核酸、氨基酸分类和蛋白质二级结构关系的分析与分子设计

Relationships of Nucleotide, Amino Acid Sequence and Secondary Structure of Protein and Molecular Design

马 鹏*,王联结

MA Peng* and WANG Lian-Jie

陕西科技大学生命科学与工程学院,咸阳 712081

Life College of Science and Engineering, Shaanxi University of Science & Technology, Xianyang, 712081, China

摘 要 核酸序列中包含一定的蛋白质结构信息。根据通常情况下遗传密码表中密码子中间位的碱基配对时产生的氢键数目,尝试将 20 种氨基酸划分为两类,并用自编的计算机软件对蛋白质二级结构数据库中两类氨基酸的类聚现象进行了统计分析。结果表明,使用这种方法对氨基酸进行划分后,氨基酸残基具有较大概率与划入同一类的氨基酸残基相邻出现,并且这种聚集体对二级结构具有一定的偏好性。最后按照该方法设计了一段氨基酸序列并给出了预测服务器预测得到的结构。

关键词 核酸,氨基酸序列,二级结构,分子设计

中图分类号 Q518.1 **文献标识码** A **文章编号** 1000-3061(2007)06-1082-04

Abstract May the structure information of protein be obtained from the corresponding nucleotide sequence? For this question, a computer program was used to conduct a statistical analysis about the clustering phenomena of amino acids. In this method, 20 kinds of amino acid were classified into 2 types according to the number of hydrogen bonds formed by middle base of their correlation codons. It can be seen that amino acid has a rather great possibility of neighboring on another of its class and this assembly has a tendency to forming specific secondary structure. A sequence was designed and its secondary structure was predicted by this prediction software.

Key words nucleotide, amino acid sequence, secondary structure, molecular design

过去的几十年中,出现了多种多样的蛋白质二级结构预测方法,最早出现的研究方法是统计序列中氨基酸残基对结构的倾向性[1-3]。但近年来,通过氨基酸序列预测蛋白质二级结构的研究又有复苏。长期以来,人们试图通过分析核酸序列找到蛋白质空间结构的信息,例如从氨基酸的密码子出发来研究序列和结构之间的关系[4-6]。对氨基酸残基聚集体的研究也有报道[3,7-9]。本文根据氨基酸密码子和反密码子配对时中间位碱基之间正常情况下形成的氢键数目的不同对氨基酸残基进行了重新分类(以下简称为氢键数法),并对分类后可能在蛋白质序列中存在的类聚现象(同一类氨基酸残基的连续分布)做了初步研究,最后依据该方法进行了氨基酸序列设计试验。

1 方法

1.1. 氢键数方法

根据 20 种氨基酸三联密码子中间位的碱基在正常情况下能够形成的氢键数目为 2 或 3 的不同,将 20 种氨基酸分为两大类,其中:第一类氨基酸残

Received: January 12, 2007; Accepted: May 18, 2007.

* Corresponding author. E-mail: mapeng_true@yahoo.com.cn

基包括 A、G、C、T、P、R、S 和 W；而第二类包括 D、E、F、I、K、L、N、Q、V、H、Y 和 M。

1.2　数据库

选用 DSSP 数据库，并使用相似性小于 25% 的蛋白质选择列表，最后取得了 923 个非同源蛋白质数据。在 DSSP 二级结构 8 态分类到 3 态分类转换中借鉴前人工作[10] 采用如下划分：α 螺旋 h(α 螺旋 H，3-10 螺旋 G，π 螺旋 I)，β 折叠 e(β 折叠 E)和卷曲 c(β 桥 B，转角 T，回折 S，无规卷曲 C)。将 B 结构划入卷曲中是因为它作为一个独立的连接键，很难被认为是一种规则结构[3]。再将 3 种二级结构按照其是否属于规则结构划为两大类：第一类为非规则结构(c)；第二类为规则结构(h，e)。

1.3　统计方法

为了研究氢键数方法在蛋白质二级结构研究中的应用意义，我们进行了一些统计计算。观察表明，分类后某些氨基酸残基在一些蛋白质中具有类聚倾向。那么这种类聚是否在蛋白质中具有普遍性？在不考虑二级结构的情况下，对蛋白质中类聚出现概率的统计对此做出了衡量。类聚的出现如果有相当大的可能性，对类聚和蛋白质二级结构之间对应关系的研究则是必要的。这种对应关系的研究包括两个方面：类聚中的残基是否倾向形成特定二级结构；具有特定二级结构的氨基酸残基是否处于对应类聚中。

(1)在不考虑二级结构的情况下，统计出类聚残基数量 N，该数值与残基总数 N_t 的比值 P 作为衡量类聚现象是否具有普遍性的统计量，表示一个氨基酸残基处于类聚的概率：

$$P = \frac{N}{N_t} \tag{1}$$

(2)氨基酸的类聚概率：令 $P_{i,j}$ ($j = 1\cdots20$)表示第 j 种氨基酸出现在其所在类别类聚中的概率，$N_{i,j}$ 表示第 j 种氨基酸出现在第 i 类类聚中的残基数，N_j 表示第 j 种氨基酸的总残基数。显然，当类聚最小长度定义为 2 时，$P_{i,j}$ 成为氨基酸 j 与同类氨基酸相邻出现概率。上述各项有以下关系：

$$P_{i,j} = \frac{N_{i,j}}{N_j} \tag{2}$$

(3)类聚呈现对应二级结构分类概率：处于 i 类($i = 1,2$)类聚中且具有相应二级结构分类(第一类氨基酸残基对应于第一类结构分类，同样第二类氨基酸残基对应第二类结构分类，下同)的残基数 N_i，处于 i 类类聚中且二级结构已知的氨基酸残基总数 $N_{i,t}$，两者的比值 P_i 作为 i 类类聚呈现对应二级结构分类的概率。若以 P_t 表示两类的总概率，则有：

$$P_i = \frac{N_i}{N_{i,t}} \tag{3}$$

$$P_t = \frac{\sum N_i}{\sum N_{i,t}} \tag{4}$$

(4)特定二级结构分类氨基酸残基处于对应类聚概率：处于 i 类($i = 1,2$)类聚中且具有相应二级结构分类的残基数 N_i，呈现该类二级结构(规则结构，非规则结构)的残基的数量 R_t，两者比值 $P_{i,i}$ 作为具有 i 类二级结构氨基酸残基处于对应类聚概率。若以 P_t' 表示两类的总概率，则有：

$$P_{i,i} = \frac{N_i}{R_t} \tag{5}$$

$$P'_t = \frac{\sum N_i}{\sum R_t} \tag{6}$$

上述处理过程中，因为类聚出现概率以及氨基酸类聚概率的统计不涉及残基的二级结构，所以在这两项统计中包括结构不明确的残基。而在类聚和二级结构对应关系的两项统计中，二级结构不明确的残基不纳入计算。

2　结果和讨论

2.1　类聚的存在

类聚长度的不同定义显然会对统计结果产生影响，较长的类聚会更少地出现。为了体现这种影响，我们计算了类聚长度定义为 2 ~ 5 的不同长度下的概率数据。计算结果见表 1，2。

表 1　不同类聚定义长度下类聚出现概率

Table 1　Probabilities of clustering of the same type amino acids on different cluster length

	$L=2$	$L=3$	$L=4$	$L=5$
P	0.751	0.524	0.355	0.240
P'	0.760	0.511	0.312	0.188

L: cluster length; P: the probability of clustering of the same classification; P': the limiting probability of random cluster.

使用氢键数方法划分后的氨基酸具有相当大的倾向出现在同类氨基酸的相邻位置，并且在自然界蛋白质中存在着大量的由同类氨基酸形成的类聚体。从表 1 可以看出：类聚长度定义越大，类聚概率越小；由于低聚集长度下随机聚集概率过高的本底值，低聚集长度下的类聚概率与随机聚集概率的极限值相差不大，随着聚集长度增加类聚概率开始明

表 2 氨基酸残基在不同类聚长度定义下处于类聚的概率
Table 2 Probability of amino acid stands in clustering

Amino acid	Classification	P_{ij}			
		$L=2$	$L=3$	$L=4$	$L=5$
A	1	0.629	0.342	0.172	0.0864
G	1	0.637	0.357	0.186	0.0943
C	1	0.653	0.351	0.193	0.0914
T	1	0.631	0.345	0.176	0.0879
P	1	0.600	0.322	0.158	0.0758
R	1	0.606	0.321	0.149	0.0735
S	1	0.640	0.351	0.183	0.0905
W	1	0.650	0.356	0.188	0.0960
D	2	0.837	0.649	0.486	0.354
E	2	0.854	0.672	0.501	0.367
F	2	0.831	0.632	0.455	0.324
I	2	0.839	0.639	0.467	0.332
K	2	0.860	0.685	0.509	0.377
L	2	0.825	0.629	0.460	0.329
N	2	0.842	0.656	0.487	0.352
Q	2	0.834	0.652	0.477	0.344
V	2	0.836	0.647	0.470	0.331
H	2	0.820	0.631	0.462	0.323
Y	2	0.831	0.628	0.458	0.325
M	2	0.835	0.637	0.458	0.333

L: cluster length; $P_{i,j}$: probability of amino acid stands in clustering.

显大于随机聚集概率,如 $L=5$ 时,随机聚集概率仅为统计数据的 78.3%。同样从表 2 中可以看出 63%左右的一类氨基酸残基都与同类氨基酸残基相连,84%左右的二类氨基酸残基与同类氨基酸残基相连出现,这两个值都明显大于随机排列下的值 40%和 60%。

2.2 类聚呈现对应二级结构分类概率

使用式(3)和式(4),我们计算了每类类聚呈现对应二级结构分类概率 P_i 以及两类的总概率 P_t,同样包含不同长度下的结果。结果见表 3。

表 3 类聚呈现对应二级结构分类概率
Table 3 Probability of clustering residue with secondary structure correlates to its clustering class

	$L=2$	$L=3$	$L=4$	$L=5$
P_1	0.390	0.427	0.454	0.473
P_2	0.784	0.802	0.811	0.818
P_t	0.662	0.714	0.750	0.775

L: cluster length; P_1: probability of 1st clustering residue with 1st secondary structure (c); P_2: probability of 2nd clustering residue with 2nd secondary structure (h, e); P_t: the total probability of 2 classes of clustering residue with secondary structure correlates to its clustering class.

统计结果说明,处于类聚中的残基对相应的二级结构分类具有一定的倾向性,较长类聚比较短类聚对相应二级结构的倾向性略有上升。从表 3 数据可以看出:对于一类氨基酸,39.0% ~ 47.3%的类聚残基呈现出了对应的非规则结构(c 结构);对于二类氨基酸,78.4% ~ 81.8%的类聚残基呈现了所对应的规则结构(h,e 结构)。总的来看,对于所有已知二级结构的类聚残基,66.2% ~ 77.5%的类聚残基呈现了对应的二级结构分类,若聚集体对结构无倾向该值应为 50%(2 结构分类下)。同时,从表 3 不同类聚长度的数据可以看到,随着类聚最短长度的增加,类聚呈现对应二级结构分类的概率出现上升,但上升幅度不是很大。

2.3 特定二级结构分类的氨基酸残基处于对应类聚的概率

依据式(5)及(6),计算每类二级结构分类的氨基酸残基处于对应类聚的概率 $P_{i,i}$,以及两类结构分类的总概率 P'_t。表 4 为不同类聚长度下特定二级结构分类的氨基酸残基处于对应类聚的概率 $P_{i,i}$。

表 4 特定结构残基处于相应类聚概率
Table 4 Probability of that residue stands in clustering corresponding to its secondary structure

	$L=2$	$L=3$	$L=4$	$L=5$
$P_{1,1}$	0.328	0.192	0.102	0.0530
$P_{2,2}$	0.570	0.456	0.343	0.251
P'_t	0.502	0.383	0.276	0.196

L: cluster length; $P_{1,1}$: probability of that 1st secondary structure (c) residue stands in 1st clustering; $P_{2,2}$: probability of that 2nd secondary structure (h, e) residue stands in 2nd clustering; P'_t: the total probability of 2 types of secondary structure residues stand in clustering corresponding to their secondary structure.

从表 4 知:处于非规则结构中的氨基酸残基,32.8% ~ 5.3%(类聚长度 2 ~ 5)出现在一类类聚中;处于规则结构中的氨基酸残基,57% ~ 25.1%(类聚长度 2 ~ 5)出现在二类类聚中。两类结构的数据加以处理得到总的概率,该值为 50.2% ~ 19.6%(类聚长度 2 ~ 5)。同样可以看出:随着类聚长度的增大,该数值出现明显的减小趋势。这主要是因为类聚概率随着类聚长度增大锐减所致。

以上分析说明:类聚现象真实存在,但并不是所有残基的普遍现象;类聚与结构之间具有一定对应关系;类聚长度增大能够更好地与随机聚集情况区分,但其类聚体总量减少。本文提出的规律并不独立作为一种预测算法,可以作为其他算法的辅助方法或者新算法的开发基础。

3 序列设计试验

使用该方法,我们设计了一段氨基酸序列(结构

形式：非规则-规则-非规则）。并将该序列提交 http://www.predictprotein.org 进行蛋白质二级结构预测，以对我们的设计进行检验。

表 5　设计的氨基酸序列及其二级结构

Table 5　The sequence designed and its secondary structure

Sequence	CCARRCPEEFKVVPTTGW
Designed	LLLLLL＊＊＊＊＊＊LLLL
PHD	LLLLLLLLLLLLLLLLL
PROF	L.LLLL..L...LLLL

"Designed": the secondary structure designed; "PHD": the result of the prediction server; "L": Loop, irregular secondary structure; ".": the structure can not be given; "＊": regular secondary structure.

从结果可以看出，对 Loop 区的设计是比较成功的，在该区域与 PHD，PROF 的判定结果高度一致。而在规则区域的设计与预测服务器返回的结果存在一定偏差，PHD 认为该区域为 LOOP 结构，而 PROF 认为该区域结构难以判定。作为一种简洁的方法得到这样的结果是相当不错的，由于其简洁性，应该在分子设计工作中有一定实用意义。后续工作对于分类和类聚的细化以及与其它方法的结合将有助于提高设计的准确性。

4　结论

在根据氢键数方法将氨基酸分类以后，我们对 923 个非同源蛋白质的氨基酸残基序列进行了统计。统计数据表明：划分后的氨基酸具有相当大的概率与同类残基相连出现；这种特定氨基酸相连出现的类聚体在自然界蛋白质中大量存在；并且这种聚集体对二级结构分类具有一定的倾向性。随着类聚最短长度的增大，类聚出现的概率以及呈现某种二级结构分类的残基处于对应类聚的概率数值降低，但出现类聚的残基呈现对应二级结构分类的概率数值增大，长类聚比短类聚具有更高的可靠性。使用这种简洁的方法，我们设计了一段具有特定结构形式的氨基酸序列，对于 Loop 区的设计与预测服务器给出的结果高度一致。考虑到方法的简洁性，应该具有一定的实用价值，当然，对于氨基酸分类以及类聚情况的进一步细化以及与其它方法的结合将是必要的。

REFERENCES（参考文献）

[1] Chou PY, Fasman G. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, 1974, **13**: 211 – 222.

[2] Fan J. Prediction of protein secondary structure with a reliability score estimated by local sequence clustering. *Protein Engineering*, 2003, **16**(9): 651 – 657.

[3] Figureau, MA Soto, J Toha. A pentapeptide-based method for protein secondary structure prediction. *Protein Engineering*, 2003, **16**(2): 103 – 107.

[4] Hu XZ(胡秀珍), Xiao Y(肖奕). Correlation of codon and double base in the protein structure. *Journal of Inner Mongolia Polytechnic University*(内蒙古工业大学学报(自然科学版)), 2003, **22**(3): 170 – 174.

[5] Wang SY(王守源), Li XQ(李晓琴), Luo LF(罗辽复). Amino acid classification based on protein secondary structures. *Acta Scientiarum Naturalium Universitatis Inner Mongol*(内蒙古大学学报(自然科学版)), 2002, **33**(4): 423 – 427.

[6] Li WJ(李炜疆), Song JN(宋江宁). How much protein secondary structural information do synonymous codons carry. *Acta Biophysica Sinica*(生物物理学报), 2003, **19**(1): 41 – 46.

[7] Zheng WM. Clustering of amino acids for protein secondary structure prediction. *Journal of Bioinformatics and Computational Biology*, 2004, **2**(2): 333 – 342.

[8] Wang YF(王艳芬), Huang JT(黄积涛). Recognition and analysis to the assembly of amino acid residues in protein structures. *Journal of Tianjin University of Technology*(天津理工大学学报), 2005, **21**(3): 30 – 33.

[9] Wu D(吴迪), Miao Q(缪强), Wang ZL(王志林). Nearest neighbor residue effect of amino acids and the secondary structure conformations of di-peptide in proteins. *Acta Chimica Sinica*(化学学报), 2005, **63**(12): 121 – 1126.

[10] Mohammed Ouali, Ross D King. Cascaded multiple classifiers for secondary structure prediction. *Protein Science*, 2000, **9**: 1162 – 1176.