氨基酸残基归类及用简化后的字符识别蛋白质结构保守区域

李 菁[①] 王 炜[①②*]

(① 南京大学固体微结构实验室, 南京 210093; ② 中国科学院理论物理研究所, 北京 100080)

摘要 序列比对是寻找蛋白质结构保守性区域的常用方法, 然而当序列相似小于 30%时比对准确度却不高, 这是因为在这些序列中具有相似结构功能的不同残基在序列比对中往往被错误配对. 基于相似的物理化学性质, 某些残基可以被归类为一组, 而应用这些简化后的残基字符可以有效地简化蛋白质序列的复杂性并保持序列的主要信息. 因此, 如果 20 种天然氨基酸残基能够正确的归类, 可以有效地提高序列比对的准确度. 本文基于蛋白质结构比对数据库 DAPS, 提出了一种新的氨基酸残基归类方法, 并可以同时得到不同简化程度下的替代矩阵用于序列比对. 归类的合理性由相互熵方法确认, 并且应用简化后的字符表于序列比对来识别蛋白质的结构保守区域. 结果表明, 当氨基酸残基字符简化到 9 个左右时能够有效地提高序列比对的准确度.

关键词 氨基酸归类 结构误别 序列对比

蛋白质序列比对是一种寻找蛋白质结构相似性的通用方法 [1,2]. 当蛋白质的序列相似性高于 30%时, 这些蛋白质通常也具有很高的结构相似性. 因此, 对于结构未知的蛋白质, 可以通过序列比对来识别它们之间的结构保守区域. 然而, 这种方法对于序列相似性低于 30%的蛋白质来说却不适用, 这是因为对于这些蛋白质来说序列比对的结果与结构比对的结果往往不同. 所以对很多序列上差异较大却具有相同结构或保守结构区域的蛋白质, 它们的结构保守区域往往不能用序列比对的方式来识别. 是否能够通过增加序列比对对这些低序列相似性蛋白质进行保守结构区域的识别能力?

以前很多研究报道了有些氨基酸残基具有相同的物理化学特性, 可以适当地根据它们在蛋白质中相似的结构和功能作用进行归类 [3~6]. 经过归类以后, 蛋白质序列各个位点中的氨基酸的相似性得到了提高. 以蛋白质系统的复杂性简化观点来看, 序列的一致性也增加了. 因此, 将氨基酸残基适当的归类有利于提高序列比对识别蛋白质之间的结构保守区域的能力. 当然, 对氨基酸残基归类结果的好坏将决定对

收稿日期: 2006-05-23; 接受日期: 2006-07-31
* 联系人, E-mail: wangwei@nju.edu.cn

蛋白结构保守区域识别的准确与否.

氨基酸残基归类的基础是氨基酸替代矩阵, 而常用的替代矩阵是对具有高序列相似性的蛋白质进行序列比对而构建的. 例如, BLOSUM矩阵是基于BLOCKS数据库构建而得的[7], 该数据库中的多数序列具有高相似性的, 因此, 由高相似性序列比对而构建的替代矩阵不能很好地描述低相似性序列中氨基酸残基之间替代关系[8,9], 所以本文基于结构比对数据库(database of aligned protein structures, DAPS)构建替代矩阵. DAPS数据库是将具有结构相似性的蛋白质进行结构比对而构建得到的, 包括了大量的低序列相似性蛋白质, 基于这样一个结构比对数据库而构建的替代矩阵能够更好地反映氨基酸残基之间的结构替代关系.

本研究中, 首先从 DAPS 数据库中构建替代矩阵, 再根据替代矩阵中的替代分值将 20 种天然氨基酸归类为 N 组. 通过归类后的各种结果分析, 发现当20种氨基酸残基归类到7组左右时, 用简化的序列保持了原来序列的大部分信息. 因此, 当 N=6~9 时, 序列比对对低序列相似性蛋白质进行保守结构区域的识别能力也得到了提高, 而对 $N\approx$20(或 N<6)时, 序列比对的准确度却不能得到提高, 这是因为序列的复杂性没有得到有效简化(或丢失了序列的主要信息).

1　方法

1.1　数据库

DAPS数据库是一个蛋白质结构比对数据库, 其详细描述可以参见网页: http://www.doe-mbi.ucla.edu/~parag/DAPS/, 该数据库是对具有不同序列相似性的蛋白质进行结构比对的数据库, 构建在FSSP, DSSP, PDB和CATH等结构数据库基础之上. 需要指出的是, 在DAPS数据库中, 蛋白质两端的结构区域如果不能被叠加的话, 在结构比对结果中将被删除, 而中间结构区域不能被叠加的话, 则在结构比对结构中作为插入或缺失的空格位点[10~12].

DAPS 数据库包含了不同序列相似性的蛋白质对, 这些的蛋白质对可以按照序列相似性进行划分, 可划分为 10 个子集 S_1, S_2, …, S_{10}. 具体的说, 子集 S_1 中所含的蛋白质对的序列相似性都不超过 10%, 也就是 S≤10%, 同样, 子集 S_2 中所含的蛋白质对的序列相似性都不超过 20%, 即 S≤20%, 以此类推. 因此不同的子集包含了不同序列相似性的蛋白质对, 统计每个子集中所有蛋白质对的结构等价位点, 可以得到各种氨基酸残基类型的替代频率, 从而构建出氨基酸替代矩阵.

1.2　基于替代矩阵将氨基酸残基归类

对于 DAPS 中的每一个子集, 如 S_3, 可以用下面的方法逐步地将 20 种天然氨基酸归类为任意 N 类. 即通过计算第 i 组和第 j 组残基字符之间的观测频率 $q_{ij}^{(N)}$ 和期望概率 $e_{ij}^{(N)}$, 可以得到一个 $N\times N$ 的残基替代矩阵($1\leqslant i, j\leqslant N$). 这里, N 代表氨基酸字符总的组数, 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 $S_{ij}^{(N)}=\log_2(q_{ij}^{(N)}/e_{ij}^{(N)})$, 也就是有效字符 $G_i^{(N)}$ 和 $G_j^{(N)}$ 之间的对数概率, 它刻画了这两组简并后的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重新计算, 因此可以得到一个新的矩阵$(N-1)\times(N-1)$, 这样所有 20 种氨基酸就可以一步一步的归并起来.

1.3　相关熵分析

根据信息论的定义, 替代矩阵中N组残基之间的平均相互信息可以用相互熵 $H^{(N)}=\sum_{i=1}^{N}\sum_{j=1}^{i}q_{ij}^{(N)}s_{ij}^{(N)}$ 来衡量[13,14], 其中 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 $H_{\max}^{(B)} = -\sum_{i=1}^{N} P_i^{(N)} \times \log_2 P_i^{(N)}$ 来描述. 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，$P_i^{(20)}$ 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信息熵值.

1.4 简化后的氨基酸字符进行序列比对

进行序列比对的程序是FASTA程序包中的ALIGN程序. 在序列比对中起始空位罚分为–11，延伸的空位罚分为–1，其他的参数采用默认的参数. 进行序列比对的蛋白质序列均取自于DAPS数据库. 当不进行氨基酸残基类型简化时，蛋白质序列不作任何变化，而当氨基酸残基类型简化为N组时(N<20)，蛋白质序列用相应简化的N个字符代替，而替代矩阵也采用于相应的$N\times N$的矩阵. 作为对比，还采用了常用的程序BLAST[15]和ClustalX[16]]对于不经过简化的蛋白质序列进行比对，与FASTA程序进行结果比较.

1.5 结构比对

结构比对是通过将两个蛋白质的三维结构在空间中平移和旋转，进行叠加找到它们的空间等价位点，从而找到这两个蛋白质的相似或相同的结构保守区域. 本研究工作中，蛋白质结构比对的结果是直接从DAPS数据库中提取的. DAPS数据库中的结构比对结果大部分是从FSSP数据库中由DALI程序计算而得 [17]，而在DALI程序中，采用弹性相似性得分，结构相似性阈值设置为 20%，其他的结构比对参数为默认参数 [18].

1.6 比较序列比对与结构比对

对一对蛋白质来说，如果序列比对所找出的保守位点正好与结构比对找到的空间等价位点相吻合，可以说序列比对能够很好识别结构保守性区域 [19]. 因此，结构比对所识别的空间等价位点作为参照位点，而序列比对所识别的保守位点将与这些参照位点相比较，如果序列比对所识别的保守位点正好是结构比对中的空间等价则认为此位点为正确识别位点，反之则为错误识别位点. 我们用两个参数来衡量序列比对与结构比对的吻合程度：C_R=$N_{correct}$ /N_{str}和D_R=$N_{correct}$ /N_{seq}. 这里$N_{correct}$是序列比对正确识别位点的个数，即序列比对和结构比对同时识别的位点个数，N_{seq}是在序列比对中排列在同一列的位点个数，N_{str}是在结构比对中空间等价位点个数. 序列比对的准确率就由参照结构比对结果而来的C_R和D_R共同刻画.

1.7 主成分分析

替代矩阵中 20 种天然氨基酸之间的相互替代分值可以看作在二十维空间中的 20 个矢量，因此这 20 个矢量间的距离矩阵可以由 $M_{ij} = \left(\sum_{k=1}^{20} (s_{ik} - s_{jk})^2 \right)^{1/2}$ 来计算，而这些距离矩阵的相互系数矩阵则可以由 $R = MM^T$ 来计算. 通过提取上述相互系数矩阵得到的两个最大特征值所对应的特征矢量，则可以将二十维空间中的矢量投影到二维空间上，即用 2 个分量的矢量最大程度的刻画了原矢量的特征 [20,21].

2 结果

2.1 从 DAPS 数据库中将氨基酸残基归类的结果

应用前面所描述的方法，可以将 20 种氨基酸残基(或 20 个字符)逐步归类为N组(2≤N<20). 为了分析不同序列相似性下氨基酸残基分类的结果，DAPS数据库中进行结构比对的蛋白质对按照不同的序列相似性分成 10 个子集. 图 1 给出了从子集S_3和S_5进行氨基酸残基归类的结果，由归类树的形式表示. 从不同序列相似性的子集得到的归类结果树中可以看出很多相同之处：如20种氨基酸残基最终都归类为2类，即疏水残基和极性残基，并且一些氨基酸残基表现了较稳定的替代关系. 这些氨基酸归类结果与以前研究的结果基本相似 [22].

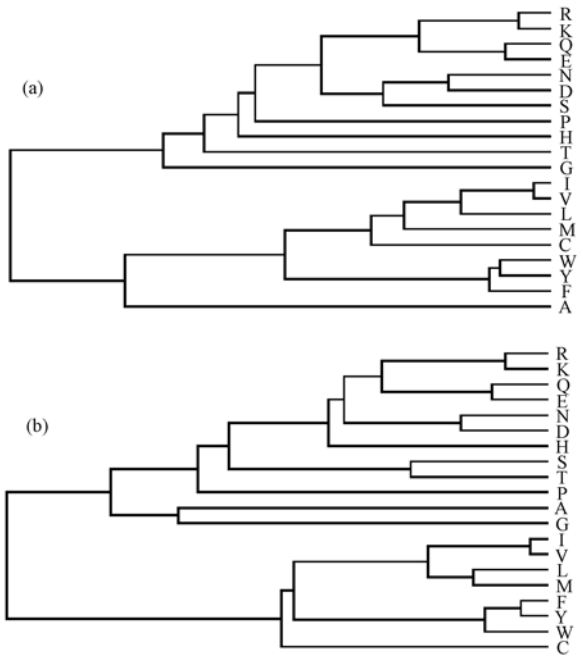


图 1　用基于替代分值的方法对 DAPS 数据库中的子集进行氨基酸残基归类的结果

(a)和(b)分别为对 S_3 和 S_5 子集进行归类的结果. 在归类树中的字符代表着 20 种天然氨基酸残基, 而归类树中组与组之间距离代表着 2 种不同归类情况下信息的丢失程度

在氨基酸归类树中, 20 种天然氨基酸的归类顺序由信息的有序度 $D^{(N)} = H_{\max}^{(N)} - H^{(N)}$ 来表示, 其中, $D^{(1)} = 0$ 时, 即所有的 20 种天然氨基酸归类为同一大组时, 作为归类树的根; 而两种不同分组的 N 和 N' 之间的距离由 $L^{(N',N)} = D^{(N')} - D^{(N)}$ 表示, 反应了两种不同简化程度的信息丢失. 因此如图 1 中所示, 两种分组之间的距离越短, 说明了这两种不同简化程度的分组信息丢失的越小.

图 2 为对氨基酸残基归类的相互熵分析. 对于从序列相似性较高的数据集S_5, 可以从图 2(a)中看出相互熵 $H^{(N)}$ 在$N \geqslant 9$ 时具有一个平台, 即当氨基酸残基简化后的组数大于 9 时, 数值基本上相等. 当氨基酸残基简化后的组数小于 9 时, 它们的相互熵急剧减少. 对于序列相似性较低的数据集, 如S_3, 可以从图 2(a)中看出对于不同简化程度时相互熵值均在 0.1 左右, 即 $H^{(N)} \approx 0.1$. 对于不同的数据子集, 20 种氨基酸残基所具有的相互熵 $H^{(20)}$ 并不相同. 在相互熵随氨基酸残基归类的变化曲线中所存在的平台, 意味着当氨基酸残基种类归类到某种程度之前, 简化后的字符所代替的归类后的氨基酸残基仍然能够很好地反映出原来 20 种天然氨基酸残基的相互关系, 即氨基酸残基简化只是丢失了很少的相互信息. 当然这种相互信息随着氨基酸简化程度的丢失情况, 蛋白质之间的序列相似性不同而不同. 这种相互熵随着氨基酸残基种类归类保持不变或提高的特性, 从相互熵与归类过程中的信息有序度变化的相关曲线能够更好的表现(图 2(b)). 对于不同相似性的蛋白质序列, 用适当归类后的 9 个字符能够保持原 20 个字符编码的序列中的大部分信息, 这个结论与我们以前的研究结果相一致 [23].

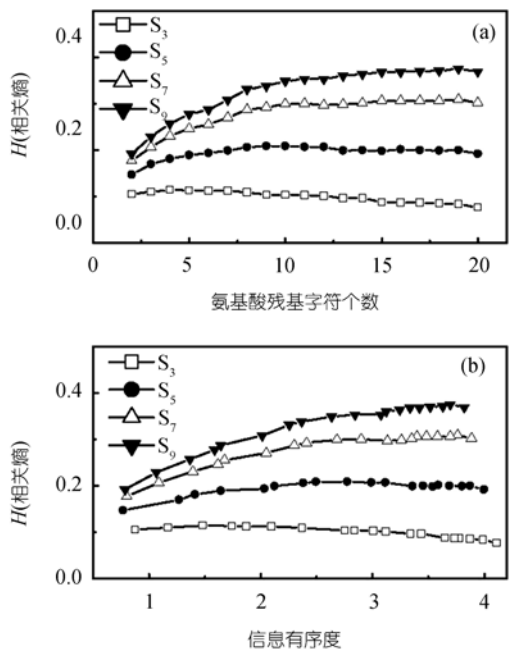


图 2

(a)为相关熵 H 与氨基酸残基字符个数的变化曲线, (b)为信息有序度与氨基酸残基字符个数的变化曲线. 在(a)和(b)中分别给出了 DAPS 数据库中 S_3, S_5 和 S_7 三个子集的相关熵和信息有序度的结果

2.2　序列比对识别蛋白质的结构保守区域

通常来说, 对于序列相似性较低的那些蛋白质对, 如 S<30%, 序列比对与结构比对的结果差异较大, 即序列上找出的保守位点并不是结构上的空间等价位点, 这是因为序列的相似性是由氨基酸残基字符的一致性刻画的. 然而, 对结构上相似但序列相似性

较低的蛋白质对，如果进行适当的氨基酸突变，可以在仍然保持它们的结构相似性条件下发现它们之间的序列差异并不是那么大. 因此，如果 20 个代表天然氨基酸残基的字符能根据它们的物理化学特性进行适当的归类简并，那么氨基酸残基字符的一致性将得到提高，从而使得序列的相似性得到提高. 这些经过适当简化的字符所编码的序列进行比对，应该能够更好地与结构比对的结果相一致，从而提高用序列比对的方法识别蛋白质结构保守区域的准确性. 值得注意的是，将 20 种天然氨基酸残基进行归类简并，并不是简化程度越高越好，因为过度的简化，即将氨基酸残基归类的过于粗糙，将会严重的丢失氨基酸残基之间明显的差异信息，从而使得序列的重要信息丢失. 所以，一个优化的氨基酸残基归类方法是应用简化后的序列和替代矩阵进行序列比对的前提条件.

为了检验简化后的序列和替代矩阵进行序列比对的准确性，我们从 DAPS 数据中随机挑选了 200 对蛋白质. 这些蛋白质已经通过结构比对进行了最优的结构叠加，而它们的序列用前面得到的不同简化程度($2 \leqslant N \leqslant 200$)下的字符进行转换，而将简化后的序列和替代矩阵进行序列比对. 相对于标准的结构比对结果，这 200 对蛋白质的序列比对的 C_R 和 D_R 的平均值随着简化后氨基酸残基的组数 N 的变化曲线在图 3 中给出. 从图中可以看出，当 N=6~9 时，C_R 和 D_R 达到最大值，即当 20 种天然氨基酸残基适当归类到 6~9 组时，用简化后的序列和替代矩阵进行序列比对识别结构保守性的准确度最高.

图 4~8 给出了一个具体例子，用于比较序列比对结果与结构比对，并分析简化和未经简化的序列和替代矩阵对序列比对准确度的影响. 进行分析的两条蛋白质链在PDB数据库中的编码分别为 1B8X_A和 2GSR_A. 图 4 显示的是这 2 条蛋白质链在DAPS数据库中结构比对结果的三维演示图，进行蛋白质三维分子结构演示的软件是MolMol[24]. 从此图可以看出，这 2 条蛋白质链确实具有很相似的空间结构. 经过结构比对后，大部分区域可以被很好地叠加，即它们均为保守的结构区域. 然而，这 2 条蛋白质链中也有 5 个区域不具有相似的空间结构，因此不能被叠加，在结构比对结果中即存在插入和缺失的空位区域. 为了方便与序列比对相比较，图 5 用序列比对的方式给出了这 2 条蛋白质链结构比对的结果，其中 2

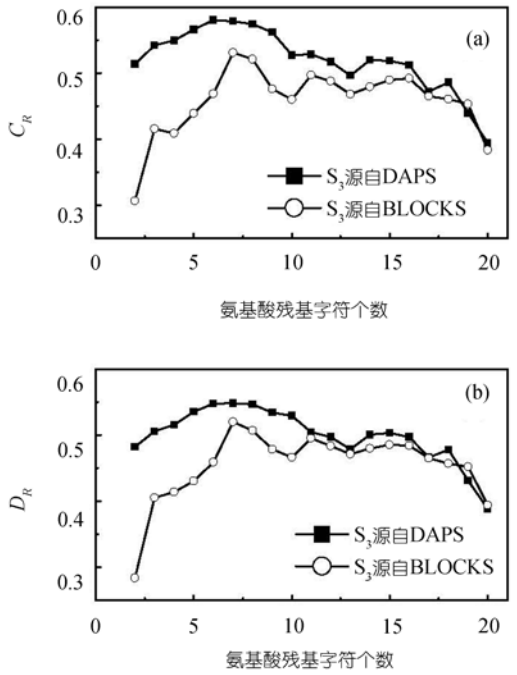


图 3 在序列比对中求得的 C_R 和 D_R 与氨基酸残基字符个数的变化曲线

(a)和(b)中的 2 条曲线分别来自与 DAPS 和 BLOCK 数据库的 S_3 子集

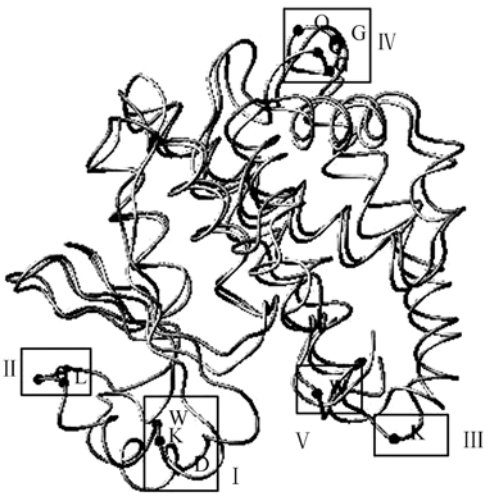


图 4 两个蛋白质进行结构比对和序列比对的实例(PDB 代码分别为 1B8X_A 和 2GSR_A)

2 条蛋白质不能叠加的区域用方框标记并给出了残基字符

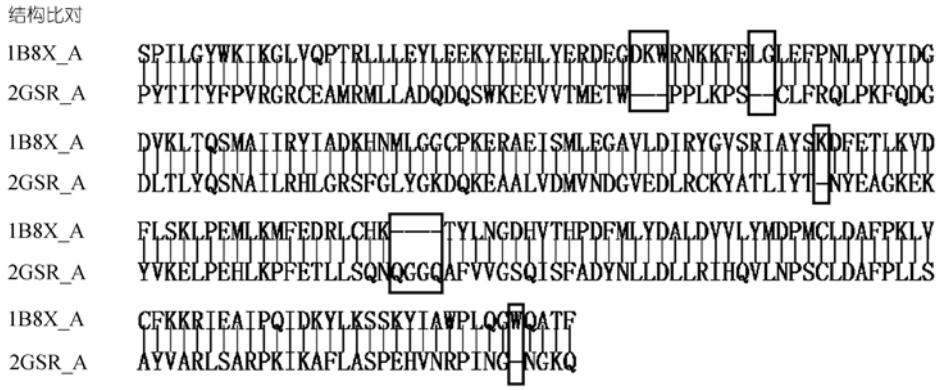


图 5　对这 2 条蛋白质链的进行结构比对的结果

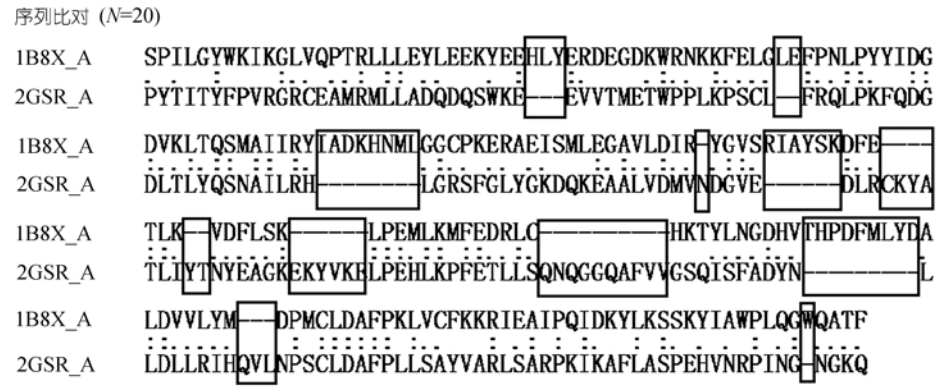


图 6　用 20 个氨基酸残基字符编码对这 2 条蛋白质序列进行序列比对的结果

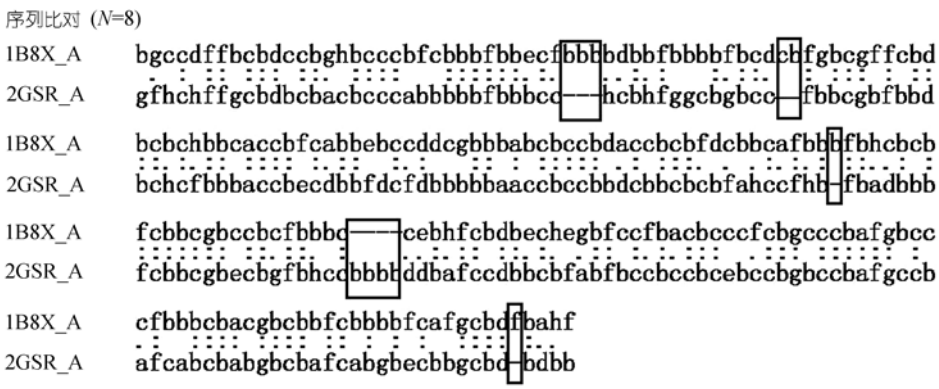


图 7　用 8 个氨基酸残基字符编码对这 2 条蛋白质序列进行序列比对的结果

条链中的空间等价位点用符号“|”相连接, 表示该字母所代表的氨基酸残基在空间中可以叠加在一起. 而这 2 条链中不可以进行空间叠加的不相似空间结构用空格表示, 并在图中用方框标示出来. 这些符号“|”相连接的空间等价位点则作为标准位点用于评价序列比对所识别的序列保守位点的准确性. 相应的, 用 20(未经简化), 8, 2(经过简化)个字母所表示的氨基酸残基组数进行序列比对的结果分别呈现在图 6~8

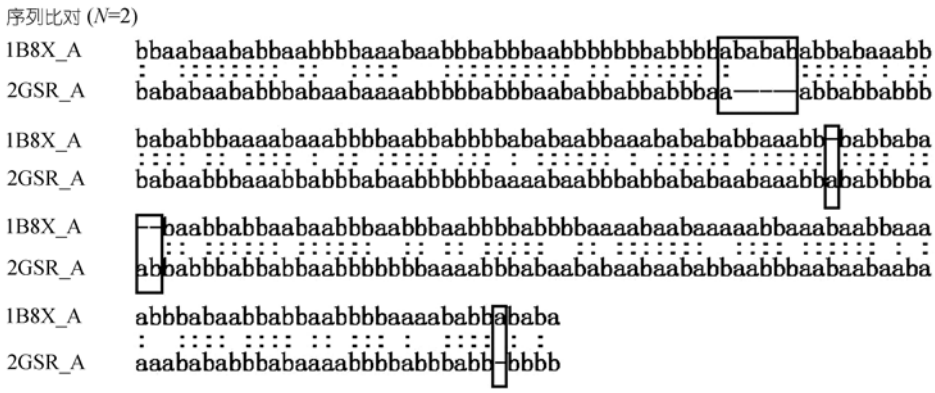


图 8 用 2 个氨基酸残基字符编码对这 2 条蛋白质序列进行序列比对的结果

中. 从这些图中, 可以看出随着氨基酸残基组数的减少, 被排在同一列中的位点, 即序列识别的保守位点个数确实增加了. 同时, 空格区域的个数和长度也减少了. 因为从直观上, 序列比对识别的保守位点个数太多, 可以通过比较序列比对中空格区域的位置和长度来间接判断序列比对的准确性. 因为空格区域, 即非保守位点, 被序列比对正确识别, 它们之间的保守区域也应该被正确识别. 在图 6~8可以明显看出, 当将 20 种天然氨基酸归类为 8 组时, 经过简化的序列和替代矩阵进行序列比对的结果最接近于结构比对的结果. 从图 7 中可以看出, 当将 20 种天然氨基酸归类为 8 组时序列比对结构同样有 5 个空格区域, 其中后 3 个区域的位置和长度与结构比对结果中所显示的空格区域完全一致, 显然这 3 个空格区域之间的序列保守位点与结构等价位点相一致. 而图 7 中前面 2 个空格区域的位置虽然和结构比对结构中空格区域的位置有所差别, 但长度是一样的. 因此, 也只有这 2 个空格区域之间的少许序列保守位点为错误识别, 而两端的序列保守位点与空间等价位点相一致, 为正确识别. 所以, 当将 20 种天然氨基酸归类为 8 组时, 经过简化的序列和替代矩阵进行序列比对有较高的准确性. 然而从图 6 和 8 可以看出, 不进行氨基酸归类简并或者归类的过于粗糙的情况下, 序列比对的准确性却不高. 当不进行氨基酸归类简并, 用 20 个残基字符进行序列比对时, 因为 2 个蛋白质之间的序列差异性太大, 导致在序列比对中过多的引入空格, 使得序列比对结果与结构比对结果完全不能相符合. 而对氨基酸残基归类过于粗糙时, 如只分为疏水和极性氨基酸 2 类, 使得序列之间的信息严重丢失, 在序列比对中不能分辨保守区域和非保守区域, 同样和结构比对的结果不符. 因此, 在对 1B8X_A 和 2GSR_A 两条蛋白质链进行不同程度简化后的序列比对结果表明, 当进行适当的氨基酸残基归类有助于提高序列比对的准确性, 这个具体的实例也与图 3 中的统计结果相一致.

2.3 用其他程序进行序列比对

图 6~8 中进行序列比对的程序是 FASTA, 替代矩阵是从 DAPS 数据库中构建而来的. 为了和其他同样的序列比对程序及替代矩阵相比较, 我们用 3 种最为常用的程序 FASTA, BLAST 和 ClustalX 进行序列比对, 替代矩阵为 BLOSUM62 矩阵. 这些程序进行序列比对的结果也与结构比对结果相比较, 得出 C_R 和 D_R, 结果见表 1. 进行比对的蛋白质序列同样是上部分所用到的 200 对蛋白质序列. 在序列比对中, 空格的引入和延伸参数为–11 和–1, 其他参数采用各个程序默认的参数. 在本部分中, 进行比对的序列均未经任何归类简化, 替代矩阵也采用原始的 20×20 的矩阵. 从此表中可以看出进行适当简化(N=8)的序列进行序列比对的准确性最高. 用 ClustalX 进行序列比对来识别结构保守区域的准确度也比较高. 在图 9 中也给出了用 ClustalX 对前面 2 个具体的蛋白质链 1B8X_A 和 2GSR_A 进行序列比对的结果. 同样, 也发现, ClustalX 对这 2 条蛋白质链的比对准确度还

是不如进行适当序列简化后的序列比对. 用 BLAST 程序进行序列比对来识别结构保守区域的准确度最低, 这是因为 BLAST 主要适用于在海量数据库中进行同源搜索.

2.4　基于 BLOCKS 数据库进行氨基酸残基归类

以前发表的很多进行氨基酸残基归类简化的工作是基于分析由BLOCKS数据库中构建的BLOSUM替代矩阵的[25,26]. BLOCKS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, 因此BLOCKS数据库中的蛋白质大部分是由高序列相似性的蛋白质组成, 而所包含的低序列相似性蛋白质比例较小[27~29], 因此, 对从BLOCKS数据库和DAPS数据库进行氨基酸残基归类结果进行了比较.

图 10(a)给出了从 BLOCKS 数据库进行氨基酸残基归类的结果. 从该图中可以看出, 所有 20 种天然氨基酸残基最终都归成 2 类: 疏水残基和极性残基, 并且氨基酸归类的大致结果与从 DAPS 数据库中进行氨基酸归类的结果相类似, 当然还是有一些细微差别.

为了进一步解释对于序列相似性较低的数据子集(S≤30%)中 20 种天然氨基酸残基之间的复杂关系, 用主成分分析方法, 把 20 种氨基酸残基的替代矩阵作为 20 个在二十维空间中的矢量, 并将复杂的二十维的矢量投影到二维空间中. 图 10(b)和(c)给出了从 BLOCKS 和 DAPS 数据库中得到的替代矩阵进行主成分分析的结果, 很明显, 用主成分分析的方法能够从 DAPS 数据库中更加清楚的显示 20 种氨基酸残基之间的关系, 例如, 这 20 种氨基酸残基更加清楚的分成了疏水和极性两个部分. 这是因为 DAPS 含有很多的结构上保守而序列上差异较大的蛋白质对, 能够更好地反映出 20 种天然氨基酸在序列相似性较低的蛋白质对中的替代关系.

更为重要的是, 从图 3 可以看出, 基于 BLOCKS 数据库进行序列简化所得到的 C_R 和 D_R 要低于基于 DAPS 数据库进行序列简化的结果, 因此用序列比对来识别结构保守性区域的能力上, 从 DAPS 数据库进行序列简化要优于从 BLOCKS 数据库.

2.5　用其他方法进行氨基酸残基归类

文中的归类方法是一种基于它们的替代分值的逐步聚类的归类方法, 正如前面所述, 这种归类方法特别适用于序列相似性较低的数据库. 为了对比, 也应用了其他的归类方法进行氨基酸残基的归类. 在

表 1　对 DAPS 数据库中任意挑选出的 200 对蛋白质序列用不同比对程序或不同的氨基酸替代矩阵进行序列比对的结果比较

Program	FASTA	FASTA	FASTA	BLAST	ClustalX
Matrix	Reduced 8×8 matrix for S_3 from DAPS	20×20 matrix for S_3 from DAPS	BLOSUM62	BLOSUM62	BLOSUM62
Number of letters (N)	8	20	20	20	20
C_R	0.58	0.38	0.41	0.36	0.52
D_R	0.55	0.39	0.42	0.34	0.49

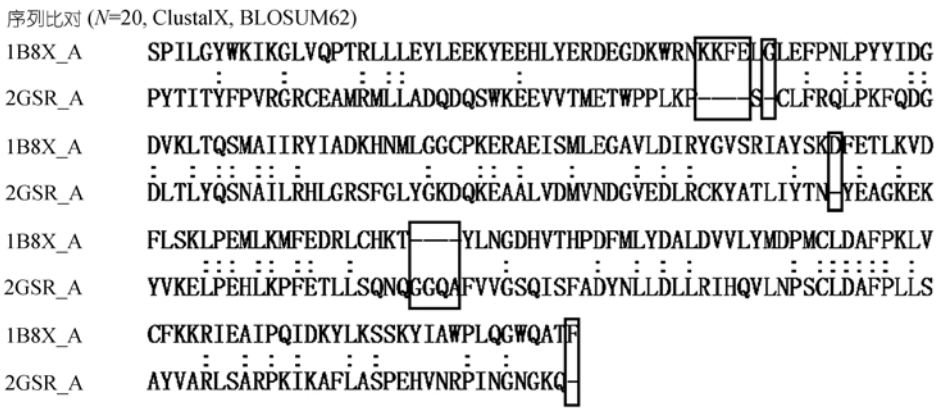


图 9　用 ClustalX 程序和 BLOSUM62 矩阵对这 2 条蛋白质序列进行序列比对的结果

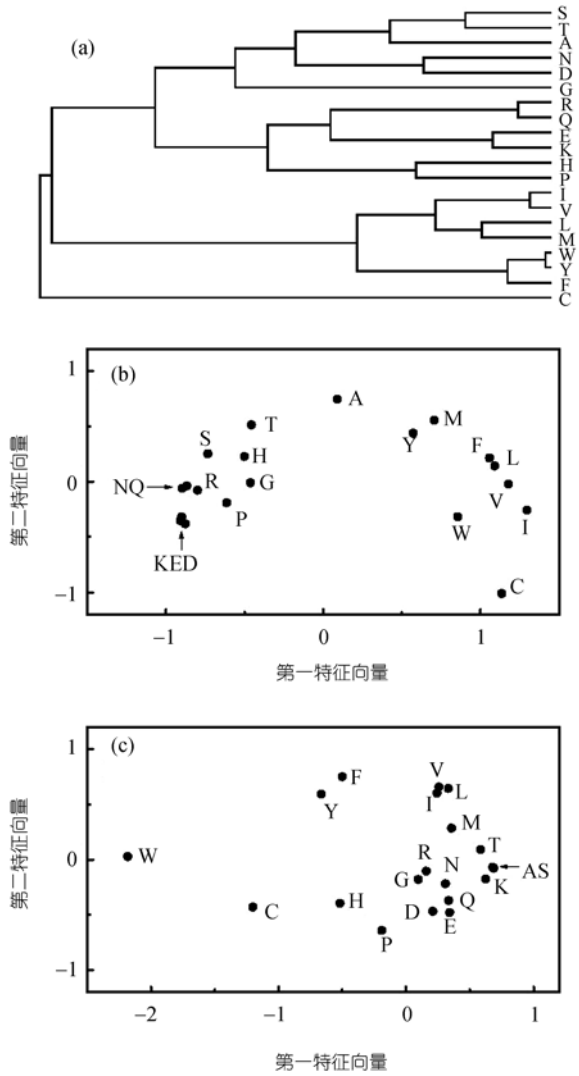


图 10

(a)为用基于替代分值的方法对 BLOCK 数据库中的 S_3 子集, 即序列相似性 S≤30%的子集进行氨基酸残基归类的结果. (b)和(c)分别为对 DAPS 数据库和 BLOCK 数据库中的 S_3 子集所构建的氨基酸替代矩阵进行主成分分析的结果

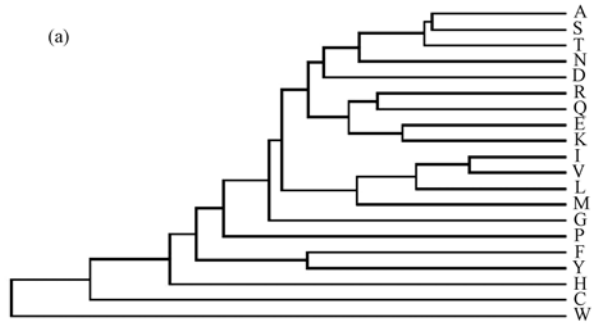
聚类分析中, 一种最为常用的方法是UPGMA的归类方法, 该方法中替代矩阵中 20 种天然氨基酸之间的相互替代分值可以看作在二十维空间中的 20 个矢量. 基于这些矢量可以计算它们之间的距离, 而具有最小距离的矢量则被归类为一组, 而 20 个矢量可以被逐步归类在一起. 应用UPGMA算法, 对DAPS数据的 S_3 数据子集中的进行氨基酸残基归类, 再用PHYLIP程序包[30]中的KITSCH程序进行归类分析, 其归类结果在图 11(a)中呈现. 在图 11 中, 可以看出用这种方法对低序列相似性的蛋白质比对数据库进行氨基酸归类的结果并不好, 甚至不能将 20 种天然氨基酸正确的归类为疏水残基和极性残基.

在氨基酸残基归类中, 另一个应用较多的方法就是动态归类法, 在这种归类方法中, 仍然是将 20×20 的氨基酸替代矩阵作为 20 个矢量, 不同的是, 在这种方法中需要指定组数N, 再计算每组的中心位置和每个矢量与这个中心位置的距离, 然后用动态规划的策略调整不同组中的成员, 使得这些矢量与中心位置的距离之和达到最小值, 得到N组条件下的最佳归类方式[31]. 图 11(b)给出了用动态归类方法对DAPS数据库中的S_3 子集(序列相似性S≤30)进行氨基酸残基归类的结果, 从该图中可以看出, 这种动态归类的方法, 对于低序列相似性的结构比对数据库DAPS进行氨基酸残基归类的结果仍然不够理想.

3 讨论

通过序列比对识别蛋白质的结构保守区域是分析蛋白质结构与序列关系的有效方法, 并在生物学和生物信息学中有着广泛应用, 如通过同源建模的方法预测蛋白质的三维空间结构. 然而, 对于序列相似性小于 30%的那些蛋白质对, 因为序列比对的准确性不高, 所以往往不能正确识别蛋白质的结构特征. 而这些结构相同但序列差异的蛋白质在自然界中分布的十分广泛, 研究它们的序列结构关系, 及其进化起源具有非常重要的生物学意义. 应该如何提高序列比对对这些结构相似而序列相似性较小的蛋白质的准确性呢? 在本研究提出通过氨基酸残基归类来合理提高蛋白质序列的相似性, 从而提高蛋白质序列比对准确性的方法.

以前的很多研究结果证明, 如果将 20 种天然氨基酸进行合理分类, 可以有效地降低蛋白质序列的复杂性而不丢失蛋白质的主要序列和结构特征. 而在应用氨基酸残基归类降低蛋白质序列复杂性的过程中, 首先需要选择一个合理的蛋白质比对数据库用于正确反映氨基酸残基在蛋白质序列中的作用和相互关系. DAPS 数据库是蛋白质结构比对数据库, 很好地反映了低序列相似性序列中氨基酸残基的相互关系. 正因为低相似性的序列的相互信息较弱, 基



(b)

2	TVMLFWYC																		ARNDQEGHKPST
3	TVLMFWC										YA								RNDQEGHKPST
4	TVLMFW						C				YA								RNDQEGHKPST
5	TVLMFW						C				YA				G				RNDQEHKPST
6	TVLMF			WY			C				AH				G				RNDQEKPST
7	TVLMF			WY			C				AH				GP			R	NDQEKST
8	TVLMF			WY			C			A			G			R		Q	NDEHKPST
9	TVLMF			WY			C			A			G			P		H	K	RNDQEST
10	TVLML		F		W		Y		C		A			H			G		RN	DQEKPST
11	TVLMF			W		Y		C		A			H			G		R	N	Q	DEKPST
12	TVLM		F		W		Y		C		A			H			G		N	Q	T	RDEKPS
13	TVLM		F		W		Y		C		A			H			G		N	Q	P	R	DEKST
14	TVLM		F		W		Y		C		A			H			G		N	Q	P	R	K	DEST
15	TVLM		F		W		Y		C		A			H			G		N	Q	P	R	K	D	EST
16	TVLM		F		W		Y		C		A			H			G		N	Q	P	R	K	S	T	DE
17	TVL	M		F		W		Y		C		A			H			G		N	Q	P	R	K	S	T	DE
18	TVL	M		F		W		Y		C		A			H			G		N	Q	P	R	K	S	T	D	E

图 11

(a)为用 UPGMA 的方法对 DAPS 数据库中的 S_3 子集，即序列相似性 S≤30%的子集，进行氨基酸残基归类的结果. (a)为用动态归类的方法对 DAPS 数据库中的 S_3 子集，即序列相似性 S≤30%的子集，进行氨基酸残基归类的结果

于计算氨基酸残基在空间中的矢量距离的方法不能很好地区别 20 种天然氨基酸残基的相似点和差异点，所以准确性不高. 而本文给出的归类方法是直接基于氨基酸残基的替代分值的，能够直观而准确的将相互关系最强，即替代分值最大的两组残基归类，从而将 20 种天然氨基酸残基逐步正确归类在一起. 相比较于以前文献中发表的氨基酸残基归类结果，可以看出在低序列相似性的蛋白质中，氨基酸残基的归类情况与高序列相似性蛋白质中的归类情况仍有很多相似之处，如均可归类为疏水性残基和极性残基 2 个大组. 当然由于蛋白质序列上的差异，也反映出细节分类上的不同之处.

简化的蛋白质序列和替代矩阵可以用于识别蛋白质的结构保守区域. 对于用简化后的序列和替代矩阵进行序列比对识别蛋白质结构保守或相似区域，我们发现当 N=6~9 时，序列比对的识别准确率确实有所提高，这说明对 20 种天然氨基酸进行适当的归类，确实有助于提高序列比对对那些低序列相似性的蛋白质的结构识别能力. 值得指出的是，我们的工作主要是研究了简化后的序列和替代矩阵对替代序

列比对对蛋白质结构保守区域识别能力的提高，使用的是最为常用的序列比对程序. 相信使用更为准确的序列比对算法，结合氨基酸残基归类的思路，能够进一步提高序列比对的准确性.

参 考 文 献

1 Bowie J U, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science, 1991, 253: 164—170 [DOI]

2 Jones D T, Taylor W R, Thornton J M. A new approach to protein fold recognition. Nature, 1992, 358: 86—89 [DOI]

3 Regan L, Degrado W F. Characterization of a helical protein designed from first principles. Science, 1988, 241: 976—978 [DOI]

4 Kamtekar S. Protein design by binary patterning of polar and nopolar amino acids. Science, 1993, 262: 1680—1685 [DOI]

5 Plaxco K W. Simplified proteins: minimalist solutions to the “protein folding problem”. Curr Opin Struct Biol, 1998, 8: 80—85 [DOI]

6 Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nature Struct Biol, 1999, 6: 1033—1038 [DOI]

7 Henikoff S, Henikoff J G. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA, 1992, 89: 10915—10919 [DOI]

8 Ogata K, Ohya M, Umeyama H. Amino acid similarity matrix for homology derived from structural alignment and optimized by the Monte Carlo method. J Mol Graph Model, 1998, 16: 178—189

9 Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins, 2005, 58: 321—328 [DOI]

10 Friedberg I, Kaplan T, Margalit H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. Protein Sci, 2000, 9: 2278—2284

11 Mallick P, Weiss R, Eisenberg D. The directional atomic solvation energy: An atombased potential for the assignment of protein sequences to known folds. Proc Natl Acad Sci USA, 2002, 99: 16041—16046 [DOI]

12 Kleiger G. PFIT and PFRIT: Bioinformatic algorithms for detecting glycosidase function from structure and sequence. Protein Sci, 2004, 13: 221—229 [DOI]

13 Karlin S, Altschul S F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA, 1990, 87: 2264—2268 [DOI]

14 Altschul S F. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol, 1991, 219: 555—565 [DOI]

15 Karlin S, Altschul S F. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci USA, 1993, 90: 5873—5877 [DOI]

16 Higgins D G, Sharp P M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 1988, 73: 237—244 [DOI]

17 Holm L, Sander C. Mapping the protein universe. Science, 1996, 273: 595—602 [DOI]

18 Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins, 1998, 33: 88—96 [DOI]

19 Blake J D, Cohen F E. Pairwise sequence alignment below the twilight zone. J Mol Biol, 2001, 307: 721—735 [DOI]

20 Dosztanyi Z, Torda A E. Amino acid identity matrices based on force fields. Bioinformatics, 2001, 17: 686—699 [DOI]

21 Johnson M S, Overington J P. A structural basis for sequence comparisons an evaluation of scoring methodologies. J Mol Biol, 1993, 233: 716—738 [DOI]

22 Li T. Reduction of protein sequence complexity by residue grouping. Protein Eng, 2003, 16: 323—330 [DOI]

23 Fan K, Wang W. What is the minimum number of letters required to fold a protein. J Mol Biol, 2003, 328: 921—926 [DOI]

24 Koradi R, Billeter M, Whrich K. MOLMOL: A Program for Display and Analysis of Macromolecular Structures. J Mol Graphics, 1996, 14: 51—55 [DOI]

25 Henikoff S. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene, 1995, 163: GC17—GC26 [DOI]

26 Pietrokovski S, Henikoff J G, Henikoff S. The Blocks database–a system for protein classification. Nucleic Acids Res, 1996, 24: 197—200 [DOI]

27 Clarke N D. Sequence “minimization”: exploring the sequence landscape with simplified sequences. Curr Opin Biotech, 1995, 6: 467—472 [DOI]

28 Riddle D S. Functional rapidly folding proteins from simplified amino acid sequences. Nature Struct Biol, 1997, 4: 805—809 [DOI]

29 Akanuma S, Kigawa T, Yokoyama S. Combinatorial mutagenesis to restricted amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA, 2002, 99: 13549—13553 [DOI]

30 Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 1985, 39: 783—791 [DOI]

31 Liu X. Simplified amino acid alphabets based on deviation of conditional probability from random background. Phys Rev E, 2002, 66: 021906-1—021906-4