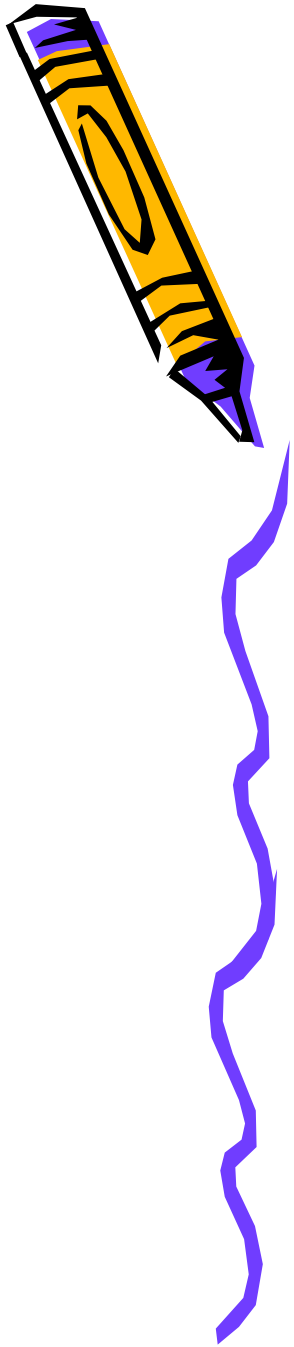


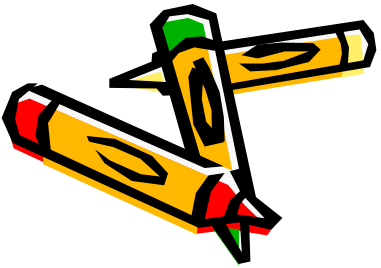
Scoring Matrices

in

Multiple Sequence Alignment

Yuan Yaxia

May 2007



Outline

- 何为打分矩阵
- 简单打分矩阵
- PAM矩阵
- Blosum矩阵
- 总结

何为打分矩阵

定义:

给不同的序列匹配定义的一系列相似性分值。

目的:

我们并不能直接计算出两条序列的最佳匹配，因此需要找到一个可以估计任何匹配的某一统计数，使生物学序列匹配最显著的匹配统计数最大。

	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

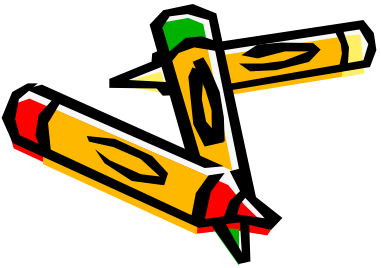
简单打分矩阵

单一打分矩阵:

如果两个氨基酸相同，就打一个分值，不同就打另一个分值，不管替换的情况。例如，相同就打1分，不同就打0分，这就是最简单常用的单一打分矩阵。

遗传密码子打分矩阵:

所有的点突变都产生于核苷酸的变化，因此氨基酸替换的分值应取决于由一个密码子转变为另一密码子所必需的点突变的数量。由这一模型而产生的打分矩阵将根据导致密码子改变所需改变核苷酸的数量来定义两个氨基酸之间的距离，此为遗传密码子打分矩阵



遗传密码子打分矩阵

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	.
3.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	A
	3.0	1.0	3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0	3.0	1.0	2.0	1.0	2.0	2.0	2.0	0.0	2.0	2.0	B
		3.0	1.0	0.0	2.0	2.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	2.0	2.0	1.0	1.0	2.0	2.0	0.0	C
			3.0	2.0	1.0	2.0	2.0	1.0	1.0	1.0	0.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	0.0	2.0	2.0	D
				3.0	0.0	2.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	3.0	E
					3.0	1.0	1.0	2.0	0.0	2.0	1.0	1.0	1.0	0.0	1.0	2.0	1.0	2.0	1.0	2.0	0.0	F
						3.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	G
							3.0	1.0	1.0	2.0	0.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	0.0	2.0	2.0	H
								3.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	0.0	1.0	1.0	I
									3.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	2.0	K
										3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	L
											3.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	0.0	1.0	M
												3.0	1.0	1.0	1.0	2.0	2.0	1.0	0.0	2.0	2.0	N
													3.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	2.0	P
														3.0	2.0	1.0	1.0	1.0	1.0	1.0	3.0	Q
															3.0	2.0	2.0	1.0	2.0	1.0	2.0	R
																3.0	2.0	1.0	2.0	2.0	1.0	S
																	3.0	1.0	1.0	1.0	1.0	T
																		3.0	1.0	1.0	2.0	V
																			3.0	1.0	1.0	W
																				3.0	1.0	Y
																					3.0	Z

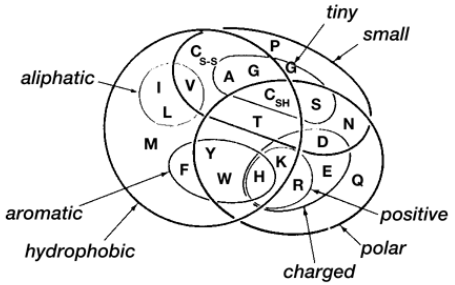
PAM矩阵

Dayhoff打分矩阵:

Dayhoff及其同事利用在70年代初期做的一个蛋白质序列和进化数据集，对一些哺乳动物蛋白质序列的比对发展出了一个精确的突变打分矩阵(mutation data matrix)。这个打分矩阵对特定蛋白质序列比对中，序列的差异是随机发生的还是源自共同祖先序列的机率作了定量。

(a) TTYGAPPWCS
TGYAPPPWS
* *** *

(b) TTYGAPPWCS
TGYAPPPWS
* * ***



PAM矩阵

表 3.11 氨基酸替换次数表 (Dayhof 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

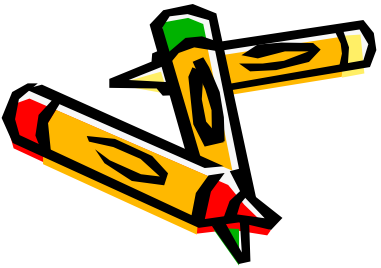
PAM矩阵

PAM是一个进化时间单位:

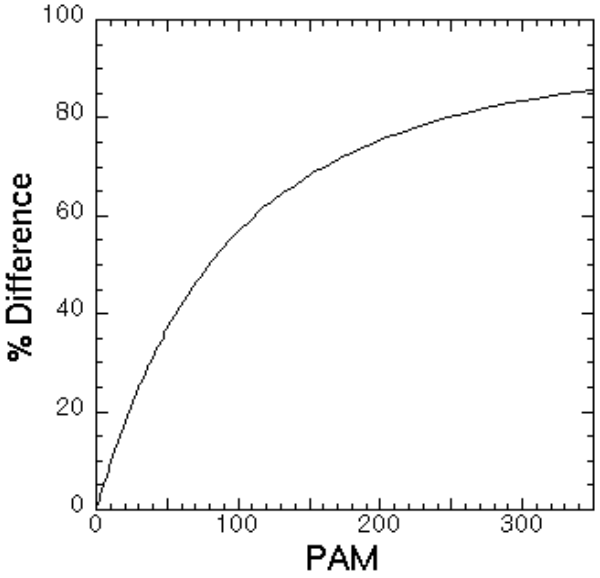
假设同一位点不会发生二次以上的突变，则1PAM等于100个氨基酸多肽链中预期发生一次替换所需的时间。1PAM相当于所有的氨基酸平均有1%发生了变化，经过100PAM的进化，并非每个氨基酸的残基均发生变化：有一些可能突变多次，甚至又变成原来的氨基酸，而另一些氨基酸可能根本没有发生过变化。因此利用大于100PAM的时间间隔可能达到区分同源性蛋白质的目的。

N PAM:

表示对原始PAM矩阵N次方。



PAM矩阵



%Difference	PAM	%Difference	PAM
1	1	45	67
5	5	50	80
10	11	55	94
15	17	60	112
20	23	65	133
25	30	70	159
30	38	75	195
35	47	80	246
40	56	85	328

PAM250矩阵相当于约20%匹配率。
而50%匹配率约为PAM80。

Blosum矩阵

出发点：

Dayhoff模型假设，蛋白质序列各部位进化的速率是均等的。但事实上并非如此，因为保守区的进化速率显然低于非保守区。

Henikoff算法：

对不同家族蛋白质序列片段的区间(blocks)进行比对，不加入gaps，这些序列区间对应于高度保守的区域。氨基酸匹配率可通过各区间可能的匹配率得到。再将这些匹配率计入匹配率表。其进化相关机率的计算方法与Dayhoff矩阵相似。

Blosum矩阵

N Blosum:

指以簇群方式将不同进化距离（N%)整合进矩阵内：当两个序列匹配的匹配率高于某个阈值时便归为一个簇群。将匹配率高于阈值的序列加入簇群内。然后将以簇群内所有序列计算匹配率表，从而也象PAM矩阵一样产生一系列的矩阵。

N表示簇群的阈值水平， N越大，表示关系越近。Blosum80指以80%匹配率为阈值将序列区间归为簇群。Blosum62最接近于PAM250。

Blosum62

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z	
4.0	-2.0	0.0	-2.0	-1.0	-2.0	0.0	-2.0	-1.0	-1.0	-1.0	-1.0	-2.0	-1.0	-1.0	-1.0	1.0	0.0	0.0	-3.0	-1.0	-2.0	-1.0	A
	6.0	-3.0	6.0	2.0	-3.0	-1.0	-1.0	-3.0	-1.0	-4.0	-3.0	1.0	-1.0	0.0	-2.0	0.0	-1.0	-3.0	-4.0	-1.0	-3.0	2.0	B
		9.0	-3.0	-4.0	-2.0	-3.0	-3.0	-1.0	-3.0	-1.0	-1.0	-3.0	-3.0	-3.0	-3.0	-1.0	-1.0	-1.0	-2.0	-1.0	-2.0	-4.0	C
			6.0	2.0	-3.0	-1.0	-1.0	-3.0	-1.0	-4.0	-3.0	1.0	-1.0	0.0	-2.0	0.0	-1.0	-3.0	-4.0	-1.0	-3.0	2.0	D
				5.0	-3.0	-2.0	0.0	-3.0	1.0	-3.0	-2.0	0.0	-1.0	2.0	0.0	0.0	-1.0	-2.0	-3.0	-1.0	-2.0	5.0	E
					6.0	-3.0	-1.0	0.0	-3.0	0.0	0.0	-3.0	-4.0	-3.0	-3.0	-2.0	-2.0	-1.0	1.0	-1.0	3.0	-3.0	F
						6.0	-2.0	-4.0	-2.0	-4.0	-3.0	0.0	-2.0	-2.0	-2.0	0.0	-2.0	-3.0	-2.0	-1.0	-3.0	-2.0	G
							8.0	-3.0	-1.0	-3.0	-2.0	1.0	-2.0	0.0	0.0	-1.0	-2.0	-3.0	-2.0	-1.0	2.0	0.0	H
								4.0	-3.0	2.0	1.0	-3.0	-3.0	-3.0	-3.0	-2.0	-1.0	3.0	-3.0	-1.0	-1.0	-3.0	I
									5.0	-2.0	-1.0	0.0	-1.0	1.0	2.0	0.0	-1.0	-2.0	-3.0	-1.0	-2.0	1.0	K
										4.0	2.0	-3.0	-3.0	-2.0	-2.0	-2.0	-1.0	1.0	-2.0	-1.0	-1.0	-3.0	L
											5.0	-2.0	-2.0	0.0	-1.0	-1.0	-1.0	1.0	-1.0	-1.0	-1.0	-2.0	M<
												6.0	-2.0	0.0	0.0	1.0	0.0	-3.0	-4.0	-1.0	-2.0	0.0	N
													7.0	-1.0	-2.0	-1.0	-1.0	-2.0	-4.0	-1.0	-3.0	-1.0	P
														5.0	1.0	0.0	-1.0	-2.0	-2.0	-1.0	-1.0	2.0	Q
															5.0	-1.0	-1.0	-3.0	-3.0	-1.0	-2.0	0.0	R
																4.0	1.0	-2.0	-3.0	-1.0	-2.0	0.0	S
																	5.0	0.0	-2.0	-1.0	-2.0	-1.0	T
																		4.0	-3.0	-1.0	-1.0	-2.0	V
																			11.0	-1.0	2.0	-3.0	W
																				-1.0	-1.0	-1.0	X
																					7.0	-2.0	Y
																						5.0	Z

总结

PAM:

对相关性未知的序列进行比对：只进行一次比对时常用PAM120矩阵。如想得到结果更全面更有效的结果则应使用多个矩阵。用三个矩阵：PAM40、PAM120、PAM250，可得出全面覆盖的结果。只用PAM80和PAM200两个矩阵也可达到较好的覆盖面。

对两个同源序列进行比对：多用几个不同的PAM矩阵会得到较好的结果。如果只进行一次比对常用PAM200矩阵。如果进行两次分析，那用PAM80和PAM250，或者PAM120和PAM320可以得到较好的结果。

作比对最好是根据序列对实际差异程度来选用相应的PAM矩阵。

总结

Blosum:

PAM矩阵从1到250PAM两极距离太远，可能引起不准确；而Blosum直接从最同源的序列的区间排比获取匹配率，不考虑进化距。因此Blosum矩阵的优点是符合实际观测结果，不足之处是它不能提供进化信息。

Blosum矩阵的突变数据来源于未加gaps的序列区间排比，相当于蛋白序列的保守区。大量试验表明，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。