

文章编号 :1009 - 038X(2001)01 - 0001 - 05

同义密码子用语与氨基酸序列 上下文的关联

李炜疆

(无锡轻工大学生物工程学院 ,江苏无锡 214036)

摘 要 :对大肠杆菌基因组的全部编码区序列进行了统计分析 ,发现同义密码子用语与氨基酸序列上下文关系密切 ,尤其是紧邻氨基酸的影响最为显著 .讨论了该结果的结构生物学意义及其在基因工程领域可能的应用 .

关键词 :同义密码子用语 ;氨基酸序列 ;统计关联 ;大肠杆菌

中图分类号 :Q 61

文献标识码 :A

Correlation between Synonymous Codon Usage and the Amino Acid Sequence Contexts

LI Wei-jiang

(School of Biotechnology , Wuxi University of Light Industry , Wuxi 214036 , China)

Abstract : All coding sequences in the genome of *Escherichia coli* were statistically investigated , which turned out that the synonymous codon usage is closely related to the amino acid sequence contexts . The influences of the nearest neighboring amino acids are the most significant . The structural biological meaning and possible applications of the finding were discussed .

Key words : synonymous codon usage ; amino acid sequence ; statistical correlation ; *Escherichia coli*

关于同义密码子用语的研究近年来引起了广泛的关注^[1] ,一个主要的原因是它与基因表达有密切关系 .在保持氨基酸序列不变的前提下 ,通过同义密码子调整适当改造编码 DNA 序列 ,可使基因表达产物量有成百倍的提高^[2] .同义密码子用语和基因表达水平关系的理论研究^[3,4]可以指导基因改造 ,具有重要的应用价值 .最近关于同义密码子用语和蛋白质二级结构关系的研究结果表明^[5~9] ,简并密码子的使用可能携带着蛋白质空间结构的信息 ,并且这种关联随生物种属而异^[6] .对于这种关

联存在与否也有不同意见^[10] ,但无论如何 ,这都是一个既有理论意义又有实用价值的重要问题 .

密码子相对使用频率受多种因素的影响 ,需要仔细分析 .比如密码子使用与蛋白质结构的关联问题 ,如果统计发现某种关联存在 ,就要首先查明这种关联是否是由密码子用语直接导致 ,还是由于序列因素使得密码子用语产生偏斜而间接导致 .要解决这个问题 ,首先要弄清密码子使用频率与序列间的关系 .由于蛋白质二级结构与紧邻和次紧邻二肽的关系非常紧密^[11] ,所以作者首先研究紧邻氨基酸

收稿日期 2000 - 05 - 30 ,修订日期 2000 - 11 - 23 .

基金项目 国家自然科学基金项目资助课题(39760027) .

作者简介 李炜疆(1964 -) ,男 ,陕西榆林人 ,理学博士 ,教授 .

对密码子相对使用频率的影响.

1 材料与方法

1.1 材料

为避免由于不同生物的同义密码子用语不同而导致的统计误差,作者选取大肠杆菌基因组的全部编码序列作为研究对象,核酸序列数据取自 GenBank,全基因组序列存取号(accession number)为 U00096.在该数据集中,有 4 千多个编码序列,共约 4 Mb 碱基.在统计时,不考虑起始密码和终止密码.

1.2 方法

令 a, b 表示氨基酸, γ 表示密码子.如果在氨基酸序列中,从某个氨基酸 a 向 C 端方向数起的第 m 个氨基酸为 b 就称 a 的相对位置 m 处为 b .若 $m < 0$ 则氨基酸 b 处于 a 的 N 端一侧.定义为

$$I_m(a|b) = \sum_{\gamma \in a} p_m^a(\gamma|b) \ln \frac{p_m^a(\gamma|b)}{p^a(\gamma)} \quad (1)$$

其中

$p_m^a(\gamma|b)$ 表示在氨基酸序列中 a 的相对位置 m 处为氨基酸 b 的条件下, a 的密码子 γ 的相对使用频率; $p^a(\gamma)$ 表示在整个数据集中,氨基酸 a 的密码子 γ 的无条件相对使用频率; $\sum_{\gamma \in a}$ 表示对氨基酸 a 的所有密码子求和.

根据定义,

$$\sum_{\gamma \in a} p_m^a(\gamma|b) = \sum_{\gamma \in a} p^a(\gamma) = 1. \quad (2)$$

可以证明,

$$I_m(a|b) \geq 0, \quad (3)$$

当且仅当

$$p_m^a(\gamma|b) = p^a(\gamma), \text{ 对所有 } \gamma \in a \quad (4)$$

时 (3) 式中的等号成立.也就是说,只有在相对位置 m 处的氨基酸 b 对于氨基酸 a 的密码子相对使用频率没有影响时,才有 $I_m(a|b) = 0$; 否则,这种影响越大,有条件使用频率和无条件使用频率的差异就越大, $I_m(a|b)$ 的值也越大.因此, $I_m(a|b)$ 可以看作是氨基酸 b 对氨基酸 a 的密码子相对使用频率影响程度的度量.

2 结果与分析

讨论紧邻 ($m = \pm 1$) 和次紧邻 ($m = \pm 2$) 氨基酸对密码子相对使用频率的影响 $I_m(a|b)$, 结果见表 1. 表中第一行为要观察密码子相对使用频率的氨

基酸(受影响氨基酸, a), 第一列为要观察氨基酸的相邻氨基酸(施加影响氨基酸, b). 与每个施加影响氨基酸 b 相对应有 4 行数据, 自上而下分别为 $I_{-2}(a|b)$, $I_{-1}(a|b)$, $I_1(a|b)$ 和 $I_2(a|b)$, 表示 b 在不同位置上对 a 的影响程度. 为方便观察, 所有 $I_m(a|b) < 0.025$ 的微小数值均未列出.

分析表 1 数据, 可以得出如下明显结论:

1) 从施加影响氨基酸的角度看: 绝大多数次紧邻氨基酸 ($m = \pm 2$) 对密码子使用频率的影响均可忽略, 极个别略有影响, 如甲硫氨酸 (Met) 对处于其 N 端一侧的次紧邻位置的几种氨基酸的密码子使用频率有些影响, 但程度很弱; 大部分氨基酸对其紧邻氨基酸密码子的使用有明显影响, 尤其是对其 N 端紧邻的影响最大, 相比之下, 对 C 端紧邻的影响要明显小一些; 对 N 端紧邻氨基酸密码子使用频率影响最大的氨基酸是色氨酸 (Trp), 它对其 N 端紧邻的所有氨基酸的密码子使用都有很强的影响; 类似的氨基酸还有酪氨酸 (Tyr), 但影响程度相对较弱.

2) 从受影响氨基酸的角度看: 赖氨酸 (Lys) 密码子用语受相邻氨基酸的影响最小, 除此以外的所有氨基酸的密码子用语都会受到与其相邻的至少一种氨基酸 (如 Trp) 的明显影响; 亮氨酸 (Leu), 丙氨酸 (Ala), 甘氨酸 (Gly) 和苏氨酸 (Thr) 是最易受影响的氨基酸, 其密码子使用频率在多种相邻氨基酸的影响下发生显著改变.

3 讨 论

通过以上分析不难看出, 密码子用语在氨基酸序列不同的上下文中有明显差异, 尤其是氨基酸对其 N 端紧邻密码子用语的影响最为显著. 这种对 N 端紧邻氨基酸密码子用语的强烈调制很可能与编码区对特定碱基顺序片段的回避现象有关, 是一种序列水平的现象, 并不一定与蛋白质结构有直接关系. 这在研究密码子用语与蛋白质结构关系时应予以足够重视.

在基因工程中, 经常可以通过将密码子调整为最适密码子显著提高表达水平^[2,4], 而作者的研究结果表明, 这样的调整必须考虑氨基酸序列的上下文关系. 例如, 大肠杆菌中, Thr 的总体密码子相对使用频率为 $ACU = 0.166$, $ACC = 0.434$, $ACA = 0.132$, $ACG = 0.268$; 而当其 C 端紧邻 Trp 时, 密码子相对使用频率则变为 $ACU = 0.017$, $ACC = 0.769$, $ACA = 0.122$, $ACG = 0.092$.

表 1 相邻氨基酸对密码子相对使用频率的影响

Tab.1 The influences of neighboring amino acids on the relative codon usage																		
	Phe	Leu	Ile	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Arg	Gly
Phe	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.08	-	.06	-	-	.06	.14	-	-	.13	-	-	-	-	-	.03	.08
	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	.09	-	.03	.03
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Leu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	.05	-	-	-	-	-	-	-
	-	.11	-	.12	.09	.06	.17	.10	.05	.04	-	-	-	-	-	-	.04	.07
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ile	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.06	-	.07	-	-	.05	.11	-	-	.05	-	-	-	-	-	.03	.06
	.08	-	-	-	-	-	.05	-	-	-	.03	.03	-	.03	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.03	-
Met	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	.07	-	-	-	-	-	-	-	-	-	.06
	.13	.10	-	.11	.08	.08	-	.09	-	.06	-	.03	-	-	.04	-	.14	.08
	.03	-	.03	.04	-	-	-	-	-	-	-	-	-	.03	-	.03	.03	-
Val	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	.03	.13	-	.05	.03	-	-	-	-	-	.10	-	-	-	.03	-	.07	.07
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ser	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	.03	.18	-	.09	.04	.06	.07	.14	.03	.04	.03	.03	-	.07	-	-	.04	.04
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pro	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	.07	-	-	-	-	-	-	-	-	-	-	-	-	-
	.08	.06	-	.05	.07	.04	.23	.06	-	.03	-	-	-	-	-	.03	-	.06
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Thr	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	.03	.03	.04	.04	-	-	-	.05	-	.04	-	.06	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ala	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.5
	.03	.09	-	-	-	.06	-	-	-	-	.08	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tyr	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	.03	.03	-	.03	.03	-	.08	.04	.07	.04	-	.10	-	.06	-	-	-	.03
	-	.13	-	.16	.11	.13	.12	.17	-	-	-	-	-	-	-	-	.05	.03
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 1(续) 相邻氨基酸对密码子相对使用频率的影响

Tab.1(Continued) The influences of neighboring amino acids on the relative codon usage

	Phe	Leu	Ile	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Arg	Gly
His	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	.03	-	-	-	-	-	-	-	-	-	-	-	-	-	.04	-	-	.11
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Gln	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.03	-	-	-	-	-	-	-	-	.04	-	-	-	-	-	.04	.05
	.05	.13	.04	.07	.05	.10	.04	.06	.04	.04	-	-	.05	-	.06	-	-	.22
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Asn	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	.03	.03	-	-	.05	-	-	-	-	-	-	-	-	-	.03
	-	.08	-	-	.04	-	.10	.05	-	-	.07	-	-	-	-.03	.06	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lys	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.7	-	-	-
	-	-	-	.04	.07	-	.03	-	-	-	.06	-	.03	-	-	-	.03	-
	-	-	.03	.03	-	-	-	-	-	-	-	-	-	-	-	.03	.03	-
Asp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.03	-	-	.04	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.20	-	.07	.07	-	.06	.08	-	.03	.10	-	.04	-	.08	-	.04	.08
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Glu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.5	-	-	-
	.03	.16	-	.03	.07	-	.03	.04	-	-	.03	-	.05	-	.07	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cys	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.03	-	-	.03	.04	-	-	-	-	.09	-	-	-	-	-	-	-
	-	.07	-	.07	.10	.14	.25	.28	-	.04	-	.04	-	.04	.05	-	.05	.05
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Trp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.04	.03	.03	.08	-	.04	.05	-	-	-	-	-	-	-	-	.07	.11
	.46	.04	.25	.32	.21	.17	.29	.24	.42	.44	-	.43	-	.70	.05	.16	.20	.22
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Arg	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	0.5	-	-	-	-	-	-	-
	.08	.07	-	.04	-	-	-	-	-	-	-	-	.03	-	.04	-	-	.12
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Gly	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	.04	-	-	-	.03	.03	.06	-	-	-	-	-	-	-	-	-	-
	-	.08	-	.06	.07	.04	.04	-	-	.07	-	-	.03	-	.08	-	-	.12
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

受影响最强烈的则是天冬氨酸(Asp),其总体密码子相对使用频率为 $GAU = 0.628$, $GAC = 0.372$; 而当其 C 端紧邻 T_{ip} 时的密码子相对使用频率则为 $GAU = 0.068$, $GAC = 0.932$.

在此情况下 , $_{Asp}$ 的原最适密码子 GAU 变成了罕见密码子 ,而原本非偏好密码子 GAC 则成为强烈偏好密码子 . 在设计高效表达编码区序列时 ,这也是一个不容忽视的重要因素 .

参考文献 :

- [1] NAKAMURA Y , WEDA K , WADA Y , *et al.* Codon usage tabulated from the international DNA sequence databases [J]. **Nucl Acid Res** , 1996 , 24 : 214 ~ 215 .
- [2] MOHSEN A-W A , VOCKLEY J. High level expression of an altered cDNA encoding human isovaleryl-CoA dehydrogenase in Escherichia coli [J]. **Gene** , 1995 , 154 : 159 ~ 166 .
- [3] POWELL J , MORIYAMA E. Evolution of codon usage bias in Drosophila [J]. **Proc Natl Acad Sci USA** , 1997 , 94 : 7784 ~ 7790 .
- [4] LI H , LUO L F. The Relation between Codon Usage , Base Correlation and Gene Expression Level in Escherichia Coli and Yeast [J]. **J theor Biol** , 1996 , 181 : 111 ~ 124 .
- [5] 解 涛 , 丁达夫 . 同义密码子用语与蛋白质结构的关系 [J]. 生物物理学报 , 1998 , 14 : 296 ~ 302 .
- [6] ORESIC M , SHALLOWAY D. Specific correlations between relative synonymous codon usage and protein secondary structure [J]. **J Mol Biol** , 1998 , 281 : 31 ~ 48 .
- [7] ADZHUBEI A A , ADZHUBEI I A , KRASHENINNIKOV I A , *et al.* Non-random usage of ' degenerate ' codons is related to protein three-dimensional structure [J]. **FEBS Lett** , 1996 , 399 : 78 ~ 82 .
- [8] THANARAJ T A , ARGOS P. Protein secondary structural types are differentially coded on messenger RNA [J]. **Protein Sci** , 1996 , 5 : 1973 ~ 1983 .
- [9] THANARAJ T A , ARGOS P. Ribosome-mediated translational pause and protein domain organization [J]. **Protein Sci** , 1996 , 5 : 1594 ~ 1612 .
- [10] BRUNAK S , ENGELBRECHT J. Protein Structure and the Sequential Structure of mRNA (Helix and Sheet Signals at the Nucleotide Level) [J]. **Proteins : Struct Funct Genet** , 1996 , 25 : 237 ~ 252 .
- [11] 李炜疆 . 蛋白质序列与二级结构的统计关联 [J]. 内蒙古大学学报 , 1991 , 22 : 248 ~ 251 .