Compositional variation in bacterial genes and proteins with potential expression level

Sabyasachi Das[a], Subhagata Ghosh[b], Archana Pan[a], Chitra Dutta[a,b,*]

[a] *Bioinformatics Center, Indian Institute of Chemical Biology, 4, Raja S.C. Mullick Road, Kolkata 700 032, India*
[b] *Human Genetics and Genomics Group, Indian Institute of Chemical Biology, 4, Raja S.C. Mullick Road, Kolkata 700 032, India*

Received 11 July 2005; accepted 22 August 2005

Available online 2 September 2005

Edited by Takashi Gojobori

Abstract **Usage of guanine and cytosine at three codon sites in eubacterial genes vary distinctly with potential expressivity, as predicted by Codon Adaptation Index (CAI). In bacteria with moderate/high GC-content, G_3 follows a biphasic relationship, while C_3 increases with CAI. In AT-rich bacteria, correlation of CAI is negative with G_3, but non-specific with C_3. Correlations of CAI with residues encoded by G-starting codons are positive, while with those by C-starting codons are usually negative/random. Average Size/Complexity Score and aromaticity of gene-products decrease with CAI, confirming general validity of cost-minimization principle in free-living eubacteria. Alcoholicity of bacterial gene-products usually decreases with expressivity.**

Keywords: Codon adaptation index; Synonymous codon usage; Amino acid usage; Average size/complexity score; Alcoholicity

1. Introduction

Gene expression plays an important role in codon adaptation in eubacteria. Highly expressed genes in eubacteria often exhibit a strong bias in synonymous codon usage – a phenomenon known as translational selection [1,2]. In a number of bacterial organisms [3–5], multivariate analysis of relative synonymous codon usage identified the gene expression level as one of the principal determinants of variation.

Gene expression may also influence the choice of non-synonymous codons. In *E. coli*, gene expression has been identified as one of the principle determinants of variation in amino acid usages [6]. Non-synonymous codon usage in *Buchnera* is strongly biased in putative high-expression genes, characterized by avoidance of aromatic amino acids, greater conservation and resistance to AT-enrichment [7], while in *Bertonella*, the strand-specific mutational pressure and gene expressivity both strongly influence the inter-strand variations in amino acid usage [8]. A trend in increase in frequency of the G-starting codons with increase in their potential level of expression was reported for certain bacterial organisms [9,10]. It is, therefore, intriguing to examine whether nucleotide/amino acid usage in bacterial genes/gene-products exhibit any definite correlation with their potential expression levels.

Expressivity is, however, not the sole determinant of gene/protein composition in eubacteria. There are various other factors that can influence the codon/amino acid usage [11–16], the most dominating one being the directional mutational bias [17,18]. Hence, in any investigation on codon/amino acid preferences, the nature of mutational pressure on the respective genomes should be taken into account. It is also necessary to check whether translational selection has any significance influence on their codon usage patterns, since there are certain bacterial genomes, especially those exhibiting extremely high [19,20] or strand-specific mutational bias [8,12], where no impact of translational selection on synonymous codon preferences could be observed. Influence of translation selection on codon bias can be assessed by several codon usage indices [1,21–23]. Among these, Codon Adaptation Index (CAI) is widely accepted as an effective measure of potential level of gene expression [1]. The present study demonstrates that in eubacterial genomes under appreciable translational selection pressure, variations in nucleotide and amino acid usage patterns exhibit distinct, significant (in some cases, unexpected) correlations with the potential gene expressivity (as measured by CAI), depending on the nature of mutational pressure.

2. Materials and methods

Annotated Open Reading Frames (ORFs) of fifty completely sequenced eubacterial genomes were retrieved from GenBank. For each genome, separate datasets for highly expressed genes (encoding ribosomal proteins and transcription/translation processing factors) [24–26] and lowly expressed genes (regulatory genes) were constructed and the correspondence analysis (COA) on Relative synonymous Codon Usage (RSCU) values were carried out using CodonW (written by John Peden and available at www.molbiol.ox.ac.uk/win95.codonW.zip).

For bacterial genomes experiencing significant impact of the translational selection on codon choice, (i) the average RSCU values of highly and lowly expressed genes should differ appreciably, (ii) CAI [1] as well as N_C [23] should vary over wide ranges, (iii) the first axis of COA on RSCU should have considerable amount of total variance (>7%) and the gene expressivity should be one of the major determinants of such variation. If all these criteria were fulfilled the respective genomes were considered to be under translational selection and were selected for further analysis. For each bacterial genome under study, CAI of individual genes were calculated taking a reference gene set comprising of genes encoding ribosomal proteins ($\geqslant$ 100 aa) and transcription-translation

[*] Corresponding author. Fax: +91 33 2473 0284/5197.
E-mail addresses: cdutta@iicb.res.in, chitradutta@hotmail.com (C. Dutta).

0014-5793/$30.00 © 2005 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2005.08.042

Table 1
Correlation coefficients: CAI vs. average size/complexity score, aromaticity, alcoholicity and GRAVY

		Organism	GC-content (%)	Average size/ complexity score	Aromaticity	Alcoholicity	GRAVY
LMB Genome	$45\% < \text{GC-content} < 55\%$	*E. coli*	50.7	-0.18^{***}	-0.17^{***}	-0.23^{***}	-0.09^{***}
		E. carotovora subsp.	51.0	-0.13^{***}	-0.09^{***}	-0.18^{***}	-0.10^{***}
		N. meningitidis	51.5	-0.28^{***}	-0.33^{***}	-0.12^{***}	0.03^{NS}
		S. typhimurium	52.2	-0.19^{***}	-0.19^{***}	-0.19^{***}	-0.11^{***}
		S. flexneri	50.8	-0.20^{***}	-0.23^{***}	-0.23^{***}	-0.03^{NS}
		Synechocystis sp.	47.6	-0.06^{**}	-0.16^{***}	-0.22^{***}	0.02^{NS}
		T. maritima	46.2	-0.11^{***}	-0.09^{***}	-0.22^{***}	-0.18^{***}
		V. cholerae (Ch1)	47.7	-0.07^{***}	-0.04^{*}	-0.16^{***}	-0.12^{***}
		W. succinogenes	48.5	-0.13^{***}	-0.09^{***}	-0.03^{NS}	-0.05^{*}
		Y. pestis	47.6	-0.13^{***}	-0.15^{***}	-0.17^{***}	-0.12^{***}
HMB Genome	$\text{GC-content} > 55\%$	*A. tumefaciens* (Ch1)	59.3	-0.20^{***}	-0.31^{***}	-0.07^{***}	-0.02^{NS}
		B. longum	60.0	-0.23^{***}	-0.29^{***}	-0.29^{***}	0.09^{***}
		B. japonicum	64.0	-0.17^{***}	0.10^{***}	-0.12^{***}	0.13^{***}
		B. melitensis (Ch1)	57.0	-0.15^{***}	-0.27^{***}	-0.20^{***}	-0.00^{NS}
		D. radiodurans (Ch1)	66.9	-0.09^{***}	-0.21^{***}	-0.04^{*}	-0.04^{*}
		G. oxydans	61.0	-0.18^{***}	-0.05^{*}	-0.08^{***}	0.00^{NS}
		M. leprae	57.8	-0.06^{*}	-0.16^{***}	-0.24^{***}	-0.07^{**}
		P. acnes	60.0	-0.08^{***}	0.08^{***}	-0.22^{***}	-0.15^{***}
		P. aeruginosa	66.4	-0.02^{NS}	-0.15^{***}	-0.10^{***}	0.16^{***}
		R. solanacearum	67.0	-0.18^{***}	-0.26^{***}	-0.18^{***}	0.20^{***}
		S. pomeroyi	64.1	-0.03^{NS}	0.05^{**}	-0.14^{***}	-0.07^{***}
		X. campestris	65.0	-0.14^{***}	-0.15^{***}	-0.12^{***}	0.07^{***}
	$\text{GC-content} < 45\%$	*B. subtilis*	43.5	-0.06^{***}	0.00^{NS}	-0.07^{***}	-0.23^{***}
		C. pneumoniae	40.5	0.02^{NS}	0.12^{***}	-0.11^{***}	-0.16^{***}
		C. acetobutylicum	30.9	-0.01^{NS}	-0.13^{***}	-0.16^{***}	-0.19^{***}
		H. influenzae	38.1	-0.10^{***}	-0.05^{*}	-0.16^{***}	-0.12^{***}
		L. johnsonii	34.5	0.03^{NS}	-0.14^{***}	0.06^{**}	-0.24^{***}
		L. lactis	35.2	-0.16^{***}	-0.15^{***}	-0.15^{***}	-0.05^{*}
		M. succiniciproducens	42.5	-0.16^{***}	-0.16^{***}	-0.04^{NS}	-0.09^{***}
		Nostoc sp.	41.3	0.06^{***}	-0.03^{NS}	-0.09^{***}	-0.09^{***}
		P. multocida	40.3	-0.18^{***}	-0.02^{NS}	-0.09^{***}	-0.19^{***}
		S. aureus	32.8	0.08^{***}	-0.03^{NS}	-0.05^{*}	-0.36^{***}
		S. pneumoniae	39.6	-0.25^{***}	-0.06^{**}	-0.13^{***}	-0.18^{***}
		V. fischeri	38.4	-0.11^{***}	-0.19^{***}	-0.27^{***}	-0.13^{***}

*Correlation coefficients significant at $p < 0.05$, **at $p < 0.01$and ***at $p < 0.001$; NS, non-significant.

processing factors (EF-Tu, EF-G, RpoB, RpoC and RpoA), which are known to be expressed at high levels in most bacterial organisms [24–26].

To check for the strand-specific compositional asymmetry, the probable oriC and termination regions were predicted using GC-skew analysis [27] and codon/amino acid usage of the leading and lagging strands were compared for each genome under study. Finally, thirty-four genomes were selected (Table 1, Supplementary Table 1) and sixteen genomes, which are characterized by either inefficient translational selection or strong compositional asymmetries in two strands, were excluded from further analysis (Supplementary Table 2).

The selected organisms were categorized into three major groups on the basis of overall GC-content – low mutational bias (LMB) group, GC-rich high mutational bias (HMB) group and AT-rich HMB group. To avoid statistical errors, the ORFs having less than 100 codons were not considered. The annotated pseudogenes, transposon genes and the ORFs reported to have authentic frameshift(s) were also excluded. For each individual annotated ORF in the datasets, the following parameters were computed: frequencies of the individual bases and GC-content at three codon sites and $(GC)_3$-skewness $[(G_3 - C_3)/(G_3 + C_3)]$. Similarly for each encoded gene-products, the frequencies of individual amino acid residues, average size/complexity score [28], alcoholicity (relative frequency of alcoholic residues, serine and threonine), aromaticity (relative frequency of aromatic amino acid residues in the protein) [6] and GRAVY score (mean hydropathy index of the encoded amino acid residues and hence $\text{GRAVY} = \sum_j \alpha_j f_j$, where α_j is the hydropathy index of jth type residue) [29] were computed.

Scatter diagrams were plotted to examine the correlation of each of these parameters with the CAI of corresponding genes. In each case, the correlation coefficients and slope were determined using STATISTICA (Version 6.0) to assess the statistical significance of the correlation, if any.

3. Results and discussion

3.1. Distinct trends in variations in G_3 and C_3 with CAI

In most of the LMB and GC-rich genomes under study, excepting *T. maritima* and *M. Leprae*, G_3 exhibits a striking biphasic relationship with CAI, where G_3 first increases and then decreases with increase in CAI (Fig. 1Ai, 1Bi). The CAI values at which the transition in correlation patterns of G_3 occurs are different for different organisms (Supplementary Table 3). C_3, on contrary, is positively correlated to CAI (Fig. 1Aii, 1Bii), the correlation coefficients and slopes differing from species to species (Supplementary Table 3). Moreover, GC_3 increases at first with CAI and then gradually reaches a plateau in most of the LMB and GC-rich bacteria (Fig. 1Aiii, 1Biii). *T. maritima* and *M. leprae* are two exceptions, which exhibit positive correlations of CAI with G_3, C_3 and GC_3 for entire range of CAI values (Fig. 1C).

In the AT-rich organisms, both G_3 and GC_3 decrease with CAI (Fig. 1Di, 1Diii), while the correlation between C_3 and CAI can be random (Fig. 1Dii), positive or negative (Supplementary Table 3). In LMB and GC-rich bacterial genomes, correlation of CAI is negative with T_3 and biphasic or random with A_3, but in AT-rich organisms, T_3, A_3 and AT_3 -all exhibit positive correlations with CAI (data not shown).

The scatter plots of G_3 and C_3 against N_c (not shown) are compatible with the observations made with CAI. For