生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(12): 1215~1222
www.pibb.ac.cn

PIBB

蛋白质序列复杂性简化与非比对序列分析 *

李 菁 李逢博 王 炜 **
(南京大学物理系固体微结构物理国家重点实验室，南京 210093)

摘要 非比对序列分析是最新发展的一种序列分析方法，具有计算效率高并适用于分析低相似性的序列，已成功用于 DNA 的序列分析中. 但是由于蛋白质序列的复杂性，非比对序列分析对于蛋白质序列分析的准确度却不高. 用将 20 种天然氨基酸残基归类的方法，简化了蛋白质序列的复杂性，并运用到对蛋白质的非比对序列分析中，有效地提高了序列分析的准确性.

关键词 非比对序列分析，氨基酸残基归类，序列复杂性简化
学科分类号 Q811.4

序列分析是生物学和生物信息学中的重要核心内容. 序列分析是指从核酸和蛋白质的序列出发，得到它们的结构和功能信息，从而了解核酸和蛋白质在生物体中的作用，并研究它们的进化起源. 例如，同源搜索是从目的核酸和蛋白质序列出发，在海量的数据库中搜索与其可能同源的核酸和蛋白质，研究目的核酸和蛋白质的结构和功能. 又例如，序列比对是应用数学原理找出两条或多条序列之间的保守区域，从而分析它们可能保守的重要结构和功能位点. 然而常用的序列比对算法的应用范围却有不少限制. 例如，基于动态规划策略的序列比对算法往往不能很好地运用在分析那些序列相似性较小的蛋白质序列上[1]. 又如，多序列比对的计算复杂度直接取决于进行比对的蛋白质长度和个数，即进行比较的蛋白质序列过多，序列长度过大，序列的复杂性过高都会需要很长的计算时间和很大的计算空间. 另外，随着分子序列数据库中所得到的序列的急剧扩增，进行一一比对的序列分析算法往往不能有效地用于这些海量数据库的查询. 因此，快速而准确又能满足要求的非比对(Alignment-free)算法越来越引起研究者的重视[2~5]. 非比对算法则是将 DNA 或蛋白质的分子序列作为字经过不同组合形成的集合，然后对字的出现率做统计，得到出现率分布的信息，再在出现率向量定义的笛卡尔空间中计算序列间距离[5,6].

据文献报道，如果直接用 20 种天然氨基酸残基所编码的蛋白质序列进行非比对序列分析的准确度不高. 这是因为在字长大于 1 的情况下，将序列转化为字频率矢量的维度过大，使得矢量的很多分量为 0，从而矢量间的距离不能很好地衡量蛋白质之间的亲缘关系[4]. 众所周知，序列是由氨基酸残基字符所编码的，也就是 20 种自然氨基酸残基的不同组合构成了蛋白质序列. 20 种氨基酸残基中某些氨基酸残基具有相似的物理化学特性，在很多情况下，不同的氨基酸残基却实行相同的结构和功能作用. 因此，经过适当的残基归并，能够在保持原蛋白质序列主要信息的同时，有效地降低蛋白质系统的复杂性[7]. 所以，基于蛋白质序列复杂性简化的观点，将 20 种天然氨基酸进行合理的归类，能够大大降低非比对序列分析中字的种类个数，解决了字频率矢量维度过大的问题，从而提高了非序列比对算法对序列分析的效率.

1 数据库

在本文中进行氨基酸残基归类的数据库为 BLOCK 数据库. 该数据库主要是有多个家族中的蛋白质进行无空位序列比对得到的保守性区域(blocks)组成[8]. 为了排除数据库中高相似性的冗余序列，选取的是相似性小于 50%的序列组成的数据子集，也即构建 BLOSUM50 矩阵的数据子集.

*国家自然科学基金资助项目(90403120, 10474041, 10021001).
** 通讯联系人. Tel: 025-83686031, E-mail: wangwei@nju.edu.cn
收稿日期：2006-05-24，接受日期：2006-06-28

在本文中进行非比对序列分析的数据库为SCOP数据库. 为了消除冗余蛋白质序列对辨别蛋白质相互作用的影响，选取的是SCOP数据库(版本1.61)中序列相似性小于40%的代表蛋白质序列所构成的数据子集，即SCOP数据库的子集ASTRAL40. 由于在ASTRAL40数据子集中，有不少蛋白质家族中只包含很少几条，甚至一条序列. 因此在序列分析测试集中排除了那些含有蛋白质序列小于5条序列的蛋白质家族，称之为ASTRAL40-v数据集. 在该数据集中共含有175个家族，1683条序列.

2 方　　法

2.1 氨基酸的归类

首先对蛋白质序列比对数据库BLOCKS中相似性过大的序列进行归并，设置归并百分比为50%，即归并后的子集中每条序列的相似性均小于50%，也即构建替代矩阵BLOSUM50的数据子集. 然后基于该子集，逐步地将20种自然氨基酸归并为N组. 具体地说，如同Henikoff等[8]的工作，通过计算第i组和第j组之间的观测频率$q_{ij}^{(N)}$，和期望概率$e_{ij}^{(N)}$一个$N \times N$的矩阵可以被获得，其中$1 \leqslant i, j \leqslant N$. 这里，$N$代表氨基酸字母总的组数，而每个组内的氨基酸均用同一个字符$G_i^{(N)}$表示. 例如，从20种自然氨基酸将I和V归并到同一组(如第10组)，这样得到的19组简并后的氨基酸，一个有效的字符$G_{10}^{(19)}$则同时代表了残基I和V，且总的组数为N=19. 矩阵中的每个元素描述了替代分值，具体定义为：

$$S_{ij}^{(N)} = \log_2(q_{ij}^{(N)}/e_{ij}^{(N)}) \tag{1}$$

也就是有效字符$G_i^{(N)}$和$G_j^{(N)}$之间的对数概率(logarithmic of odds ratio)，它刻画了这两组简并后氨基酸之间的替代频率. 在矩阵的所有元素中，最大的替代分值假设为字母$G_k^{(N)}$和$G_l^{(N)}$之间的替代分数$S_{kl}^{(N)}$，那么可以认为在所有组中，$G_k^{(N)}$和$G_l^{(N)}$为最佳的替代，也就是这两组可以归并为一个新的组. 同样的，这个新的组可以被分配为一个新的字符，而原来两个组中的所有残基将用新的字符表示. 这样，又可以得到一个新的字符表，它们之间的观测频率和期望概率可以被重新计算. 因此一个新的矩阵$(N-1) \times (N-1)$可以被得到. 这样所有20种氨基酸就可以一步一步地归并起来.

2.2 相互熵分析

根据信息论的定义，替代矩阵(包括未简化的矩阵和简化后的矩阵)中N组残基之间的平均相互信息可以用相互熵来衡量：

$$H^{(N)} = \sum_{i=1}^{N} \sum_{j=1}^{i} q_{ij}^{(N)} s_{ij}^{(N)} \tag{2}$$

相互熵$H^{(N)}$值越大，则替代矩阵中所包含的N组残基之间的平均相互信息越大，反之，$H^{(N)}$越小，替代矩阵中包含的平均信息越小. 经过简化后，N组残基中可能含有的最大信息可以用下式描述：

$$H_{\max}^{(N)} = -\sum_{i=1}^{N} P_i^{(N)} \times \log_2 P_i^{(N)} \tag{3}$$

这里，$P_i^{(N)}$为当20种天然氨基酸残基归类为N组时，有效字符$G_i^{(N)}$在所有蛋白质序列中出现的频率. 因此当N = 20时，即不做任何简化时，$P_i^{(20)}$代表20种天然氨基酸残基在所有蛋白质中的频率，即天然丰度. 而当N < 20时，$P_i^{(N)}$意味着同处于第i组中所有氨基酸残基的天然丰度之和. 因此，$H_{\max}^{(20)}$代表着20种天然氨基酸残基在自然界分布的最大相互熵值.

2.3 序列的字出现率描述

一个长度为n的序列X，可看作是n个字符组成的线性连续集，字符取自长度为r的有限字符集A[9]. 从该序列中抽取一个长度为L的片段定义为字. 所有可能的字将组成一个字集合W_L={$w_{L,1}$，$w_{L,2}$，…，$w_{L,K}$}，集合中有$K = r^L$个元素. 在序列中可以计算字集合中各个字出现的次数$c^X_{L,i}$，并用矢量C^X_L=($c^X_{L,1}$，…，$c^X_{L,K}$)表示. 并且，通过计算各字的相对含量就可得到字的出现率：

$$f_L^X = \frac{c^X_{L,i}}{n-L+1} \tag{4}$$

2.4 矢量距离计算方法

1986年Blaisdell[5]首先运用L-元组进行序列比较，这也是最早的非比对序列比较方法. 他使用马尔科夫链模型，马尔科夫链的转移矩阵等同于所有L-元组的出现率，通过转移矩阵欧几里德距离的平方来量化两个序列之间的差异. 欧几里德距离方法已在生物基因组序列的比较中有成功的应用. 该方法也被用来进行预筛选，通过滤除低相似性的序列来提高数据库的搜索速度，这对于指数增加的序列数据库意义很大. 该方法已在生物信息学中有广泛使用[10]. 下面介绍3种常用的距离计算方法.

2.4.1 欧几里德距离方法

即上面提到的最早用于非比对序列分析的距离计算方法. 序列X和Y欧几里德距离由下式决定：

$$d_L^{eu} = \sqrt{\sum_{i=1}^{K} (c_{L,i}^X - c_{L,i}^Y)^2} \quad (5)$$

这里得到的序列差异值和用序列比对得到的失配量是对应的.

2.4.2 协方差的欧几里德距离. 本方法用协方差与欧几里德距离相组合来计算序列间距离，序列 X 和 Y 之间的协方差距离如下[11]：

$$d_L^{se}(X,Y) = \sum_{i=1}^{K} \frac{(c_{L,i}^X - c_{L,j}^Y)}{S_{i,i}} \quad (6)$$

该方法在字的交迭量的计算上非常有效，因为字的交迭，反映了序列中字的周期性，而具有重复 motifs 的字更容易一起出现，使得协方差结构发生变化.

2.4.3 信息论距离法. 两个序列 X 和 Y 可以看作两个离散概率分布 p 和 q，它们之间的差异可用相互熵来表示，Kullback-Leibler(KL)偏差是其一种表示方法[12, 13]，用该方法表示的两个序列距离为：

$$d_L^{ku}(X,Y) = \sum_{i=1}^{k} f_{L,i}^X \cdot \log_2(\frac{f_{L,i}^X}{f_{L,i}^Y}) \quad (7)$$

2.5 同源搜索

在序列分析中，常用同源搜索的方法来评价分析的准确度. 同源搜索就是在序列数据库中搜索同源序列，即属于同一个家族的序列对. 具体地说，就是对 ASTRAL40-v 中所有的 1 683 条序列进行一一的距离计算(all-against-all)，可以得到 1 683×1 683 = 2 832 489 个距离值. 将蛋白质序列对按照距离值进行排序，距离越小的序列对之间的关系越近，可以被认为具有同源关系.

在同源搜索中，通过对 ROC 曲线的分析来评价和比较用 Alignment-free 算法进行蛋白质序列分析的准确度[14]. 即，如果两个蛋白质序列被程序作为同源序列，则它同源关系为阳性(positive). 如果在 ASTRAL40-v 数据库中它们确实属于同一家族，则为真阳性(true positive, TP)，反之则为假阳性(flase positive，FP). 同样的，两个蛋白质序列被判定为非同源，即同源关系为阴性 (negative)，如果在 ASTRAL40-v 中，这两个蛋白质确实不属于同一家族，则为真阴性 (true negative，TN)，反之则为假阴性(true negative，FN). 因此，同源搜索的敏感性和专一性则可有下式表示：

$$Sensitivity = \frac{TP}{TP + FN} \quad (8)$$

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

而

$$1-Specificity = \frac{FP}{TN + FP} \quad (10)$$

在同源搜索中可以设置不同的阈值，即小于某个距离值相对应的序列对被判定为同源. 这时就会有不同的 TP，FP，TN，FP 值，也就会有不同的敏感性和专一性. 因此，ROC 曲线就是在各种不同阈值条件下 *Sensitivity* 与(1−*Specificity*) 的曲线. 而曲线下的面积则定义为 AUC(area under a ROC curve). 在归类算法中，AUC 是衡量归类质量的一个非常重要的参数[15]. 如果所有的归类结果完全正确，则 AUC 为 1，而如果归类结果为随机归类，则 AUC 大致为 0.5[16]. 在本文工作中，也主要是用 AUC 参数来衡量归类的准确程度. 因此对蛋白质中字出现率的矢量后，进行非比对序列分析的归类过程中，AUC 值越大，则对基于蛋白质序列信息的归类准确度越高，也就是对蛋白质亲疏远近关系的分辨准确度越高. 所以通过 AUC 值在简化蛋白质序列的非比对序列分析的变化，不但可以得到蛋白质序列信息与字出现率的关系，还可以得到序列简化过程中序列信息保持的程度.

需要指出的是，用简化的蛋白质序列进行非比对序列分析，即将蛋白质序列用氨基酸残基归类后的字符代替原 20 种天然氨基酸字符，进行序列的简化即可进行非比对的序列分析.

3 结果与讨论

3.1 氨基酸归并结果

用我们的方法，可以将 20 种自然氨基酸残基按照不同的简化程度，逐步地归并到不同的组中. 这里对 BLOCKS 数据库序列相似性小于 50% 的序列组成 0 的子集，计算不同字符个数 N 时的替代频率，从而将氨基酸进行归并. 图 1 显示了归类结果. 从图 1 中可以清楚地显示，一些具有明显的物理化学或立体化学相似性的氨基酸被归类在一起，如(I，V)，(Q，E)，(R，K)，(N，D)，和(W，F，Y). 并且最终所有 20 种氨基酸可以归并为疏水氨基酸和极性氨基酸两大类. 这个归类结果大致与我们以前的归类结果和其他文献报道的结果相一致[7, 17]. 但归类树的形式更能直观形象表示出各种氨基酸不同的相似性距离.

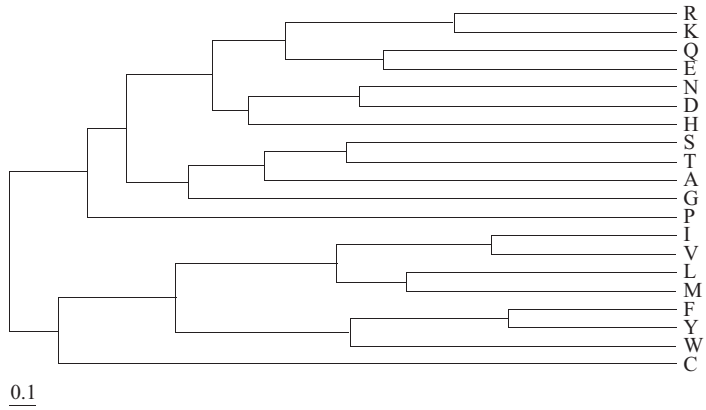

Fig. 1 The tree-like distribution of residues based on substitution scores for subset from BLOCKS database with clustering percentage setting as 50%

对于归类树中各个氨基酸归类的远近距离是由信息有序度所刻画的，即 $D^{(N)}=H^{(N)}_{\max}-H^{(N)}$. 这里 $D^{(1)}=0$，也就是当所有 20 种氨基酸归为同一大组时，它们之间的距离为 0，设定为数的根部. 而当不同分组条件下的简化程度的差别，即不同组数 N 和 N' 之间的距离为 $D^{(N)}=H^{(N)}_{\max}-H^{(N)}$，此式刻画了两种不同分组的信息丢失，也就是说这个距离越短，则两个分组之间的信息丢失越小.

3.2 氨基酸归类相互熵的变化

正如在构建氨基酸归并树的过程中所描述的，可以用氨基酸归并后的信息丢失，即相互熵的变化，来刻画每次归并后的距离. 因此，观测替代矩阵的相互熵随着字母数减少的变化能直接地观测与原 20 种氨基酸比较，不同氨基酸分组带来的信息变化. 从图 2 中我们可以看出，当 $N\geqslant 9$ 时，相互熵 $H^{(N)}$曲线为一个平台，而之后随着归类水平的减少相互熵 $H^{(N)}$下降. 这个存在的平台暗示着当 $N\geqslant 9$ 时，归类后的氨基酸组能保持足够的替代关系，并且相应于 20 个字母时只有很少的信息被丢失. 这个结论与我们以前所研究的结果相一致[17, 18].

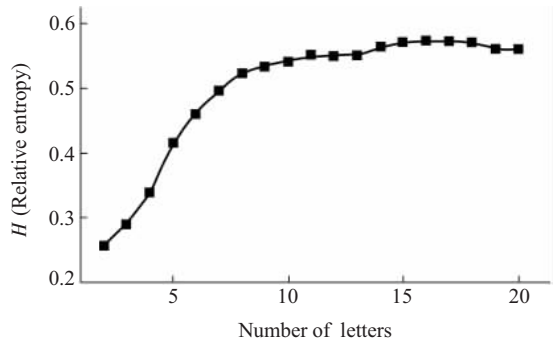

Fig. 2 The relative entropy H versus the number of letters N in reduced alphabets

3.3 对非简化序列进行同源搜索

首先，研究的是用 Alignment-free 算法对非简化的蛋白质序列，即由 20 种天然氨基酸残基编码的蛋白质序列进行归类分析. 为了研究不同算法对蛋白质序列归类准确程度的影响，采用了 3 种不同的算法，即欧几里德距离算法(Euclidean，Eu)，协方差欧几里德算法(Standard Euclidean，Se)，信息论距离法(Kullaback-Leibler，Ku). 图 3 给出的是用最基本的欧几里德距离算法对 ASTRAL-40v 数据子集中的 1 683 条蛋白质序列进行同源搜索得到的 ROC 曲线.

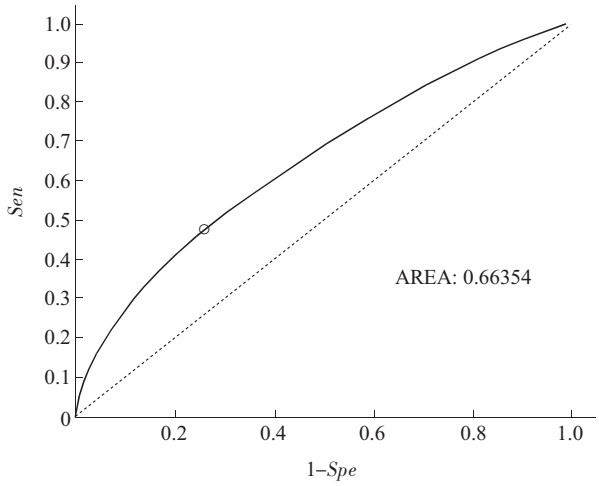

Fig. 3 The ROC curve of homology detection using Euclidean algorithm

The protein sequences for detection are encoded by 20 kinds of natural occurring residues. The AUC is the area under the ROC curve.

将 20 种天然氨基酸残基字符编码的蛋白质序列转换为不同的“字”出现频率 (或个数) 所编码

的矢量，能够反映出原来蛋白质序列的某些物理化学信息，如当字长为 1 时，反映原蛋白质序列的长度，所含氨基次残基的种类和丰度. 当字长大于 1 时，还反映了原蛋白质的部分序列信息. 然而，在实践计算并不是字长越大越好，反而是字长取得越大，准确度越小. 图 4 是运用 3 种不同的算法，在家族层次上，取用字长分别为 1，2，3 时进行同源搜索的准确程度. 从图 4 中发现，对于各种不同的算法，字长取越大，同源搜索准确度越小. 这是因为当不在氨基酸残基简化时，字长为 1，每个矢量(即序列)的分量个数为 20^1=20 个，而当字长为 2 时，每个矢量的分量个数为 20^2=400 个，当字长为 3 时，每个矢量的分量个数为 20^3 = 8 000 个. 这两种字长所得到矢量维数都要大大超过 ASTRAL-40v 中的蛋白质平均长度，175 个残基，从而字出现频率(或个数)所编码的矢量为稀疏矢量，反而不能很好地刻画蛋白质的特征. 因此，在用非比对算法对序列长度较长的核酸序列进行序列分析时，可以采用较长的字长，而对序列长度不长的蛋白质序列进行非比对的序列分析时，如果不作氨基酸残基种类归类的序列简化的话，字长取 1 反而最为合适. 很多文献报道也证实了这个结论[1, 9, 19].

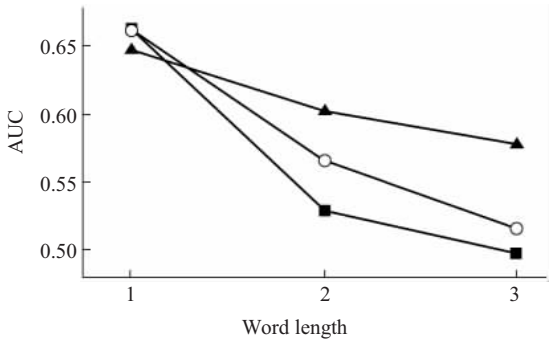

Fig. 4 The comparison of *AUC* value of three different methods: Euclidean (Eu), Standard Euclidean (Se), Kullback-Leibler (Ku), which are obtained at different word length: 1, 2, and 3

■—■: Eu; ○—○: Se; ▲—▲: Ku.

3.4 序列简化后的非比对序列分析

用非比对算法对未经简化的蛋白质序列，即由 20 种天然氨基酸残基编码的蛋白质序列进行分析中，可以看出，不同的字长进行编码，不同的算法计算矢量间的距离，以及在不同的结构层次上进行序列分析的准确度都是不同的. 这其中的主要原因之一便是将字符组成的蛋白质序列转换为由字频率组成的矢量，造成的序列信息的冗余或者丢失. 例如在前面的描述中，字长为 2，3 时用非比对算法进行序列分析的准确度反而不如字长为 1 的时候，这主要是因为字频率所形成的可能维度空间太大，由序列转换而来的矢量中大部分分量都为 0 的缘故. 并且，字频率所形成的可能维度空间过大，还会造成计算空间、时间的巨大花费. 因此，基于简化蛋白质序列复杂性的原理，将 20 种天然氨基酸残基进行合理归类，简并字符，从而在指数降低字频率维度空间的同时，大大降低了进行序列分析所需的计算空间和计算时间. 但是在用非比对算法进行序列分析的过程中，还需要关心的是随着蛋白质序列的简化，序列分析的准确度是否能够保持甚至提高呢？下面，用不同的字长，不同的距离计算方法和不同的氨基酸残基简化程度，来比较序列同源搜索的准确程度.

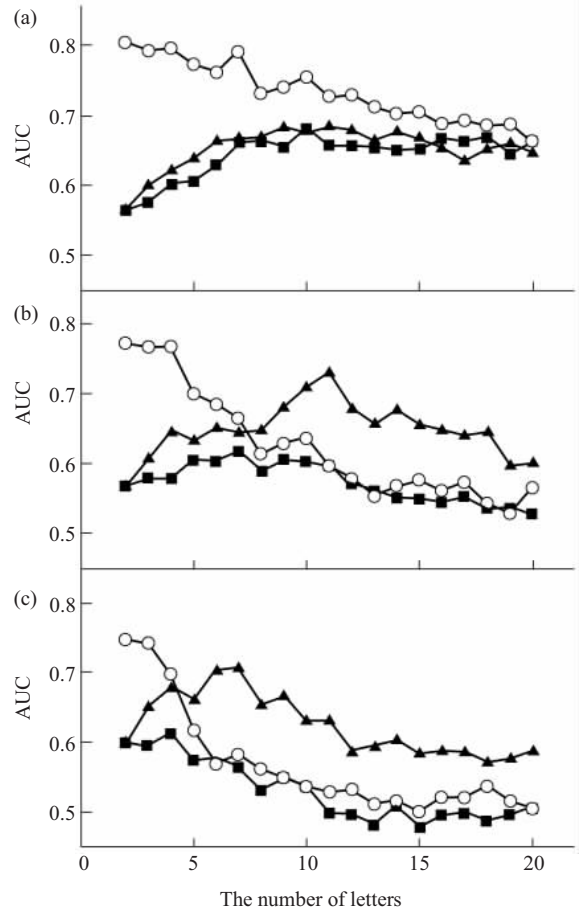

Fig. 5 AUC values versus the number of letters *N* in reduced alphabets

Detailedly, the plots in figure are obtained by three kinds of word length: (a) *L*=1, (b) *L*=2, (c)*L*=3, and three algorithms: Euclidean (Eu), Standard Euclidean (Se), Kullback-Leibler (Ku). ■—■: Eu; ○—○: Se; ▲—▲: Ku.

图 5 是在家族层次上，不同简化程度上进行同源搜索的结果，其中图 5a～c 分别为字长为 1，2，和 3 时同源搜索的结果.

当字长为 1 时，从图 5a 可以看出，用欧几里德距离法和信息论距离法进行同源搜索的结果相似，即在 $9 \leqslant N \leqslant 20$ 的范围内，这两种方法在不同序列简化程度进行同源搜索的 AUC 值基本保持在 0.65 左右，而当 $N < 9$ 时，AUC 值则随着字符个数的减少而降低. 这个结果也正好与相互熵随着氨基酸残基归类的变化规律 (图 2) 相吻合. 这说明在字长为 1 的条件下，当蛋白质序列简化到一定程度时，即 $N > 9$ 时能保持原蛋白质序列信息. 因此用欧几里德距离法和信息论距离法进行同源搜索的准确率能保持与字符 $N = 20$ 时相一致. 与这两种距离计算方法不同，协方差欧几里德距离法随着 N 的减少搜索的准确度反而提高，即当 $N = 2$ 时 (分为疏水和极性氨基酸残基两大类)，同源搜索的准确度最高，超过了 0.8. 这是一个相当高的 AUC 值，要超过文献报道中，进行比对的 Smith-Waterman 算法能够得到的 AUC 值[19]. 我们分析，这是因为协方差欧几里德距离法引入了协方差因子，即综合了矢量中各分量的空间距离和它们在不同序列中的协方差. 随着蛋白质序列的简化，即具有相似物理化学氨基酸残基被归并为一组，具有同源关系的序列间各个字的协方差将变小，而非同源序列间各个字的协方差将扩大，因此氨基酸残基字符 N 越小，越能分辨序列的同源关系，同源搜索的准确率也越高.

当字长为 2 时，从图 5b 可以看出，用欧几里德距离法进行同源搜索的准确度相比较，字长为 1 时的准确度有所下降，在 0.55～0.6 左右，且搜索的准确度基本上不受氨基酸残基简化程度(N)的影响. 而协方差欧几里德距离法进行同源搜索的准确度相比较，字长为 1 时相类似，即氨基酸残基字符 N 越小准确度越高. 在图 5b 中，与图 5a 差异最大的就是用信息论算距离法计算的曲线，即用该法在氨基酸残基字符个数 N=11 时，准确度 AUC 达到最大值 0.75，这同样是一个接近用比对的 Smith-Waterman 算法能够得到的较高准确度.

当字长为 3 时，从图 5b 可以看出，用欧几里德距离法和协方差欧几里德距离法所得到曲线与字长为 2 时差不多，而用信息论算距离法进行同样得到的最大准确度在氨基酸残基字符个数 $N = 5$ 达到最大值.

从上面的分析中可以看出，用欧几里德距离法进行同源搜索的准确度受序列简化的影响较小，这可能是因为其本身的准确度就较差的缘故. 而协方差欧几里德距离法随着 N 的减少搜索的准确度反而提高，即当 $N = 2$ 时，同源搜索的准确度最高，而当 $N = 20$ 时，同源搜索的准确度反而最低. 与上面两种方法不同的是，用信息论距离法在合适的字长与合适的 N 值时($L = 2$，$N \approx 10$ 和 $L = 3$，$N \approx 5$)，能够达到最大的 AUC 值. 即矢量的维度为 100 左右，大约为蛋白质序列的平均长度时，进行同源搜索的准确度最高.

4 结　论

非序列比对(Alignment-free)是近几年发展非常迅速的一种序列分析方法，它有着明确的理论体系，并且有很多成功应用于序列分析的实际例子[20~22]. 相比较于传统的基于动态规划策略的比对序列分析算法，非比对序列分析方法的最大优点就是大大增加了计算效率. 然而，在实际应用不同的距离计算方法对 ASTRAL-40v 数据库进行同源中却发现，当用 20 个天然氨基酸残基字符编码的蛋白质序列转换为字频率矢量时，序列分析的准确度却不高，且字长取的越长，准确度却越低. 经过分析，这主要是因为“字”频率所形成的维度空间太大，字频率矢量的大部分分量都为 0，从而矢量间的距离不能有效衡量蛋白质之间的远近关系的缘故. 并且，字频率所形成的可能维度空间过大，还会造成计算空间、时间的巨大花费.

在应用不同的距离计算方法(欧几里德距离法，协方差欧几里德距离法，信息论距离法)，不同的字长 (L=1，2，3) 条件下研究归类后的残基字符个数 N ($2 \leqslant N \leqslant 20$) 对序列分析准确度的影响. 在分析中，不同的距离计算方法对字长和序列简化程度的敏感度是不一样的. 具体来说，欧几里德距离法受字长 L 和残基字符个数 N 的影响较少，进行序列分析的准确度也较低. 然而，运用协方差的欧几里德距离法进行非序列分析的准确度则随着残基字符个数 N 减少而提高，在 L=1，N=2 时达到最大值. 而运用信息论距离法进行非序列分析的准确度在选取合适的字长和残基个数时达到最大值，如 L=2，$N \approx 10$ 或 L=3，$N \approx 5$. 这两种方法进行序列分析的准确度接近和达到了比对的 Smith-Waterman 算法同样的准确度. 因此可以认为根据 20 种天然氨基酸中某些氨基酸残基相似的物理化学特性进行归类，

在保持序列信息的同时有效的简化的蛋白质序列的复杂性，从而非比对序列分析中代表蛋白质序列的字频率矢量的维度空间，不但大大地降低了进行序列分析所需的计算空间和计算时间，而且提高了某些距离计算方法的有效度，从而在提高了计算时间的同时，保持了高的准确度.

在本文的工作中，给出的是对 SCOP 数据库中 ASTRAL40-v 数据集的非比对序列分析结果. SCOP 数据库是对所有已知三维结构的蛋白质进行分类的二级数据库，而 ASTRAL40-v 数据集则是收集了 SCOP 数据库中序列相似性较低的序列子集，因此本工作对 ASTRAL40-v 数据集进行非比对序列分析的结果具有普遍性的意义. 我们也对其他序列分类数据库，如 CATH 数据库中的序列进行简化并进行非比对序列分析，也得出了相类似的结论，即对序列进行合理的简化并且选取适当的字长能够得到最佳的归类准确度. 当然，最佳的氨基酸字符个数 N 和最佳的字长 L 限于不同数据库中序列规模的限制略有不同.

非比对序列分析方法的研究发展得很快，将蛋白质序列中的信息转换为矢量的新编码方法和距离计算方法在不断提出[23, 24]. 从我们的分析中也发现，由于蛋白质序列的冗余性，各种非比对序列分析在序列未经简化的条件下都达到最好的准确度. 而根据氨基酸残基的相似性，进行合理归类，是简化蛋白质序列复杂性的有效方法. 而不同的氨基酸简化程度可能对不同的编码方式和不同的距离计算方法产生不同的影响. 所以氨基酸残基归类和蛋白质序列简化的思想，也是提高非比对序列分析方法性能的有效方法.

参 考 文 献

1 Borosy A P, Balogh B, Matyus M. Alignment-free descriptors for quantitative structure-rate constant relationships of [4+2] cycloadditions. J Mol Struc-Theochem, 2005, **729** (3): 169～176

2 Pham T D, Zuegg J. A probabilistic measure for alignment-free sequence comparison. Bioinformatics, 2004, **20** (18): 3455～3461

3 Mantaci S, Restivo A, Rosone G, *et al*. A new combinatorial approach to sequence comparison. Lect Notes Comput Sc, 2005, **3701** (4): 348～359

4 Vinga S, Almeida J. Alignment-free sequence comparison - a review. Bioinformatics, 2003, **19** (4): 513～523

5 Blaisdell B E. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci USA, 1986, **83** (14): 5155～5159

6 Pevzner P A. Statistical distance between texts and filtration methods in sequence comparison. Bioinformatics, 1992, **8** (2): 121～127

7 Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nature Struct Biol, 1999, **6** (11): 1033～1038

8 Henikoff S, Henikoff J G, Alford W J, *et al*. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene, 1995, **163** (1): GC17～GC26

9 Fechner U, Franke L, Renner S, *et al*. Comparison of correlation vector methods for ligand-based similarity searching. J Comput Aid Mol Des, 2003, **17** (10): 687～698

10 Lapinsh M, Gutcaits A, Prusis P, *et al*. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci, 2002, **11** (4): 795～805

11 Wu T J, Burke J P, Davison D B. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics, 1997, **53** (4): 1431～1439

12 Li M, Badger J H, Chen X, *et al*. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics, 2001, **17** (2): 149～154

13 Chen X, Kwong S, Li M. A compression algorithm for DNA sequences. IEEE Eng Med Biol Mag, 2001, **20** (4): 61～66

14 Bradley A P. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn, 1997, **30** (7): 1145～1159

15 Green R E, Brenner S E. Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE, 2002, **90** (12): 1834～1847

16 Baldi P, Brunak S, Chauvin Y, *et al*. Assessing the accuracy of prediction algorithms for classification. An overview. Bioinformatics, 2000, **16** (5): 412～424

17 Li T, Fan K, Wang J, *et al*. Reduction of protein sequence complexity by residue grouping. Protein Eng, 2003, **16** (5): 323～330

18 Fan K, Wang W. What is the minimum number of letters required to fold a protein?. J Mol Biol, 2003, **328** (4): 921～926

19 Vinga S, Gouveia-Oliveira R, Almeida J S. Comparative evaluation of word composition distances for the recognition of SCOP relationships. Bioinformatics, 2004, **20** (2): 206～215

20 Christoffels A, van Gelder A, Greyling G, *et al*. Sequence Tag alignment and consensus knowledgebase. Nucleic Acids Res, 2001, **29** (1): 234～238

21 Zhu J, Liu J S, Lawrence C E. Bayesian adaptive sequence alignment algorithms. Bioinformatics, 1998, **14** (1): 25～39

22 Almeida J S, Vinga S. Universal sequence map (USM) of arbitrary discrete sequences. Bmc Bioinformatics, 2002, **3** (1): 11～16

23 Hirons L, Holliday J D, Jelfs S P, *et al*. Use of the R-group descriptor for alignment-free QSAR. Qsar Comb Sci, 2005, **24** (5): 611～619

24 Carpenter J E, Christoffels A, Weinbach Y, *et al*. Assessment of the parallelization approach of d2_cluster for high-performance sequence clustering. J Comput Chem, 2002, **23** (7): 755～757

Simplification of Protein Sequence and Alignment-free Sequence Analysis*

LI Jing, LI Feng-Bo, WANG Wei**

(*National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing* 210093, China)

Abstract Alignment-free comparison is a recently developed method for sequence alignment, which has high computational efficiency and suitable to the low identical sequences. Alignment-free comparison was successfully applied in the DNA analysis. However, the accuracy of analysis is not high when it was applied in protein analysis because the complexity of protein is larger than DNA by consisting of 20 types of residues. Thus, residues are clustered into a few groups based on their similarity of physicochemical features. Using such simplified alphabets, the complexity of protein sequences is reduced and at the same time the key information encoded in the sequences remains. Therefore, the accuracy of alignment-free comparison is improved.

Key words alignment-free comparison, grouping of amino acids, simplification of protein sequence

---

*This work was supported by grants from The National Natural Science Foundation of China (90403120, 10474041, 10021001).

**Corresponding author . Tel: 86-25-83686031, E-mail: wangwei@nju.edu.cn

Received: May 24, 2006 Accepted: June 28, 2006