序列会列在矩阵的左侧。图1-4.7 列举了 PROSITE 功能域数据库(http://prosite.expasy.org)的一个概型记录(PIWI,记录号 PS50822)。该功能域比较长(297 个氨基酸),图中仅列出该记录总体信息及其部分概型结果。概型中首先是说明行,说明该概型的总体信息,如长度、默认参数值等(本例为 6 行),然后是矩阵横列,为 22 个氨基酸,纵列为功能域序列纵排,矩阵中标出功能域中特定位点出现各种氨基酸的频率(以对

数转换),同时包括特定位点出现(插入)空格的频率等。

Entry: PS50822

General information about the entry	
Entry name [info]	PIWI
Accession [info]	PS50822
Entry type [info]	MATRIX
Date [info]	MAY-2002 (CREATED); OCT-2013 (DATA UPDATE); MAR-2016 (INFO UPDATE).
PROSITE Doc. [info]	PDOC50822
Associated ProRule [info]	PRU00150

Name and characterization of the entry	
Description [info]	Piwi domain profile.
Matrix / Profile [info]	

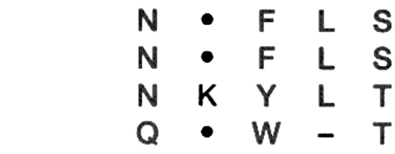
```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=297;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=292;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=2.1406; R2=0.01583904; TEXT='NScore';
/CUT_OFF: LEVEL=0; SCORE=402; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=276; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: MO=-8; D=-20; I=-20; B1=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105;

                A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:          B1=0; BI=-105; BD=-105;
/M: SY='L'; M=-11,-16,-11,-18,-19, -4,-19,-18,  7,-23, 11,  3,-16,-25,-18,-22,-17,-11,  2,-21, -1,-20;
/M: SY='L'; M= -8,-30,-19,-33,-26, 13,-32,-25, 27,-27, 28, 15,-27,-28,-25,-22,-21, -7, 24,-20,  0,-26;
/M: SY='M'; M=  3,-20,-20,-26,-14,  1,-22,-17, 14,-17, 11, 15,-18,-19,-13,-18,-12, -7, 10,-19, -4,-14;
/M: SY='V'; M= -8,-25,  6,-29,-24,  3,-30,-25, 11,-20, 11,  5,-23,-29,-24,-18,-17, -7, 17,-25, -6,-24;
/M: SY='I'; M= -6,-27,-19,-31,-25,  8,-30,-25, 16, 11,-22,-25,-23,-22,-13, -4, 26,-22, -2,-25;
/M: SY='L'; M= -9,-29,-21,-32,-23,  6,-31,-21, 28,-27, 37, 23,-27,-27,-19,-21,-25, -9, 18,-21, -1,-22;
/M: SY='P'; M= -7, -3,-29,  2,  3,-27,  1,-13,-27, -5,-26,-19, -3, 19, -5, -7,  2, -3,-24,-29,-23, -3;
/M: SY='E'; M= -4,  8,-21, 10, 14,-24,-12, -7,-21,  8,-24,-16,  8,-11,  4,  0,  9,  0,-15,-33,-17,  9;
/M: SY='B'; M=-12,  9,-26,  9,  8,-12,-17,  3,-15, -3,-17,-14,  9, -8, -2, -7, -4, -6,-18,-20,  2,  2;
/I:          I=-4; MI=0; MD=-19; IM=0; DM=-19;
/M: SY='N'; M= -7,  7,-22,  0, -3,-19, -2, -5,-16,  7,-21,-11, 16,-15, -1,  6,  2, -4,-16,-25,-14, -3; D=-4;
/I:          I=-4; DM=-19;
/M: SY='D'; M= -3,  7,-23, 12,  8,-26, -3, -8,-26,  3,-23,-17,  3,  0,  0,  0,  3, -6,-20,-26,-18,  3; D=-4;
/I:          I=-4; DM=-19;
/M: SY='E'; M= -8, 10,-22,  8, 13,-12,-13, -5,-16, -5,-18,-15, 13,-12,  0, -8,  7, -1,-17,-30,-13,  6;
/M: SY='R'; M=-10, -6,-23, -6, -4, -8,-20,-13,-14, -1, -7, -7, -7, -7, -7, -7,  1, -5,  0,-11,-24, -8, -6;
/M: SY='Y'; M=-16,-22,-30,-20,-15, 12,-28,  3, -1,-14,  3, -1,-22, -2,-12,-14,-20,-10,-11,  6, 39,-18;
```

图 1-4.7 概型(Profile)列举(具体说明见正文)。

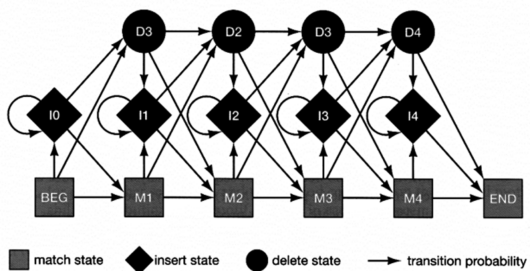
图中:I: Profile insert position; M: Profile match position; D: Deletion extension score; MI: State transition score from state M to I; M0: Match extension score for a character not included in thealphabet; B1: Internal initiation score; E0: External termination score

很明显,构建好的功能域概型可以用于序列比对和搜索,确定未知序列中可能与该概型矩阵高度相似的区段。反之,如果某一序列包含特定功能域(统计上可以给出推测),则该蛋白质序列可能具有该功能域相同的功能。这是目前基因功能预测的主要生物信息学途径之一。

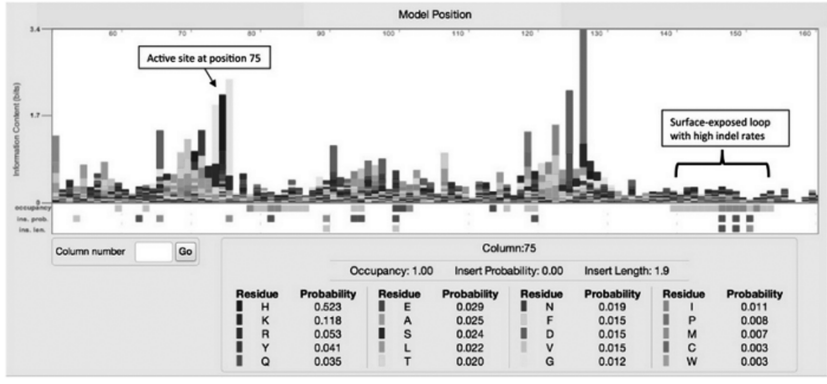
HMM(隐马尔科夫模型)方法则是通过对多序列联配构建隐马尔科夫概率模型,进行功能域描述(有关 HMM 的具体介绍见第 1-5 章基因预测模型一节)。图 1-4.8 给出了一个被广泛应用的“左-右”(left-right)结构模型——标准线性结构 HMM 模型。所谓“左-右”结构是指该结构中不存在从一种状况回复到已有状况的情况。对于 HMM 模型,将一个功能域(图 1-4.8a 给出多序列联配结果)视为一个从左开始到右结束各个状态(氨基酸匹配或错配、插入和删除)之间的转换(图 1-4.8b)。该模型各个“态”之间转换会有一个频率,每个“态”所处的具体状态(如各种氨基酸)存在一个概率分布(图 1-4.8c),具体状态是未知的(所谓“隐”)。具体而言,可以从标有 BEG 的状态开始,然后沿任意一条路径(如状态转换箭头所示)行走,最后在标有 END 的状态结束。任意一个多序列联配结果都可以用该模型生成,而且每种可能路径都有一个概率。例如生成“NKYLT”序列的一种方法:BEG-> M1-> I1-> M2-> M3-> M4-> END。每个态间转换都有一种概率,且离开一种状态的转换概率之和为 1。就像其他统计方法一样,氨基酸的分布和转换概率可以转换为对数几率计分。



(a)



(b)



(c)

图 1-4.8　功能域 HMM 模型例举(部分引自 Mount, 2004)

a.一个多序列联配结果。对于一个多序列联配结果,可能存在氨基酸匹配或错配(如第 1、3 和 5 列)、插入(第 2 列)和删除(第 4 列);b.多序列联配的马尔科夫模型。每个方形表示氨基酸联配状态,菱形表示插入状态,圆形表示缺失状态,箭头表示转移概率; c.来自 Pfam 功能域数据库的一个 HMM 记录(具体说明见文中)。

一个功能域 HMM 记录见图 1-4.8c。这是功能域数据库 pfam 的一个记录(http://pfam.xfam.org/)。该记录以类似序列信息量徽标(logo)形式呈现。徽标图内包括每个位点上 20 种氨基酸的出现概率(用比特表示),徽标图下面三行给出了 HMM 模型三个态(匹配或错配、插入和删除)之间的转移概率(有关信息量和序列徽标详见下节说明)。作为一个说明,图中下方特别给出了第 75 位氨基酸的相应转移概率和 20 种氨基酸出现概率。根据该序列徽标,可以清楚看到该功能域蛋白质序列的保守区域,如第 75 位点附近区域。

一个功能域 HMM 构建完成后,其可以利用的领域或方式与概型一样。

第三节 熵与信息量

一、不确定性与信息量

信息量或信息熵的概念(来自信息论)。当我们说一条信息或消息,我们会问其信息量有多大?或它可以提供我们多少明确的信息?当一条消息的信息量越大,其不确定性就越小。例如,我们说"今年将在中国召开 G20 峰会"和"2016 年 9 月将在杭州召开 G20 峰会",这两条信息的信息量明显不同,后一条消息包含更大信息量,它涵盖了前一条消息。

不确定性(uncertainty)可以用必须提问的次数来度量,即你为了获得明确信息你不得不问相关问题的次数(答复只有"是"和"不是"两种)。例如 64 个反扣的杯子,其中只有一个杯子里有一个乒乓球。我们的问题是我们需要问多少次才能知道哪个杯子中有乒乓球(答复同样只有"是"和"不是"两种)?我们最少要提问多少次才能获知乒乓球在哪个杯子里?答案是 6 次(通过把杯子等分成 2 组来不断提问)。

一个序列方面的例子(表 1-4.4):一个多序列联配结果,其 3 个位点(列)碱基构成不同。很明显,这三列碱基构成不确定性不同:第一列我们不需要问问题,就知道其碱基构成为 G,而第二列,我们需要至少需问一个问题才能明确其是 A 还是 T,而第三列则需要问 2 次才能知道特定碱基类型。也就是说,从左到右三列的不确定性逐渐增大,反之,不确定性增大,它可以提供的信息量就减少了。

表 1-4.4 一个多序列联配结果三列碱基构成比例

	1	2	3
G	1.0	0.0	0.25
A	0.0	0.5	0.25
C	0.0	0.0	0.25
T	0.0	0.5	0.25

用概率来估计不确定性标志着信息论的起始(Hartly, 1928):

$$H=\log N=-\log P,\quad P=1/N$$

H:不确定性,N:事件可能发生的总数,P:事件发生的概率

香农(Claude Elwood Shannon)的一个重要贡献是将不同事件发生概率作为权重,重新定义不确定性,即香农信息熵(H):

$$H=-\sum p_i\log_2(p_i)$$

p_i 是特定事件 i 发生的概率

上述杯子的例子,找到特定杯子(内有乒乓球)的概率是 1/64,所以其不确定性 H 为以 2 为底的负对数$[-\log_2(1/64)]=6$ 比特。对于上述多序列联配结果,某一列仅观察到一个碱基(第一列),则其不确定性或熵值为 0,因为没有其他可能性。而对于第二列存在两个等概率的碱基,你必须问一个问题后才能得到答案,这样其熵值为$[-(0.5\times\log_2 0.5+0.5\times\log_2 0.5)]=1$.

熵的概念最早由克劳修斯(Rudolf Clausius,1822—1888,德国)提出并应用于描述热力

学第二定律,后来香农第一次将熵的概念引入信息论中。其德文(entropie)或英文(entropy)均与“能”有关。1923 年,我国物理学家胡刚复在翻译该词时找不到一个确切的词,于是造了一个新字——“熵”:因其为热量与温度之商,且与火有关(象征着热),因此商字上加火字旁。目前熵在控制论、概率论、数论、天体物理、生命科学等领域都有重要应用,且不同学科中有更为具体的定义,成为各领域十分重要的参量。总体上,熵是指体系的混乱程度。在生物信息学领域,其与不确定性(uncertainty)等价。

二、信息熵的应用

在生物信息学领域,信息熵有两个广泛应用,如计分矩阵信息量的估计和序列保守性的图形描述等。

1. 计分矩阵

我们构建了一个计分矩阵(如 PSSM、PAM 和 BLSUM 矩阵),我们经常会问的一个问题是,该矩阵用于序列搜索或保守区段的搜索效果如何? 该问题等于问该矩阵的信息量如何。生物信息学领域往往用不确定性参数(H)来度量。

一般来说,PSSM 特定列(如 c 列)各个氨基酸或碱基不确定性的平均数(H_c)(以比特为单位)由下式给出

$$H_c = -\Sigma p_{ic} \log_2(p_{ic})$$

其中 p_{ic}是 c 列第 i 个氨基酸或碱基频率。对于整个 PSSM 矩阵,其不确定性为各列之和

$$H = \Sigma H_c$$

H 在信息论中被称为 PSSM 的信息熵。因为这个值越高,不确定性就越大;相反,不确定性 H 值越低,表明其信息量越大,PSSM 用于从随机匹配鉴别真实基序的能力就越强,可应用性越好。同样,PAM 和 BLSUM 计分矩阵的不确定性或信息量也可以用 H 值估计。例如不同 PAM 距离的矩阵,从 PAM10 到 PAM250,其 H 值为从高到低,表明其不确定性在不断降低,其区分随机匹配的能力在不断增强。同时,作为序列联配计分系统的一部分——空位罚分方式也会影响整个计分系统的 H 值(详见 Mount,2004)。

同时我们也可以用信息量(information content,IC)来评估计分矩阵。信息量与不确定性 H 值关系:

对于核苷酸 PSSM 的特定列,$IC = 2 - H$

整个 PSSM 的信息量为各列之和。式中“2”是针对核苷酸序列,针对氨基酸序列等用更大值(如 4.32),该值往往为 H 理论最大值。

根据上式,我们知道计分矩阵的信息量越大,其不确定性或信息熵值越小。

以表 1-4.4 的多序列联配结果为例,进行信息量计算:各列四种碱基组成不同,其相应的信息量也相应发生变化(表 1-4.5):

表 1-4.5　多序列联配位点碱基构成比例及其信息量估计

A%	C%	G%	T%	信息量
100	0	0	0	2
95	5	0	0	1.71
90	10	0	0	1.53
85	5	5	5	1.15
80	10	5	5	0.98
70	10	10	10	0.64
50	50	0	0	1
50	40	5	5	0.54
45	45	5	5	0.53
50	30	10	10	0.31
35	35	15	15	0.12
25	25	25	25	0

2. 序列信息量徽标

序列徽标或标志(sequence logo)是一种描述功能域或保守区段信息量的可视化图形方式,它将保守区段每列的信息量大小通过氨基酸或碱基字母大小方式进行表示。序列徽标的纵坐标为信息量(比特),横坐标为各列依次出现的字母。

例如图 1-4.9,入噬菌体 *cI* 和 *cro* 蛋白结合绑定位点在碱基水平上表现出保守性(图上部),特别是在其+5 和+7 等位点仅观察到一种碱基构成。计算其不确定性,说明+5 和+7 等位点 *H* 熵值为 0,其信息量 IC 为最大。反之,其他位点 *H* 值较大,则其信息量就偏低。

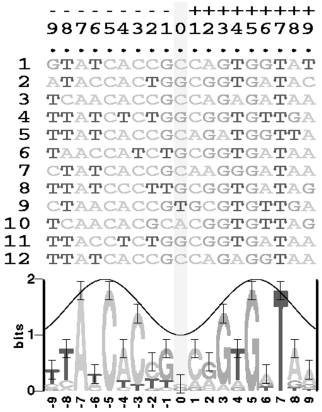


图 1-4.9 序列信息量徽标(sequence logo)列举

图中为 12 条 **λ** 噬菌体 ***cI*** 和 ***cro*** 蛋白结合位点序列联配结果(上)和各位点信息量情况(下)

习 题

1. 简述渐进多序列联配算法(ClustalW 算法)
2. 什么是功能域和基序(motif)?
3. 简述几种功能域的描述方式(模型)
4. 请构建多序列联配结果(功能域)的马尔科夫模型并简要说明
5. 说明 PSSM 等矩阵的熵(*H*)和信息量(IC)的概念