46 卷 6 期
2006 年 12 月 4 日

微 生 物 学 报
Acta Microbiologica Sinica

46(6):895 ~ 899
4 December 2006

大肠杆菌 mRNA 编码区长度、形成二级结构倾向与密码子偏好性的关系

王 侃[1],刘次全[1,2,3,4*],曹 槐[1],陈雪峰[1]

([1] 云南大学现代生物学中心 昆明 650091) ([2] 中国科学院动物研究所 昆明 650223)
([3] 北京大学理论生物学中心 北京 100871) ([4] 大连理工大学高科技研究院 大连 116024)

摘 要:从 GenBank 获得大肠杆菌 K-12 MG1655 株的全基因组序列,计算了与基因密码子偏好性相关的多个参数(Nc、CAI、GC、GC3s),对其 mRNA 编码区长度、形成二级结构倾向与密码子偏好性之间的关系进行了统计学分析,发现虽然翻译效率(包括翻译速度和翻译精度)是制约大肠杆菌高表达基因的密码子偏好性的主要因素,同时,mRNA 编码区长度及其形成二级结构的倾向也是形成这种偏好性的不可忽略的原因,而且对偏好性有一定程度的削弱。另外对 mRNA 编码区形成二级结构倾向的生物学意义进行了讨论分析。

关键词:大肠杆菌;密码子偏好性;mRNA 编码区;mRNA 二级结构;翻译效率
中图分类号:Q78 **文献标识码**:A **文章编号**:0001-6209(2006)06-0895-05

大肠杆菌的密码子使用具有简并性(codon degenerate),即一个氨基酸由两个以上的密码子编码。但是,大肠杆菌并不是同等地使用每个密码子来编码氨基酸的,对某一个氨基酸而言,大肠杆菌通常倾向于更多地使用它所对应的同义密码子中的一种或数种,而不是均衡地或随机地使用每一种同义密码子来编码它,即大肠杆菌的密码子使用具有偏好性(codon usage bias)。对于密码子具有偏好性的解释,目前获得较为广泛认可的是"突变-选择平衡"假说(mutation-selection balance)[1]。它认为,由于选择压力的存在,生物体倾向于选用最优密码子(optimal codon)来编码氨基酸,但由于突变的发生,仍会有非最优密码子的存在,不同物种的基因组的密码子偏好情况主要就是在这两个力量的动态平衡中形成的。

关于选择压力,目前取得较多共识的是对翻译效率(包括翻译速度和翻译精度)的选择压力,尤其对高度表达的基因更是如此[2,3]。这个观点认为,不同的同义密码子在翻译时分别使用不同的 tRNA,而不同 tRNA 在细胞中的含量是不同的,那些对应着更高含量 tRNA 的密码子在核糖体中翻译时能用更短的时间与正确的 tRNA 相匹配,因此提高了翻译的速度和精度,这些和高含量 tRNA 相匹配的密码子就是最优密码子。基因的表达水平越高,所受的对翻译效率的选择压力就越大,因此密码子的偏好性也就越高;而表达水平较低,则偏好性也相应较低。因而现在经常把基因的密码子偏好程度作为该基因的表达水平的标志。但是,对翻译效率的选择压力,并不是造成密码子偏好性的全部原因,基因的密码子偏好程度也不是完全和其表达水平有关,已有工作表明,低偏好性的基因也可以有较高的表达水平[4]。本文试图对这个问题作更深一步的探讨。

长久以来,对 mRNA 结构的关注主要集中在非编码区,但近来已有研究认为,mRNA 编码区总体来讲有较低的自由能,具有形成二级结构的倾向[5,6]。因此,本文也将讨论 mRNA 编码区的长度、二级结构倾向以及基因的密码子偏好性之间的关系。

1 材料和方法

1.1 材料

从 GenBank 取得大肠杆菌 K-12 MG1655 株全基因组序列(Accession No. U00096),从中获得 4242 个编码蛋白质的基因(包括已确认的和假设的),去掉其中序列不完整的基因和内部含有终止密码子的基因。为减小小基因中存在随机变异的影响,序列长度短于 150 个核苷酸(nt)的基因被去掉。计算了剩余 4199 个基因的密码子偏好性指数 CAI、Nc,以及其它指数 GC、GC3s,所用软件为 codonW(John Peden,可获得:http://codonw.sourceforge.net/Readme.html);统计了每个基因的长度;统计分析所用的软件为

基金项目:国家自然科学基金重大研究计划项目(90208018,90303018);国家自然科学基金数学天元基金重点项目(A0324101)
* 通讯作者。Tel:86-871-5033496;Fax:86-871-5036373;E-mail:liucq@ynu.edu.cn
作者简介:王 侃(1977 -),男,云南人,硕士研究生,现从事结构分子生物学研究。E-mail:maiersi2004@163.com
收稿日期:2005-12-29;接受日期:2006-02-22;修回日期:2006-05-05

spss for windows 13.0。

1.1 密码子适应指数(Codon adaption index, CAI)

该指数以一组具高表达水平的基因为参考,测量某一个基因的密码子偏好情况和这些高表达基因密码子偏好情况的接近程度,如果一个基因完全使用高表达基因中所用的密码子,则其 CAI 值为 1。目前这个指数已被广泛用来预测基因的表达水平[7]。

1.2 有效密码子数(Effective Number of Codon, Nc)

CAI 测量的是某个基因所用的密码子与高表达基因所用密码子的接近程度。和 CAI 不同,Nc[8] 测量的是某个基因的密码子偏好程度,如果一个基因平均使用每一个密码子,则其 Nc 为 61,如果一个基因只使用每组同义密码子中的一个,则其 Nc 为 20。理论上讲,一个具有低 CAI 的基因也可以同时具有低 Nc 值,换句话说,该基因具有较强的密码子偏好性,只不过其偏向的并不是高表达基因所用的密码子。

1.3 GC 和 GC3s

GC 测量的是基因中 G 和 C 的含量。GC3s 则计算密码子第三个碱基中出现 G 或 C 的频率。一般认为这两个因素对基因的密码子选择有重要影响。

1.4 mRNA 编码区二级结构形成倾向

Katz 等[6] 的工作表明,包括大肠杆菌在内的许多生物的 mRNA 编码区具有较低的自由能,他们据此认为 mRNA 编码区有形成二级结构的倾向,这种结构倾向可能并不是要形成具体的、保守的结构,而是形成一种非特定位点的结构倾向。本文拟从密码子偏好性和 mRNA 编码区长度的角度对这种结构倾向进行分析。大肠杆菌是共转录翻译的,即在转录完成前,翻译就已经开始,因此计算 mRNA 的整体结构没有意义。在翻译过程中,相邻核糖体间间隔的 mRNA 核苷酸数大约是 150nt,可以想象,在这样长度的区间里,mRNA 完全有可能形成局部的二级结构,而且这种二级结构应该是动态变化的,随着核糖体的移动,当核糖体接近时,该结构解开,而远离核糖体的地方又折叠形成新的结构。这里我们以 50nt 为可能形成局部二级结构的核苷酸长度的近似(以其它长度 25nt、100nt、125nt 也可以得到同样的结果[6])。因此,以 50nt 为窗口,沿 mRNA 编码区滑动,每次移动 10nt,对 mRNA 进行"扫描",每次移动截取得到一条 50nt 的 mRNA 片段,这样每个 mRNA 可以获得一系列 50nt 的片段,用 RNAstructure 4.1[9,10] 软件预测这些片段的二级结构,取最优结构时的自由能值,对这些自由能值求算术平均,即得到该 mRNA 50nt 的平均自由能。我们认为,平均自由能较低的 mRNA 形成二级结构的倾向也较强,而平均自由能高的则有避免或较少形成二级结构的趋势,因此,本文以平均自由能(E)作为 mRNA 形成二级结构强弱的标志来加以分析。

2 结果

Reis 等[4] 通过分析大肠杆菌全基因组的 CAI 和表达水平间的关系,将大肠杆菌基因组分为 3 组:组Ⅰ基因的 CAI 值范围为:0.307 ~ 0.849,共 1398 个基因,该组基因有高平均 CAI 值,同时也有高平均表达水平,这符合人们的通常认识,即高表达的基因所受的对翻译效率的选择压力越大,因此密码子偏好程度越高;组Ⅱ基因的 CAI 值范围为:0.123 ~ 0.404,共 2164 个基因,有中度的平均 CAI 值和 3 组中最低的平均表达水平;组Ⅲ基因的 CAI 值范围为:0.110 ~ 0.441,共 727 个基因,这是最让人出乎意料的,它有 3 组中最低的平均 CAI 值,但是却有高的平均表达水平。这是一个很有意义的现象,我们将在本文中按照这种分类对大肠杆菌全基因组的偏好性等特征进行分析。

2.1 对 3 组基因的 mRNA 编码区长度与其结构倾向的方差分析

对 3 组基因的 E 值及 mRNA 编码区长度进行一元方差分析。由图 1-A、C、E 可以看出,3 组基因的 E 值均呈正态分布,对之进行方差齐性检验,发现各组方差不等(Levene Statistic 为 47.35, $P < 0.01$),因此分别用 Welch[11]、Brown-Forsythe[12] 两种方法做方差分析,两种方法均显示 3 组的 E 值有显著差异(Welch 统计量:923.743, $P < 0.01$;Brown-Forsythe 统计量:1148.939, $P < 0.01$),进一步用 Tamhane 方法[13] 做多重比较,结果见表 1。

由图 1-B、D、F 可以看出,3 组基因的 mRNA 编码区长度均不呈正态分布,对之进行方差齐性检验,发现各组方差不等(Levene Statistic 为 33.90, $P < 0.01$),由于方差分析方法对正态分布的要求不严格[14,15],因此仍分别用 Welch、Brown-Forsythe 两种方法做方差分析,两种方法均显示 3 组的 mRNA 编码区长度有显著差异(Welch 统计量:77.556, $P < 0.01$;Brown-Forsythe 统计量:78.672, $P < 0.01$),进一步用 Tamhane 方法做多重比较,结果见表 1。

对 3 组的 mRNA 编码区长度与其形成二级结构

倾向的方差分析表明,组Ⅰ和组Ⅱ的 50nt 平均自由能之间没有显著差异,但组Ⅲ的 50nt 自由能比组Ⅰ、组Ⅱ的要高出很多。而 3 组间的 mRNA 编码区长度均有显著差异,平均长度由大到小依次是组Ⅰ、组Ⅱ、组Ⅲ(表 1,图 1)。组Ⅲ是 3 组中最为特殊的,它的平均长度比另外两组均明显要短,同时它形成二级结构的倾向也最低;组Ⅰ和组Ⅱ形成二级结构的倾向没有显著差异,同时它们的平均长度虽有差异,但明显没有与组Ⅲ的差距大。

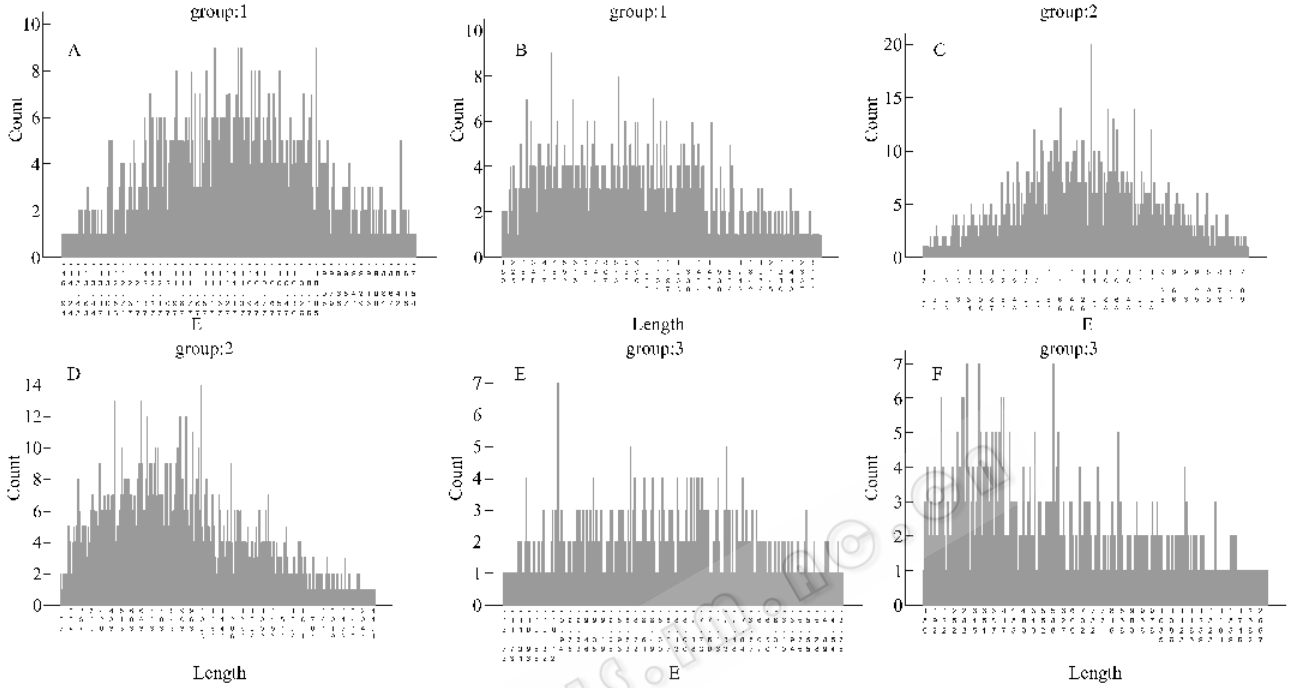

图 1 (A、C、E)组Ⅰ、Ⅱ、Ⅲ mRNA 编码区 50nt 自由能的分布图;(B、D、F)组Ⅰ、Ⅱ、Ⅲ mRNA 编码区长度的分布图

Fig.1 (A,C,E) Distribution of free energy of 50nt mRNA coding regions in group Ⅰ, Ⅱ and Ⅲ respectively; (B,D,F) Distribution of length of mRNA coding region in group Ⅰ, Ⅱ and Ⅲ respectively.

表 1 对 3 组基因的 mRNA 编码区长度与其结构倾向的方差分析

Table 1 One way ANOVA of mRNA coding region length and their tendency of folding

Parameter	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	P
E	Ⅰ	Ⅱ	0.06951	0.04464	>0.05
		Ⅲ	-2.92943	0.07372	<0.01
	Ⅱ	Ⅰ	-0.06951	0.04464	>0.05
		Ⅲ	-2.99894	0.07168	<0.01
	Ⅲ	Ⅰ	2.92943	0.07372	<0.01
		Ⅱ	2.99894	0.07168	<0.01
Length	Ⅰ	Ⅱ	121.760	21.947	<0.01
		Ⅲ	360.129	29.036	<0.01
	Ⅱ	Ⅰ	-121.760	21.947	<0.01
		Ⅲ	238.369	25.562	<0.01
	Ⅲ	Ⅰ	-360.129	29.036	<0.01
		Ⅱ	-238.369	25.562	<0.01

E: Means of free energy.

2.2 Nc 与各参数的关系

为了了解造成大肠杆菌密码子偏好的原因,对 Nc 的负对数作多元逐步回归统计分析(表 2)。对 Nc 的负对数而不是直接对 Nc 做回归分析仅仅是为了使结果更为直观:-lg(Nc)越大,基因的密码子偏好程度越高。

表 2 表明,各变量对全基因组及组Ⅰ的-lg(Nc)变化的影响程度较高(校正的 R^2 分别为 0.593,0.655),但对组Ⅱ及组Ⅲ-lg(Nc)变化的影响程度则很小(校正的 R^2 分别只有 0.325,0.224)。对全基因组和组Ⅰ、组Ⅱ而言,在几个变量中对-lg(Nc)影响程度最高的都是 CAI,而其它变量的影响度较低。E 和 lg(length)的影响程度较低,但都达到了统计学显著性,而且都与其呈负相关。与其它不同的是,在组Ⅲ中,对-lg(Nc)的影响,GC3s 超过了 CAI,另外,E 及 GC 造成影响的显著程度未能达到 0.05。

在各因子中,对大肠杆菌基因密码子偏好性影响较大的是 CAI,尤其在组Ⅰ中,CAI 说明了偏好性的较大部分,这说明,在高表达基因中,影响密码子偏好状况的主要原因是对翻译效率的要求,但在表达水平较低的基因中,翻译效率不是影响密码子偏好状况的主要因素,另外,mRNA 编码区长度及其形成二级结构倾向的影响虽然不及翻译效率,但却是不能忽略的因素,而且从它们与偏好性呈负相关来

看,它们对为满足翻译效率而形成的偏好性有一定程度的削弱。

表 2 全基因组及其分为 3 组后的-lg(Nc)的多元逐步回归统计分析

Table 2 Multiple linear regression of -lg(Nc) in genome and three group

Group	Model	Regression coefficients	t-test	P	Adjusted R^2
Genome	CAI	0.405 ± 0.006	68.564	< 0.01	0.593
	GC3s	0.174 ± 0.013	13.608	< 0.01	
	GC	− 0.434 ± 0.034	− 12.630	< 0.01	
	E	− 0.008 ± 0.001	− 11.146	< 0.01	
	lg(length)	− 0.007 ± 0.002	− 3.379	< 0.01	
Group Ⅰ	CAI	0.437 ± 0.009	50.082	< 0.01	0.655
	GC3s	0.193 ± 0.016	12.422	< 0.01	
	E	− 0.004 ± 0.001	− 5.047	< 0.01	
	lg(length)	− 0.013 ± 0.003	− 4.145	< 0.01	
Group Ⅱ	CAI	0.405 ± 0.019	21.681	< 0.01	0.325
	GC3s	0.233 ± 0.017	13.636	< 0.01	
	E	− 0.008 ± 0.001	− 8.411	< 0.01	
	GC	− 0.256 ± 0.046	− 5.538	< 0.01	
	lg(length)	− 0.256 ± 0.003	− 2.331	< 0.05	
Group Ⅲ	GC3s	− 0.277 ± 0.023	− 12.092	< 0.01	0.224
	lg(length)	− 0.040 ± 0.006	− 6.720	< 0.01	
	CAI	0.219 ± 0.037	5.973	< 0.01	

E: Means of free energy; lg(length): Logarithm of length.

2.3 mRNA 编码区长度与其形成二级结构倾向的关系

对 mRNA 编码区长度及其 50nt 平均自由能进行相关分析,发现二者之间呈现出弱负相关:Pearson 相关,$r = -0.239$,$P < 0.01$;Spearman 相关,$R = -0.232$,$P < 0.01$。说明当编码区长度较长时,其 50nt 平均自由能较低,即其形成二级结构的倾向较强。Katz 及其合作者的工作表明[6],mRNA 编码区具有形成非特定位点的遍布 mRNA 编码区的二级结构倾向,这种倾向虽不强,但有较高的统计学显著性。

根据当前的研究,mRNA 的降解可能主要是由核糖核酸酶 E 起始的,其具体的机制目前尚不清楚,一般认为核糖核酸酶 E 的切割发生在某些具有特殊序列特征的位点[16]。我们认为,mRNA 编码区形成局部的二级结构有助于掩蔽核糖核酸酶 E 作用的位点,降低 mRNA 被过早降解的可能,尤其 mRNA 长度越长,其暴露在核糖核酸酶 E 下,遭受其攻击的可能性越大,越需要形成局部的二级结构。有研究表明,在室温下一个单链 mRNA 的寿命只有 0.1ms,但形成发夹后其寿命则可以延长到 10s[17]。但这种二级结构又不能太强,否则会影响核糖体翻译的进行,这也可以解释为何 mRNA 编码区长度与其形成二级结构的倾向二者之间的相关性不强。局部二级结构的形成应当是既能阻止 E 酶对靶点的识别、攻击,但又不影响核糖体的正常阅读、翻译。

3 讨论

同义密码子偏好是在复杂因子作用下形成的,对翻译效率的适应是其中一个重要的因子,组Ⅰ的基因就是这一观点的明证,在这类基因中,CAI 和表达水平的变化是相一致的,高表达水平对应着高的 CAI 值,而且 CAI 的变化说明了 Nc 变化的相当大一部分,表明在这些基因中密码子的偏好主要是在对翻译效率的选择压力下形成的。但组Ⅲ则不同,这类基因有 3 组中最低的平均 CAI 值,却有着高的表达水平,而且 CAI 对 Nc 的说明度较低。组Ⅱ的情况则介于这二者之间。这些结果均说明,在大肠杆菌基因组中,对翻译效率的适应并不是造成其密码子偏好分布的唯一原因,尤其在组Ⅲ中,也许突变偏好或是其它尚未发现的因素才是其密码子偏好形成的主要原因。

基因的转录和翻译是个复杂的过程,当转录形成 mRNA 开始,它就受着被降解的威胁,如何在翻译完成前避免被过早降解掉,这对基因也是一个重要的选择压力。阻止核糖核酸酶 E 的识别与攻击是防止 mRNA 被过早降解的方法之一。前边我们已经提到,单链 mRNA 的寿命比发夹状态下 mRNA 的寿命要明显短很多,再结合 mRMA 编码区有较低自由能的证据,可以合理地推测当 mRMA 进行翻译期时,在核糖体之间的 mRNA 片段(即未受核糖体保护的部分)正是通过形成局部的、较弱的二级结构来避免被过早降解的。尤其越长的 mRNA,越需要形成这样的二级结构来防止被过早降解,在 mRNA 编码区长度和形成二级结构倾向之间存在着一定程度的负相关对此是一个旁证。

这样对形成二级结构的要求势必将影响到同义密码子的选择,即选择那些更加容易形成一定二级结构的密码子。mRNA 编码区长度和形成二级结构倾向确实影响了大肠杆菌密码子偏好的一部分,虽然影响度不高,但却是不能忽略的。对此我们认为,施加在大肠杆菌上的选择压力是一个交织的、错综复杂的(有时甚至可能是相矛盾的)压力网络,对比 mRNA 编码区长度和形成二级结构的倾向,对翻译效率的选择是一个更加强大的压力,因此一般情况下大肠杆菌的密码子选择首先要满足后者,然后才及其它。

致谢:感谢云南大学现代生物学中心刘海桑博士对此文的建议。

参 考 文 献

[1] Bulmer M. The selection-mutation-drift-theory of synonymous codon usage. *Genetics*, 1991, **129**: 897 – 907.

[2] Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol*, 1981, **146**: 1 – 21.

[3] 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 id optimal for the *E. coli* translational system. *J Mol Evol*, **151**: 389 – 409.

[4] Reis DM, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Research*, 2003, **31**(23): 6976 – 6985.

[5] Seffens W, Digby D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research*, 1999, **27**(7): 1578 – 1584.

[6] Katz L, Burge BC. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Research*, 2003, **13**: 2042 – 2051.

[7] Sharp PM, Li W. The codon adaptation index-a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Research*, 1986, **15**: 1281 – 1295.

[8] Wright F. The effective number of codons' used in a gene. *Gene*, 1990, **87**: 23 – 29.

[9] Mathews DH, Disney MD, Childs JL, *et al*. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences USA*, 2004, **101**: 7287 – 7292.

[10] Mathews DH, Sabina J, Zuker M, *et al*. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 1999, **288**: 911 – 940.

[11] Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 1938, **29**: 350 – 362.

[12] Brown MB, Forsythe AB. Robust tests for the equality of variances. *J Amer Statist Assoc*, 1974, **69**: 364 – 367.

[13] Tamhane AC. A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, 1979, **74**: 471 – 480.

[14] 薛 薇. 统计分析与 SPSS 的应用. 第一版. 北京:中国人民大学出版社, 2002.

[15] 东方人华等. 统计基础和 SPSS 11.0. 第一版. 北京:清华大学出版社, 2004.

[16] Coburn GA, Mackie GA. Degradation of mRNA in *Escherichia coli*: an old problem with some new twists. *Prog Nucleic Acid Res Mol Biol*, 1999, **62**: 55 – 108.

[17] Tinoco I Jr. Force as a useful variable in reactions: unfolding RNA. *Annu Rev Biophys Biomol Struct*, 2004, **33**: 363 – 385.

Insight into relationship among mRNA coding region length, folding tendency and codon usage bias in *Escherichia coli*

WANG Kan[1], LIU Ci-quan[1,2,3,4*], CAO Huai[1], CHEN Xue-feng[1*]

([1] *Modern Biology Research Center, Yunnan University, Kunming* 650091, *China*)

([2] *Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming* 650223, *China*)

([3] *Center of Theoretical Biology, Peking University, Beijing* 100871, *China*)

([4] *High-technological Academy, Dalian University of Technology, Dalian* 116024, *China*)

Abstract: *Escherichia coli* has been regarded as a model organism in the study of codon usage bias. Most studies in this organism regarding this topic have focused their mind on translation efficiency (translation speed and translation precision). However, some genes with low codon usage bias and high expression can not be explained by translation efficiency. And some evidence of local RNA secondary structure in coding region of *Escherichia coli* genes have been given. All of these need explanation from a new point of view. The genomic sequence of *Escherichia coli* K-12 MG 1655 was obtained from GenBank. Several parameter (Nc, CAI, GC, GC3s) of 4199 genes on codon usage bias were computed. Folding tendency of mRNA coding region was represented by its average energy of 50 nucleotides. And then, relationship among mRNA coding region length, folding tendency and codon usage bias in *Escherichia coli* were studied. It is argued that though translation efficiency is a primary cause that shapes the codon usage bias of highly expressed genes in *Escherichia coli*, mRNA coding region length and folding tendency are also unneglectable factor for codon usage bias and weakening of bias. In addition, biological significance of folding tendency in mRNA coding regions was discussed.

Keywords: *Escherichia coli*; Codon usage bias; mRNA coding region; Secondary structure of mRNA; Translation efficiency

Foundation item: National Nature Science Fund of China (90303018, 90208018); The Mathematics Tian Yuan Foundation (A0324101)

* Corresponding author. Tel: 86-871-5033496; Fax: 86-871-5036373; E-mail: liucq@ynu.edu.cn

Received: 29 December 2005/Accepted: 22 February 2006/Revised: 5 May 2006