阅读日期：20190422

发表日期：19861109

发表杂志：Biochemistry

影响因子：

平均影响因子：

标签：

|  |  |
| --- | --- |
| 文章名 Profile analysis: Detection of distantly related proteins | |
| 中文译名 | 远亲蛋白的谱分析检测 |
| 网址 |  |
| 关键词 |  |
| 摘要 |  |
| 启发 | 1，新的蛋白质出现了，若它的序列和已知结构蛋白的序列足够相似，则可以推测他们有相似的结构。问题在于如何确定“相似度”。  2，本篇论文所提出的问题是：检测远缘蛋白的相似度。  3，文章用globin举例说明，为什么要研究它。Globin在人，鼠，树上有不同的体现，长度为150的序列中只有2个位置对应的氨基酸是相同的。130个氨基酸是不相同的。而每一段氨基酸代表一种折叠，我们怎么区分或找到这种折叠的侧面描写呢？于是开发出了这个矩阵来完成。  4其他的解决方案：dot matrix method; dynamic programming methods; rapid database searching methods.等等。相比它们，本文提出的“Profile”方法在结构和家族方面都节省了计算时间。  5  6  7 |
| 摘录 | 1，profile analysis有2步，  (i)construction of the profile with the program PROFMAKE, and  (ii)comparison of the profile with database of sequences or a single sequence (program PROFANAL).  第一，建立矩阵；第二，与数据库中的序列或者单独序列进行比较  2，第一步中输入的可以是这两种情况：一条序列，或者是一组序列（probe）  Probe:通常是功能相似的或结构相似的典型的蛋白质。  3，profile 是一个序列位置特异性得分矩阵。M(p,a).有21列，N行。N是probe的长度。  【问题】那要是输入的是一组蛋白，这个N是其中一条的长度还是所有累加起来的长度？  答：一组蛋白的话，应该是用来算出现概率的。  4计算方案：    The profile is a sequence position-specific scoring matrix M(p,a)composed of 21 columns and N rows ( N = length of probe).  Profile是一个位置特异性得分矩阵，有21列N行（N取决于序列长度。）  The row p corresponds to a sequence position of the probe.  行p,对应于prob的位置  The first 20 columns of each row specify the score for finding, at that position in the target, each of the 20 amino acid residues.  每行中靠前的20列，是该行每种氨基酸在目标里指定位置中的得分。  An additional column contains a penalty for insertions or deletions at that position  多出来的这一列包含了penalty,因发生在这个位置的插入或删除。  In PROFMAKE, the profile is generated from the probe by using a comparison table derived from the mutational distance matrix (MDM78) of Dayhoff .  在这个程序中，profile是通过使用一个对比表产生的，这个对比表来自于变异距离矩阵。Dayhoff的研究。  The value of the profile for amino acidu at positionp is  这里给出了每位数值的计算公式。  （我很好奇整数是怎么计算出来的。）  where Y(a,b) is Dayhoff s matrix 其中一项很好获得，是Dayhoff矩阵中对应位置的数值。  and W ( p， b )is a weight for the appearance of amino acid b at position p.  而W则是氨基酸b出现在位置p的权重。  This weight is determined as follows: 这个权重这么算：  Suppose that amino acid b appears n(b.p) times in position p in the NR probe sequences. Then a simple average weight is given by W(b,p) = n(b,p)/NR.  假设b在p处出现n 次，那么这个权重就是n/NR。  这里NR是参与比较的序列的个数。  Another useful weighting (19) is to take Wasproportional to log[n(b,p)/NR], setting n(b,p)= 1 for any amino acid that never appears at p .  This logarithmic weighting may be better when there are several closely related sequences, but the simple weighting is used for the examples given in this paper.  还有一种权重计算方法是这样的，它适宜非常相近的序列中。本文中使用的示例是简单的权重。  5，插入和消失更倾向于发生在功能结构域之间，而不是功能结构域之内。  6，penalty 用于限定是否可以发生插入或消失。分数越大，则相似度越高，不允许发生插入和消失。分数越低，则可以发生。若设置为0，则任意大小的插入和消失都是可以存在的。  就是这个数值是用来设置可以产生的gap的大小的。  <https://www.ebi.ac.uk/seqdb/confluence/display/THD/PSI-BLAST> 这里列举了psi-blast工具的一些默认数值。  7，  A consensus sequence C（p）is generated at each position of the profile to aid the display of alignments of target sequences with the profile.  The consensus residue c is the amino acid at P that has the highest score M(p,c)  It is the amino acid most mutationally similar to all the aligned residues of the probe sequences at its position,  rather than merely the most common residue present in the collection of sequences.  一致性序列，对应的每一个都是该行中得分最高的氨基酸。而不是出现次数最高的氨基酸。 |
| 相关论文 | Dayhoff, M. ***0.*(1979)**in ***Arlas of Prorein Sequence undSrrucrure.*** eds. Schwartz, R. M. & Dayhoff, M. 0. (Natl. Biomed. Res. Found., Washington, DC). **Vol. *5 ,* Suppl. 3,** pp. Gribskov, M. & Burgess, **R. R. (1986)*Nucleic Acids Res.* 14, 6745-6763.** |

其实不需要关注它是怎么算的，而是需要关注它能用来干什么。

<http://bioinformatica.upf.edu/T12/MakeProfile.html> 这个网站中包含pssm的介绍。

位置特异性得分矩阵，出现在这个位置上，得分高。用于寻找其他序列中，得分高的：那么这两个序列就非常相近。这个矩阵能表达出这个位置特异性，就是好的。

总结：

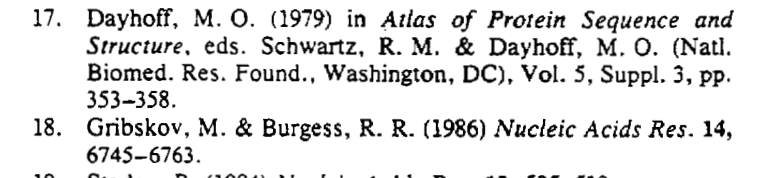
它如果是一条序列，则是突变频率。如果是多条序列，则是突变频率乘以出现的次数。

目的是用这个去寻找存在的相似的序列。那么“突变频率”从哪来？

<http://blog.sina.com.cn/s/blog_5dc95ca6010131vr.html>

Position-specific scoring matrix (PSSM)（特定位点记分矩阵，PSI-BLAST等搜索程序使用）

The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence. Represents the variation found in the columns of an alignment of a set of related sequences. Each subsequent matrix column corresponds to the next column in the alignment and each row corresponds to a particular sequence character (one of four bases in DNA sequences or 20 amino acids in protein sequences). Matrix values are log odds scores obtained by dividing the counts of the residue in the alignment, dividing by the expected number of counts based on sequence composition, and converting the ratio to a log score. The matrix is moved along sequences to find similar regions by adding the matching log odds scores and looking for high values. There is no allowance for gaps. Also called a weight matrix or scoring matrix.



In atlas of protein sequence and structure

<https://proteinstructures.com/Sequence/Sequence/amino-acid-substitution.html>

氨基酸替换矩阵！！！！

介绍blast矩阵的链接。

<https://wenku.baidu.com/view/7a716acca1c7aa00b52acb14.html>