生物物理学报 第十八卷 第一期 二〇〇二年三月
ACTA BIOPHYSICA SINICA Vol.18 No.1 Mar.2002

不同结构的蛋白编码基因的密码子偏性研究

顾万君, 马建民, 周 童, 孙 啸, 陆祖宏
(东南大学吴健雄实验室,江苏 南京 210096)

摘要:利用聚类分析方法,对两类具有不同三级结构的75个蛋白的编码基因的密码子使用偏性进行了分析。75个基因样本序列按照对应蛋白的三级结构被很清晰的分成了两类,从而发现密码子的使用与蛋白质的三级结构有很大的相关性。这一重要结果证实了DNA的一维信息中蕴含着蛋白质的三级结构信息。

关键词:密码子使用;蛋白质三级结构;聚类分析;mRNA
中图分类号:Q617 文献标识码:A 文章编号:1000-6737(2002)01-0081-06

在密码子表中,每个氨基酸至少对应1种密码子,最多有6种对应的密码子,此即密码子的简并性。在基因中,同义密码子的使用并不是完全均匀的[1,2]。不同物种、不同生物体的基因密码子使用存在着很大的差异[3-5]。进一步的研究表明,不同的密码子使用模式的形成可能与基因的GC含量有关[1,2]。在一些单细胞生物如*Escherichia coli*、*Saccharomyces cerevisiae*中,高表达的基因密码子的使用偏性一般比较大,这主要是由于基因的碱基组成和mRNA翻译时的tRNA选择两大因素造成的[5-7]。最近的一些研究表明,基因密码子的使用与基因编码的蛋白的结构和功能有关。基因密码子的使用与基因表达的生理功能有着密切的联系[8,9]。mRNA中的稀有密码子的使用与蛋白质结构域的连接区和规则二级结构单元的连接区有关,翻译速率在连接区会降低[10];同时,在表达具有不同二级结构的蛋白质时,mRNA区段的翻译速率有所不同[11]。这些均说明蛋白质折叠方式与mRNA序列之间存在一定的相关性[12]。丁达夫等[13]的研究显示,在*Escherichia coli*中,基因的密码子使用与蛋白的二级结构相关性并不是很大,而在哺乳动物中也存在着一些相关性。Oresic等[14]也发现,*Escherichia coli*中AAC编码的Asn一般位于β-折叠的C末端,而在人蛋白质中,GAU编码的Asp多数位于α-螺旋的N端。最近,我们应用聚类分析方法研究了哺乳动物MHC基因密码子的使用偏性,发现了不同功能的MHC分子对应mRNA的密码子使用偏性有显著不同[15]。在本文中,我们报道了蛋白的三级结构与密码子使用概率的关系,发现用聚类分析方法可以很清晰地将具有不同三级结构蛋白质的编码基因分成不同的类,而具有相似三级结构蛋白的编码基因则大致聚在同一类中,从而证明基因密码子偏性与蛋白质三级结构间具有密切的相关性。不同基因的密码子的使用与所编码的蛋白的三级结构的相关性的研究,可以用来对未知基因的密码子使用偏性情况进行预测,从而用来进行蛋白mRNA检测型基因芯片的探针设计。同时,这一相关性还可用来对未知空间结构的蛋白质进行空间结构的预测,从而预测蛋白质的功能,这在后基因组时代有着非常重要的研究意义。

1 样本和方法

1.1 选取的基因样本

我们从蛋白库PRINTS(http://bioinf.man.ac.uk/dbbrowser/PRINTS)中随机选取了所有具有4-二硫化物核心单位蛋白指纹(4-DISULPHCORE)和肾上腺素受体蛋白指纹(ADRENERGICR)的两组蛋白。所谓蛋白指纹,是指蛋白质中一系列具有特定氨基酸组成特征与功能的保守性蛋白模式。对于特定的蛋白结构域,它可能由多个蛋白模式组成。这些蛋白模式在三维空间上可能是相连的,能够实现特定的功能,但是在一级结构上它们可能相差很远,也有可能相互重叠。表1列出了选取的两组蛋白对应的编码基因序列,包括每一条序列的Genebank登记号、编码基因的序列

收稿日期:2001-03-29
基金项目:国家重大基础研究项目(G1998051200),国家自然科学基金资助项目(69525102)
作者简介:顾万君,博士研究生,电话:(025)3792245,E-mail:wanjungu@seu.edu.cn

Table 1 Genes coding for proteins with two different fingerprints

ID	Access No.	Length	Gene definition
1	AF023459	4287	Haliotis rufescens lustrin A mRNA, complete cds.
2	AF036161	2712	Trichuris trichiura putative porin precursor (TT95) mRNA.
3	AJ000221	399	Sus scrofa domestica partial mRNA for porcinewhey acidic protein.
4	D83667	564	Sus scrofa mRNA for preproSPAI-2, complete cds.
5	J00801	414	Rat whey acidic protein mRNA.
6	J00802	414	Rat whey phosphoprotein mrna clone.
7	L12144	2031	Chicken KAL mRNA, complete cds.
8	S60085	2040	ADMLX = putative adhesion molecule [human, mRNA].
9	M57446	348	Porcine antileukoproteinase mRNA, complete cds.
10	M97252	2043	Homo sapiens Kallmann syndrome (KAL) mRNA, complete cds.
11	S77395	375	CE4 = epididymal secretory protein [dogs, epididymides, mRNA].
12	U03890	366	Salvelinus fontinalis antileukoproteinase precursor, mRNA.
13	U67854	789	Salvelinus fontinalis ovulatory protein-2 precursor, mRNA.
14	X04470	399	Human mRNA for antileukoprotease (ALP) from cervix uterus.
15	X04503	399	Human SLPI mRNA fragment for secretory leucocyte protease inhibitor.
16	X60299	2043	H.sapiens KALIG-1 mRNA for neural cell adhesion.
17	X93037	225	M.musculus mRNA for WDNM1 protein.
18	Z18538	354	H.sapiens encoding skin-derived antileukoproteinase.
19	AF013261	1368	Homo sapiens alpha 1A adrenergic receptor isoform 4 mRNA.
20	AF031431	1401	Mus musculus alpha 1A-adrenergic receptor mRNA, complete cds.
21	AF193027	1215	Mus musculus beta-3b adrenergic receptor splice variant (B3bar).
22	AF200596	1257	Macaca mulatta beta 3 adrenergic receptor (B3AR) mRNA.
23	AF200597	1206	Canis familiaris beta 3 adrenergic receptor (B3AR) mRNA.
24	D25235	1401	Human mRNA for alpha1C adrenergic receptor, complete cds.
25	D32201	1290	Human mRNA for alpha 1C adrenergic receptor isoform 3.
26	D32202	1500	Human mRNA for alpha 1C adrenergic receptor isoform 2.
27	J03019	1434	Human beta-1-adrenergic receptor mRNA, complete cds.
28	J03024	1257	Rat beta-adrenergic receptor mRNA, complete cds.
29	J03853	1386	Human kidney alpha-2-adrenergic receptor mRNA, complete cds.
30	J04084	1548	Syrian golden hamster alpha-1B adrenergic receptor mRNA.
31	J05426	1401	Cow alpha-1C-adrenergic receptor mRNA, complete cds.
32	L20333	483	Mouse alpha-1A adrenergic receptor homologue mRNA, partial cds.
33	L31771	1686	Rat alpha 1a/d adrenergic receptor mRNA, complete cds.
34	L31772	1719	Human alpha-1a/d adrenergic receptor mRNA, complete cds.
35	L31773	1560	Human alpha-1B-adrenergic receptor mRNA, complete cds.
36	L31774	1401	Human alpha-1C-adrenergic receptor mRNA, complete cds.
37	L38905	1248	Macaca mulatta beta-2 adrenergic receptor (B2AR) mRNA.
38	M14379	1452	Turkey beta-adrenergic receptor mRNA, complete cds.
39	M15169	60	Human beta-2-adrenergic receptor mRNA,complete cds.
40	M32061	1362	Rat alpha-2B-adrenergic receptor (RNG-alpha-2) mRNA, complete cds.
41	M58316	1377	Rat alpha-2B-adrenergic receptor mRNA, complete cds.
42	M60654	1683	Rat alpha-1A-adrenergic receptor mRNA, complete cds.
43	M60655	1548	Rat alpha-1B adrenergic receptor mRNA, complete cds.
44	M74716	1203	Rat beta-adrenergic receptor mRNA, complete cds.
45	M76446	1506	Human alpha-A1-adrenergic receptor mRNA, complete cds.
46	S53290	72	beta 3-adrenergic receptor {3' region} [mice, white adipose tissue].
47	S53291	54	beta 3-adrenergic receptor {3' region} [human, SK-N-MC cells].
48	S56481	1203	beta 3-adrenergic receptor {spliced version} [rats, colonic tissue].
49	S70782	1719	Homo sapiens alpha adrenergic receptor subtype alpha 1a mRNA.
50	S73473	1203	beta 3-adrenergic receptor [rats, Genomic/mRNA, 1383 nt].
51	S76001	429	alpha 1c-adrenoceptor subtype [human, brain, mRNA Partial, 432 nt].
52	S80044	1689	alpha 1d-adrenergic receptor [mice, brain, mRNA, 1902 nt].
53	S80219	117	alpha 1b-adrenergic receptor [mice, brain, mRNA Partial, 120 nt].
54	U02569	1401	Human alpha1C adrenergic receptor mRNA, complete cds.
55	U03864	1719	Human adrenergic alpha-1a receptor protein mRNA, complete cds.
56	U03865	1563	Human adrenergic alpha-1b receptor protein mRNA, complete cds.
57	U04310	1410	Didelphis virginiana alpha-2c adrenergic receptor mRNA.
58	U07126	1401	Rattus norvegicus alpha1c adrenergic receptor mRNA, complete cds.
59	U13368	1401	Rattus norvegicus Sprague Dawley alpha1C-adrenergic receptor mRNA.
60	U13977	1287	Meleagris gallopavo adrenergic beta-4c receptor mRNA, complete cds.
61	U53185	327	Sus scrofa beta 2 adrenergic receptor mRNA, partial cds.
62	U56425	1110	Sus scrofa beta 1 adrenergic protein mRNA, partial cds.
63	U64032	1731	Oryctolagus cuniculus alpha 1d adrenoceptor mRNA, complete cds.
64	U79031	1353	Rattus norvegicus alpha2D adrenergic receptor mRNA, complete cds.
65	U81982	1401	Oryctolagus cuniculus alpha 1a-adrenoceptor mRNA, complete cds.
66	X03804	60	Hamster mRNA for beta-adrenergic receptor.
67	X04827	1242	Human mRNA for brain beta-adrenergic receptor.
68	X51585	1548	Rat mRNA for the alpha-1B adrenergic receptor.
69	X67213	375	B.taurus mRNA for adrenergic receptor beta 2.
70	X70811	1227	H.sapiens mRNA for beta 3 adrenergic receptor.
71	X85961	1218	B.taurus mRNA for beta-3 adrenergic receptor.
72	X94608	1248	C.familiaris mRNA for beta2-adrenergic receptor.
73	Y09213	1158	X.laevis mRNA for beta-1-adrenergic receptor.
74	Y12738	1545	M.musculus mRNA for alpha-1B adrenergic receptor.
75	Z86037	1257	B.taurus mRNA for beta-2-adrenergic receptor.

长度、基因说明。为了分析方便，本文对每个基因都给定一个对应的序列号(ID)。在选取的基因中，第 1 条到第 18 条是含有 4 – 二硫化物核心单位蛋白指纹的蛋白编码基因，第 19 条到第 75 条是含有肾上腺素受体蛋白指纹的蛋白编码基因。

1.2　基因密码子相对使用概率

与密码子使用概率计算相关的统计量有密码子使用的频次、绝对密码子使用概率、相对密码子使用概率、基因的密码子适应系数和基因的有效密码子数量等。其中密码子使用的相对概率（relative synonymous codon usage，RSCU）是指对于某一特定的密码子在编码对应氨基酸的同义密码子间的相对概率[16]。它去除了氨基酸组成对密码子使用的影响，如果密码子的使用没有偏性的话，该密码子的 RSCU 值等于 1。当某一密码子的 RSCU 值大于 1 时，代表该密码子为使用相对较多的密码子，反之亦然。由于它计算简便，而且直观地反映了密码子使用的偏性，在本文的密码子使用偏性分析中都用它作为衡量的标准。在这里，相对使用概率 $RSCU_{ij}$ 的计算公式为：

$$RSCU_{ij} = \frac{obs_{ij}/aa_{ij}}{1/n}$$

其中 $RSCU_{ij}$ 为表 1 中第 i 个基因序列、第 j 个密码子的相对使用概率，obs_{ij} 代表密码子 j 在基因 i 中出现的次数，aa_{ij} 代表密码子 j 编码的氨基酸在基因 i 编码的蛋白中出现的次数，n 代表与密码子 j 同义的密码子的个数。

1.3　系统聚类分析

系统聚类的基本思想是：先将所有的样品各自看成一类，规定样品之间的距离和类与类之间的距离。刚开始由于每个样品自成一类，所以类与类之间的距离与样品与样品之间的距离是相等的。选择距离最小的一对并成一个新类，接着计算新类与其他类之间的距离，再将距离最近的两类合并。这样循环每次减少一类，直至所有的样品合并成一类为止。

在对基因密码子使用概率进行聚类分析的过程中，我们将每一条基因作为一个对象，将密码子的相对使用概率统计值作为变量。由于编码蛋氨酸(M)的密码子 AUG 和编码色氨酸（W）的密码子 UGG 的 RSCU 值始终为 1，再除去三个不编码氨基酸的终止子，我们取基因的 59 个密码子的 RSCU 值对基因的密码子使用偏性进行分析。

基因间的距离我们规定为基因密码子相对使用概率的欧拉平方距离。对于基因 i 与基因 j，它们的密码子使用距离 d_{ij} 的计算公式为：

$$d_{ij} = \sum_{k=1}^{59} (RSCU_{ik} - RSCU_{jk})^2$$

对于类与类间的距离，我们采用离差平方和法（Ward 法）。该法定义类与类之间的距离时，当两类 G_p 与 G_q 合并成新类 G_r 时，任一类 G_i 与 G_r 之间的距离 D_{ir} 为：

$$D_{ir}^2 = \frac{n_i + n_p}{n_i + n_r} D_{ip}^2 + \frac{n_i + n_q}{n_i + n_r} D_{iq}^2 - \frac{n_i}{n_i + n_r} D_{pq}^2$$

其中 n_i 指第 i 类中样本的个数，D_{ij} 代表第 i 类和第 j 类的距离。

利用离差平方和法进行聚类的效果比较好，因为它能使同类样品之间的离差平方和最小，而类与类之间的离差平方和最大。

本文中我们利用 SPSS 9.0 系统多元分析软件来实现对基因密码子使用偏性的聚类。

2　结果和讨论

2.1　结果

对表 1 所列的 75 条基因序列计算出了所有的编码区的密码子相对使用概率，并根据 59 个密码子的相对使用概率（RSCU）进行聚类分析，分析结果列入表 2。图 1 给出了 75 条基因样本序列的分层聚类树状图，图 1 中给出的样本序列间的密码子使用距离已经标准化。

Table 2　Cluster results of genes in Table 1

ID	类号	ID	类号	ID	类号	ID	类号	ID	类号
1	I	16	I	31	II	46	I	61	II
2	I	17	I	32	II	47	I	62	II
3	I	18	I	33	II	48	II	63	II
4	I	19	II	34	II	49	II	64	II
5	I	20	II	35	II	50	II	65	II
6	I	21	II	36	II	51	II	66	I
7	I	22	II	37	II	52	II	67	II
8	II	23	II	38	II	53	II	68	II
9	I	24	II	39	I	54	II	69	II
10	I	25	II	40	II	55	II	70	II
11	I	26	II	41	II	56	II	71	II
12	I	27	II	42	II	57	II	72	II
13	I	28	II	43	II	58	II	73	II
14	I	29	II	44	II	59	II	74	II
15	I	30	II	45	II	60	II	75	II

从图 1 可以看出，75 条基因样本序列被很清晰的分成了 I、II 两大类，其中基因 1 – 7、9 – 18 和基因样本 39、46、47、66 聚成了类 I，其他基因样本聚成类 II。在我们选取的基因样本中，基因样本 1 – 18 是编码具有 4 – 二硫化物核心单位蛋白指纹蛋白的基因序列，而基因样本 19 – 75 是编码肾上腺素受体

蛋白指纹蛋白的基因序列。从聚类的结果可以看出,除了基因样本 8 被聚到了类 II 中,基因样本 39、46、47、66 被聚到了类 I 中,其余的基因样本的聚类结果与我们所选取的基因编码蛋白的三级结构完全吻合。因此,从聚类分析的结果可以看出,蛋白质的三级结构与基因的密码子使用偏性有着明显的相关性。

基因样本 39、46、47、66 并没有与具有相似三级结构的蛋白编码基因一样聚在类 II 中,而是聚在了类 I 中,并且它们聚成了类 I 中的小类 I – B。从表一中可以看到,基因样本 39、46、47、66 的序列长度分别为 60、72、54、60。这些基因序列的序列长度太小,每一个密码子的使用频次非常小,实际密码子相对使用概率的统计值并不能代表基因密码子使用的统计特性。因此,这些基因样本在聚类过程中没有与具有相似结构的蛋白编码基因聚在类 II 中是完全可以解释的。同时,基因样本 8 在聚类过程中也出现了偏差,这可能是所有具有 4 – 二硫化物核心单位蛋白指纹蛋白的编码基因密码子使用的一个特例。

2.2 密码子使用与蛋白三级结构相关

遗传中心法则指出,基因首先转录成 mRNA,然后经过 mRNA 的翻译,最后翻译出的肽段通过空间卷曲折叠而形成具有生物功能的蛋白质。

首先,mRNA 的翻译速率与下列三个因素有关:(a) 密码子的使用频率以及密码子与反密码子的相互作用,(b) 密码子的上下文关系,(c) mRNA 不同区域的二级结构 [2]。不难看出,这三个影响 mRNA 翻译速率的因素都与密码子的使用相关。因而,mRNA 的翻译速率与基因的密码子使用有着很大的相关性。

同时,翻译速率与蛋白质的空间结构有着密切的关系。一方面,新生肽链在被释放出核糖体前便开始了肽链的折叠过程,这一过程被称作共翻译折叠。翻译速率的差异会通过共翻译折叠过程影响到蛋白质的空间构象 [12]。另一方面,新生肽链与核糖体的相互作用对于肽链的折叠也有着很大的影响[10,11]。

因此,基因密码子的使用偏性对蛋白结构的影响主要体现在从 mRNA 翻译到蛋白质这一阶段。不同三级结构的蛋白的编码基因通过使用不同的密码子偏性,影响到 mRNA 不同区域的翻译速度,从而影响翻译过程中新生肽链的折叠,进而影响翻译出的蛋白的空间构象。

从功能来说,蛋白的功能与蛋白的空间结构有着密切的联系。大量的密码子使用研究表明,密码子的使用与基因的功能有一定的相关性。我们在样本选取时选取的蛋白结构单位是蛋白指纹,它在很大程度上也是一种蛋白功能单位。因此,基因密码子使用与蛋白三级结构之间的相关性同时也证实了基因密码子使用与基因功能之间的相关性。

2.3 密码子使用与物种的相关性

相当多的研究表明,不同物种的基因密码子使用模式各不相同 [3-5]。从图 1 和表 1 可以看出,类 II – AA 中都为大鼠属 (Genus Rattus)、鼠属 (Genus Mus) 基因,类 II – AB 中大多为人属 (Homo sapiens) 基因,它们的物种大致相同。

从图 1 可以看出,聚在类 I 中的基因对应的物种包括鲍属 (Haliotis)、鞭虫属 (Trichuris)、猪属 (Sus)、大鼠属 (Genus Rattus)、鼠属 (Genus Mus)、大马哈鱼属 (Oncorhynchus) 和人属 (Homo sapiens) 等,这些物种中不但有脊椎动物如猪属、鼠属、人属等,而且还有软体动物如鲍属和原生动物如鞭虫属。聚在类 II 中的基因对应的物种包括人属 (Homo sapiens)、大鼠属 (Genus Rattus)、鼠属 (Genus Mus)、牛属 (Bovine)、兔属 (Rabbit)、猪属 (Sus) 和猕猴属 (Macaca),这些物种都属于脊椎动物。虽然类 I 中的基因对应的物种很广,包括从比较原始的原生动物到较高等的软体动物再到更高等的哺乳动物,但这些物种的基因聚在了同一类中。同时,虽然类 I 和类 II 中选取的基因都包括人属、大鼠属、鼠属和猪属的基因序列,但在聚类的过程中相同种属的基因还是分别聚在了两类中,并不一定聚在同一类中。从上面的分析可以看出,基因对应的物种并不完全决定基因的分类。但在类 I 和类 II 中,每一类中基因编码的蛋白的三级结构大致相似。因此,我们可以认为基因密码子的使用最主要的还是与蛋白的三级结构和功能相关,而物种对密码子使用的影响相比较于蛋白的结构和功能来说并不是很明显。

2.4 研究意义

基因的密码子使用与蛋白三级结构的相关性的重要结果,说明了基因序列的一维信息中蕴含着蛋白的三级结构和蛋白功能的信息,同时它还具有很大的应用价值。一方面,对于未知结构的蛋白片段,可以通过研究它的编码基因的密码子使用,预测出它的三级结构和功能;另一方面,对于未知编码基因的蛋白片断,可以通过它的结构预测出它的编码基

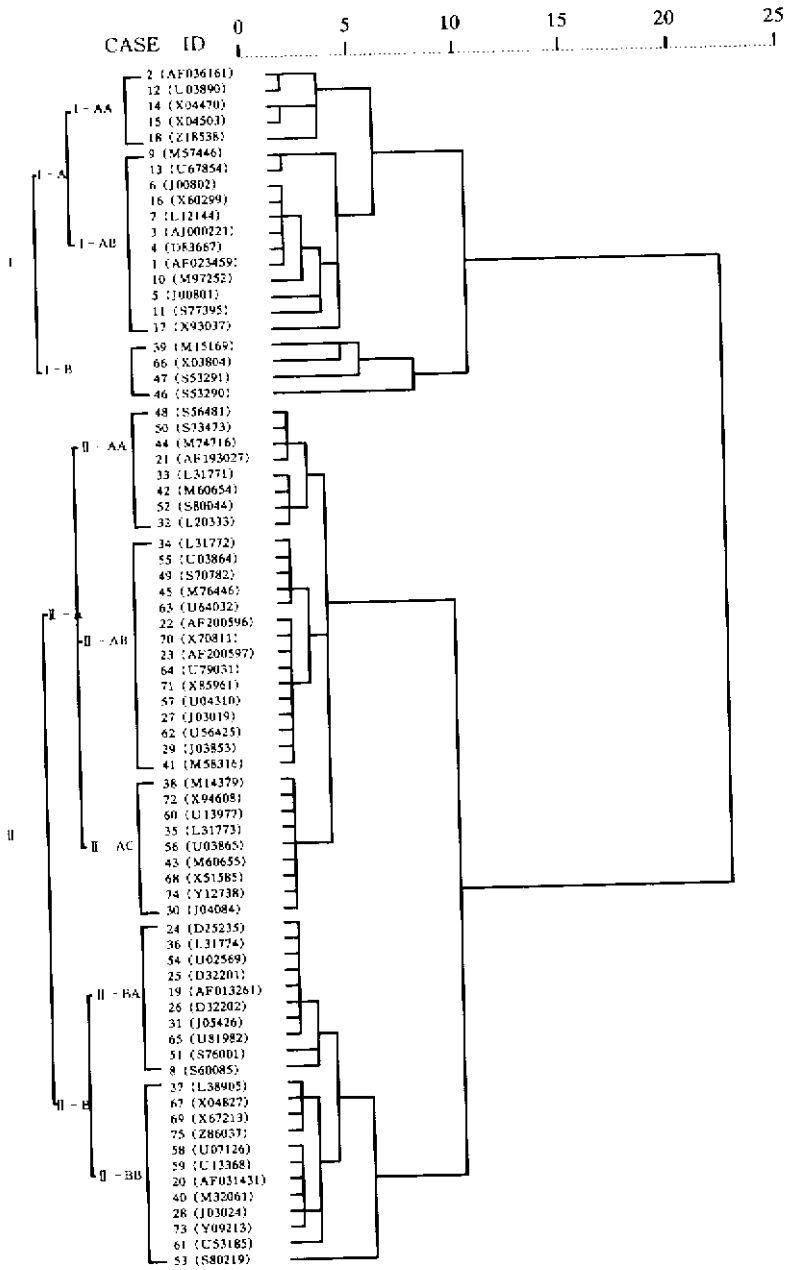

Fig.1 Cluster analysis dendrogram of RSCU values for genes coding for proteins with two different fingerprints

因的密码子使用，从而对蛋白序列进行反翻译，设计出检测编码基因的寡核苷酸探针或者检测型基因芯片的探针阵列。

参考文献：

[1] Ghosh T. Studies on codon usage in Entamoeba histolytica[J]. *International Journal of Parasitology*, 2000,30: 715－722.

[2] Karlin S, Mrazek J. What drives codon choices in human genes[J]? *Journal of Molecular Biology*, 1996,262: 459－472.

[3] Grantham R, Gautier C, Gouy M, et al. Codon catalog usage and the genome hypothesis[J]. *Nucleic Acids Research*, 1980,8:49r－62r.

[4] Grantham R, Gautier C, Gouy M, et al. Codon catalog usage is a genome strategy modulated for gene expressivity[J]. *Nucleic Acids Research*, 1981,9:43r－74r.

[5] Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity[J]. *Nucleic Acids Research*, 1982,

10:7055 – 7074.

[6] Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system[J]. *Journal of Molecular Biology*, 1981,151:389 – 409.

[7] Sharp P, Tuohy T, Mosurski K. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes[J]. *Nucleic Acids Research*, 1986,14: 5125 – 5143.

[8] Helene C, Frederique L, Michel C, et al. Codon usage and gene function are related in sequences of Arabidopsis thaliana[J]. *Gene*, 1998,209:GC1 – GC38.

[9] Richard J, Lin K, Tan T. A functional significance for codon third bases[J]. *Gene*, 2000,245:291 – 298.

[10] Thanaraj T, Argos P. Ribosome – mediated translational pause and protein domain organization[J]. *Protien Science*, 1996,5:1594 – 1612.

[11] Thanaraj T, Argos P. Protein secondary structural types are differentially coded on messenger RNA[J]. *Protien Science*, 1996,5:1973 – 1983.

[12] 柳树群, 刘次全. mRNA 的序列、结构以及翻译速率与蛋白质结构的关系[J]. 动物学研究, 1999,20:457 – 461.

[13] Xie T, Ding D. The relationship between synonymous codon usage and protein structure[J]. *FEBS Letters*, 1998,434:93 – 96.

[14] Oresic M, Shalloway D. Specific correlations between relative synonymous codon usage and protein secondary structure[J]. *Journal of Molecular Biology*, 1998,281: 31 – 48.

[15] 周童, 马建民, 顾万君,等. 哺乳动物 MHC 密码子使用概率的聚类分析[J]. 东南大学学报(自然科学版), 2001,31(2): 1 – 5.

[16] Comeron J, Aguade M. An evaluation of measure of synonymous codon usage bias[J]. *Journal of Molecular Evolution*, 1998,47:268 – 274.

CODON USAGE IN GENES CODING FOR PROTEINS WITH DIFFERENT TERTIARY STRUCTURES

GU Wan – jun, MA Jian – min, ZHOU Tong, SUN Xiao, LU Zu – hong

(Chien – Shiung Wu Laboratory, Southeast University, Nanjing 210096, China)

Abstract: Based on gene codon usage distance 75 genes coding for proteins with two different fingerprints were clustered, to analyze the relationship between gene codon usage and protein tertiary structure. It was found that 75 genes coding for proteins with different tertiary structure can be clearly classified into two groups according to the difference of codon usage bias, which suggests that codon usage is lighly corrclated to protein tertiary structure. The relationship between codon usage and protein tertiary structure is firstly found in the research of codon usage, which proves that gene sequence contains protein tertiary structure information.

Key Words: Codon usage; Protein tertiary structure; Cluster analysis; mRNA