Nucleic Acids Research, 1996, Vol. 24, No. 13 **2525–2527**

Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency

Gabriel Gutiérrez, Lorenzo Márquez and Antonio Marín*

Departamento de Genética, Universidad de Sevilla, Apartado 1095, E-41080 Sevilla, Spain

Received March 25, 1996; Revised and Accepted May 21, 1996

ABSTRACT

The variation in base composition at the three codon sites in relation to gene expressivity, the latter estimated by the Codon Adaptation Index, has been studied in a sample of 1371 *Escherichia coli* genes. Correlation and regression analyses show that increasing expression levels are accompanied by higher frequencies of base G at first, of base A at second and of base C at third codon positions. However, correlation between expressivity and base compositional biases at each codon site was only significant and positive at first codon position. The preference for G-starting codons as gene expression level increases is discussed in terms of translational optimization.

INTRODUCTION

The expression level of *Escherichia coli* protein-coding genes is accompanied by a remarkable change in the usage of synonymous codons. Highly expressed genes have a strong preference for a subset of codons, while lowly expressed genes have a more uniform pattern of codon usage (1,2). Changes in codon usage accompanying higher gene expression in *E.coli* are a result of selection for translational efficiency, as shown by the positive correlation between optimal codons and tRNA abundance (1). Lobry and Gautier in a recent paper (3) have widened the effects of translational constraints by finding that the variation in amino acid composition among a sample of 999 *E.coli* genes correlates with their expression levels as estimated by the Codon Adaptation Index or CAI (4).

There is not an obvious link between the variation in third codon base choices, upon which is based the CAI computation, and the variation at first and second codon positions, which is related to amino acid composition. That is why we have undertaken the present study to analyze the variation in individual base frequencies at each codon site in a sample of *E.coli* genes ranked according to their CAI value.

DATA AND METHODS

A sample of 1371 protein coding genes longer than 100 codons was retrieved from the *E.coli* database ECD (5). The sample contains only those entries with the system number EGxxxx (structural gene), genes lacking standard initiation or termination codons and genes containing internal in frame termination codons were removed. For each gene we computed: (i) the CAI value as defined in (4), (ii) the relative frequency of each amino acid (with distinction, in the case of sextets, between the fraction encoded by the quartet and the duet), and (iii) the base composition at each codon site.

First, we investigated the relationship between CAI value and amino acid composition, and second between CAI value and base composition by codon site. For this purpose, we constructed for each amino acid a scatter diagram, in which every gene is represented by one point whose coordinates are the CAI value (x-axis), and the relative content of the amino acid under analysis (y-axis). The same procedure was used to analyze the relationship between CAI value (x-axis) and the frequency of each base at each codon site (y-axis). In every diagram, Pearson's and Kendall's rank correlation coefficients were computed; a regression line, y = ax + b, was calculated only when the Pearson's correlation was statistically significant.

RESULTS

Variation in amino acid composition

The variation profile of amino acid abundance encoded by *E.coli* genes in relation to CAI value can be appreciated in Table 1, where the different amino acids are sorted out according to their regression line slope. Three amino acid groups can be distinguished: (i) amino acids which increase its frequency as CAI value does (Lys, Glu, Gly, Asp, Val, Ala, Asn and Met), (ii) amino acids which decrease its abundance as CAI value rises (Leu, Ser, Gln, Arg, Trp, Pro, His, Cys and Phe), and (iii) amino acids whose abundance is not affected by CAI value variation (Thr, Tyr and Ile). This pattern closely agrees with that based upon correspondence analysis (3). It is interesting to note the negative correlation shown by the amino acids Leu, Arg and Ser, otherwise abundant in *E.coli* proteins; such a trend is caused by the decrease of their respective duet components, while their quartet components remain insensitive to CAI variation.

Variation in base composition by codon site

We address specifically here the changes at the nucleotide level by remarking the consistent pattern displayed in Table 1: the

[*] To whom correspondence should be addressed

amino acids which increase their abundance as CAI value increases are all encoded by codons starting with a purine; more specifically, all the amino acids with G-starting codons increase their frequency as CAI value does.

Table 1. Correlation coefficients (r Pearson's) and slopes obtained by plotting the frequency % of each amino acid against the CAI value

AA	Codons	r	Slope
Lys	AAR	+0.359 **	+6.57
Glu	GAY	+0.242 **	+4.85
Gly	GGN	+0.264 **	+4.83
Asp	GAR	+0.271 **	+4.27
Val	GTN	+0.192 **	+3.29
Ala	GCN	+0.143 **	+3.27
Asn	AAY	+0.059 *	+0.75
Met	ATG	+0.060 *	+0.60
Cys	TGY	–0.117 **	–1.02
His	CAY	–0.166 **	–1.77
Arg2	AGR	–0.389 **	–1.80
Pro	CCN	–0.143 **	–1.93
Trp	TGG	–0.237 **	–2.13
Gln	CAR	–0.184 **	–2.87
Ser2	AGY	–0.358 **	–3.54
Leu2	TTR	–0.695 **	–9.67
Thr	ACN	+0.012 NS	
Tyr	TAY	+0.007 NS	
Arg4	CGN	–0.027 NS	
Ile	ATH	–0.051 NS	
Leu4	CTN	–0.052 NS	
Ser4	TCN	–0.055 NS	
Phe	TTY	–0.063 NS	
Arg6	CGN+AGR	–0.122 **	–2.29
Ser6	TCN+AGY	–0.277 **	–4.14
Leu6	CTN+TTR	–0.426 **	–10.71
	RNY	+0.556 **	+27.72
	RNR	–0.044 NS	
	YNR	–0.494 **	–17.61
	YNY	–0.156 **	–4.69

R = A or G; Y = C or T; H = A, C or T not G; N = A, C, G or T.
*Statistical significance P<0.05, ** idem P <0.001, NS = non-significant.

The nucleotide frequency changes accompanying increasing CAI are shown in Table 2 as correlation and regression coefficients; these have been computed in scatter diagrams obtained by plotting the base frequency at each codon site against CAI value. To give an idea of the heterogeneity of base frequencies, the genes pertaining to the top and bottom 10% of the CAI distribution were extracted, and the average base frequencies at each codon site were computed in these two extreme classes (Table 3).

At the first codon position, the most conspicuous change consists in the increasing frequency of base G while base T decreases. At the second codon position, base A increases while bases G and T decrease and base C shows no variation. At the third codon position, a strong increase in the frequency of base C is observed at the expense of a decrease in bases A, G, and T.

Table 2. Correlation coefficients (r) and slopes obtained by plotting the base % frequency at each codon position against the CAI value

	Codon position					
	First		Second		Third	
	r	slope	r	slope	r	slope
A	+0.055 *	+1.87	+0.243 **	+12.04	–0.230 **	–8.87
C	–0.208 **	–8.05	+0.031 NS	–	+0.454 **	+21.62
G	+0.504 **	+20.44	–0.168 **	–4.38	–0.207 **	–10.15
T	–0.429 **	–14.26	–0.197 **	–8.57	–0.061 *	–2.61

*Statistical significance P <0.05, ** idem P <0.001, NS = non-significant.

Table 3. Mean base frequency (%) and standard deviation (SD) by codon site in the first (L) and last (H) 10% of CAI distribution

		Codon Position					
		First		Second		Third	
		Mean	SD	Mean	SD	Mean	SD
A	L	27.30	5.93	28.14	6.85	22.87	5.82
	H	26.17	2.95	32.19	3.95	17.73	3.64
C	L	23.42	6.17	21.56	3.88	21.65	6.18
	H	21.81	3.75	22.64	3.27	30.86	4.42
G	L	30.45	4.62	18.63	3.99	26.31	7.69
	H	39.64	4.28	16.89	3.16	24.28	4.63
T	L	18.83	4.33	31.68	5.82	29.18	7.40
	H	12.39	3.30	28.28	3.25	27.14	4.54

These results merge as the well known general pattern of overrepresentation of RNY codons, whose count exceeds those of RNR, YNR and YNY codons (6). Actually, the overrepresentation of RNY codons strongly correlates to increases in CAI, while the frequencies of RNR, YNR and YNY are negatively correlated to CAI (bottom of Table 1).

Base composition bias by codon site

We have shown that variation in base frequencies occurs to different extent at each codon site. A second point concerns the departure from base equifrequency at each codon site. The measurement of this departure is particularly suitable in the *E.coli* genome, whose overall base composition is fairly equifrequent and no marked regional compositional variation exits.

To disclose the relationship between the degree of base usage bias at each codon position and the variation in CAI value, we have computed for each codon position in every gene an index, *fi* (*i* = a, c, g, t) previously described (7). The *fi* index is defined as the frequency of the *i* base divided by the expected frequency if all bases are equifrequent (i.e. the relative frequency of base *i* multiplied by 4). By definition, the mean of *f*a, *f*c, *f*g and *f*t equals unity. The standard deviation (*sigma-f*) of the *fi* may be a good measurement of the degree of base utilization bias, being larger in heavily biased codon positions than in less-biased ones. As an example, let us consider the gene ECOAAS (GenBank accession

L14681), where A-, C-, G- and T- starting codons appear 173, 193, 246 and 107 times. Thus, *f*a is 173/[(173+193+246+107)/4] = 0.96, *f*c =1.07, *f*g =1.37 and *f*t = 0.60; the standard deviation (*sigma-f*) of *f*a, *f*c, *f*g and *f*t is 0.27.

We have computed the correlation coefficient between *sigma-f* and CAI value. Interestingly, only the first codon position compositional bias is positively correlated to CAI (r = 0.442, P <0.001), while compositional biases at second and third codon positions do not (r = –0.0003, P = 0.992; and r = –0.002, P = 0.929, respectively). To illustrate these results, the average value of *sigma-f* at each codon site has been computed in the genes pertaining to the four quartiles of the CAI distribution (Table 4). It can be seen that only at first codon position is there a clear increasing trend in compositional bias, while no variation is observed at second and third codon positions. Thus, with regard to nucleotide composition the main changes which accompany increasing gene expression level do occur at first codon positions.

Table 4. *Sigma-f* averages and standard deviations (SD) in the four quartiles of the CAI distribution

	Codon Position					
	First		Second		Third	
	Mean	SD	Mean	SD	Mean	SD
Q1	0.27	0.08	0.26	0.10	0.25	0.11
Q2	0.31	0.08	0.25	0.09	0.25	0.09
Q3	0.33	0.08	0.25	0.08	0.26	0.09
Q4	0.38	0.09	0.26	0.07	0.25	0.08

The CAI cutpoints were 0.295, 0.354 and 0.428.

DISCUSSION

Lobry and Gautier discussed (3) that proteins encoded by genes with high CAI values are enriched in amino acids carried by the most abundant major tRNA; this implies that the forces shaping codon usage can also influence protein sequences, i.e. the best codon for one amino acid may not be as good as that for another (8).

However, some discrepancies were noted regarding the pattern exhibited by Lys, which is enriched as CAI increases in spite of the low concentration of its major tRNA, and by Leu and Arg, whose major tRNAs were higher than expected (3); rather, it was suggested that the reduction in the diversity of amino acid choices encoded by highly expressed genes should be a strategy to increase translation efficiency (3).

Concomitant to the above explanation, our results would suggest that some force leads to a preferential presence of base G at first codon positions as the gene expression level rises. Such a force might be directed to optimize translational efficiency not necessarily mediated by tRNA abundances, but through a mechanism related to the three base (GNN) periodicity which arises from first codon position compositional bias.

It has been suggested that the repeating nature of RNY codons could simplify transcription and reduce frameshifting during translation, and that the preference for GNN codons might arise from translational advantages of such G-first codons (9–10). The above statement has been refined by proposing that the repeating GNN pattern found in the mRNA may be responsible for monitoring the correct reading frame during translation, based on the complementarity of G-periodical mRNA to the C periodical sites in the *E.coli* 16S rRNA sequence (11,12). Our results substantiate that the G-periodicity of mRNA sequences and thus their stickiness to the ribosome (11, see also 13 for review) might influence the rate of translation. In this framework, sequences with a stronger GNN periodical pattern should bind better to the framing sites, thus favouring processivity and avoidance of reading frame errors.

In connection with the above hypothesis, we note the variation in amino acid composition observed in proteins encoded in the bacteriophage lambda genome related to their expressivity (14). Although not statistically significant, the frequency of amino acids with G-starting codons is higher (36.8%) in the heavily expressed proteins encoded by the late operon (head and tail synthesis and assembly), than in the less expressed proteins corresponding either to the left operon (regulation and recombination) or to the right early operon (regulation and replication), where the frequencies of G-starting codons are 30.6 and 33.2%, respectively. On the other hand, the differences between the three major lambda operons are smaller when the C-starting codons are considered (late operon 22.1%, left operon 18.8% and right early operon 21.6%). This observation might be valuable to separate the variation in amino acid composition from tRNA availability, given the poor correlation between lambda codon usage and host tRNA abundance (1). Indeed, the codon usage of lambda head and tail genes resembles that of weakly expressed host genes (15) and has been considered an example of how high expression levels can be achieved for genes with a relatively high content of rare codons (13).

As a concluding remark, we would like to note that if the hypothesis is true, this would be an interesting example of how a selective force acting on a mechanical process (translational framing) might promote neutral amino acid substitutions with regard to protein function. Thus, amino acid replacements which are neutral for protein function might be advantageous with regard to protein synthesis mechanics.

ACKNOWLEDGEMENTS

Helpful comments by Josep Casadesús, Luis M. Corrochano, Francisco González and José L. Oliver are acknowledged.

REFERENCES

1 Ikemura (1981) *J. Mol. Biol.*, **146**, 1–21.
2 Gouy,M. and Gautier,C. (1982) *Nucleic Acids Res.*, **10**, 7055–7073.
3 Lobry,J.R. and Gautier,C. (1994) *Nucleic Acids Res.*, **22**, 3174–3180.
4 Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
5 Wahl,R., Rice,P., Rice,C.M. and Kröger,M. (1994) *Nucleic Acids. Res.*, **22**, 3450–3455.
6 Shepherd,J.C.W. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 1596–1600
7 Miyata,T. and Hayashida,H. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 5739–5743.
8 Sharp,P.M. and Matasi,G. (1994) *Curr. Opin. Genet. Dev.*, **4**, 851–860.
9 Eigen,M. and Schuster,P. (1979) *The Hypercycle.* Springer-Verlag, Berlin, Heidelberg, NY.
10 Wong,J.T. and Cedergren,R. (1986) *Eur. J. Biochem.*, **159**, 175–180.
11 Trifonov,E.N. (1987) *J. Mol. Biol.*, **194**, 643–652.
12 Lagúnez-Otero,J. and Trifonov,E. N. (1992) *J. Biomolec. Struct. Dynam.*, **10**, 455–464.
13 Andersson,S.G.E. and Kurland,C.G. (1990) *Microbiol. Rev.*, **54**, 198–210.
14 Daniels,D.L., Sanger,F. and Coulson,A.R. (1983) *Cold Spring Harbor Symp. Quant. Biol.*, **47**, 1009–1024.
15 Holm,L. (1986) *Nucleic Acids Res.*, **14**, 3075–3087.