Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Unusual codon usage bias in low expression genes of *Vibrio cholerae*

Surajit Basak[1,] *, Indranuj Mukherjee[1], Mayukh Choudhury[2] and Santasabuj Das[1]

[1]Biomedical Informatics Center, National Institute of Cholera and Enteric Diseases, P-33, C.I.T Road, Scheme-XM, Beliaghata, Kolkata 700010, India; [2]School of Biotechnology, Chemical and Biomedical Engineering, VIT University, Vellore 632014, India;
Surajit Basak* - Email: basaksurajit@gmail.com; Fax: 91 33 23705066; * Corresponding author

received December 01, 2008; accepted December 02, 2008; published December 31, 2008

Abstract:
Positive correlation between gene expression and synonymous codon usage bias is well documented in the literature. However, in the present study of *Vibrio cholerae* genome, we have identified a group of genes having unusually high codon usage bias despite being low potential expressivity. Our results suggest that codon usage in lowly expressed genes might also be selected on to preferably use non-optimal codons to maintain a low cellular concentration of the proteins that they encode. This would predict that lowly expressed genes are also biased in codon usage, but in a way that is opposite to the bias of highly expressed genes.

Keywords: CAI; RSCU; effective number of codons; tRNA copy number; correspondence analysis

Background:
In most species synonymous codons are not used with equal frequencies; the phenomenon known as codon usage bias. Codon bias is generally governed by a balance between mutation, genetic drift and natural selection **[1-5**]. In various organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, synonymous codon usage bias has been shown to be correlated with the abundance of isoaccepting tRNA [6]. An optimal codon is thought to increase translation rate **[7-9]**. On the other hand, the presence of non-optimal codons has been postulated to reduce translation rate **[10**], probably due to a relative scarcity of cognate tRNA species. Non-optimal codons have selective advantage to maintain a low cellular concentration of the proteins that they encode **[11]**. It was reported previously that non-optimal codons occur at high frequency in the signal sequence of secretory genes in *Escherichia coli* **[12]**. The high occurrence of non-optimal codons in the signal sequence of secretory proteins has also been observed in the gram-positive bacterium *Streptomyces coelicolor* **[13]**.

Apart from gene expression level, gene length also has important role in affecting synonymous codon usage bias. Several earlier studies have also documented strong effects of gene length on codon bias in a variety of organisms. The level of synonymous codon usage bias has been shown to be positively correlated to gene length in *Escherichia coli* **[14]**. In *Drosophila* genome, longer genes had lower codon usage bias **[15]**. However, Hou and Yang **[16]** reported that in *S. penumoniaes* genome, the longer genes had higher expression level and higher codon usage bias.

Cholera remains a heavy burden to human health in some developing countries including India where sanitation is poor and health care is limited **[17-20]**. After the publication of the complete genome sequence of *Vibrio cholerae* **[21]**, the etiological agent of cholera, extensive possibilities, earlier unavailable, have opened up to understand the genetic organization of *Vibrio cholerae*. The present study demonstrates an unusual trend in synonymous codon usage pattern of lowly expressed genes of *Vibrio cholerae* genome. Contrary to the usual expectation, we have identified 138 genes that are highly biased yet lowly expressed. Moreover, the usage pattern of non-optimal codons in lowly expressed genes depends on the gene length. Our results clearly suggest that translational selection has significant influence on the codon usage pattern of lowly expressed genes depending on gene length.

Methodology:
The complete genome sequence of *Vibrio cholerae* has been downloaded from ftp://ftp.ncbi.nih.gov/genbank/genomes and the coding sequences were extracted. To minimize the sampling errors **[22],** only those coding sequences that are more than or equal to 30 amino acids has been retained for our analysis. Correspondence analysis **[23]** available in CodonW 1.4.2 (J. Peden, 2000; http://www.molbiol.ox.ac.uk/cu/) was used to investigate the major trend in relative synonymous codon usage variation among the genes. We have also used CodonW to calculate Relative Synonymous Codon Usage (RSCU) values and gene length. Synonymous codon usage bias was measured by calculating the 'effective number of codons used in a gene' (Nc) **[22, 24]**. The values of Nc range from 20 (when one codon is used per amino acid) to 61 (when all the codons are used with equal probability). In the present study, a gene is designated as highly biased if Nc< 36, and lowly biased if Nc> 44. We have used Codon Adaptation Index (CAI) to calculate gene expressivities. CAI is widely accepted as an effective measure of potential level of gene expression **[25]**. CAI of individual genes were calculated taking a reference gene set of all the ribosomal proteins, which are known to be

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 213-217 (2008)

highly expressed in most bacterial genomes **[26-28]**. We have sorted our dataset according to the CAI values. We have taken out genes from extreme 20% of population from both ends of the sorted dataset. Using the above criteria, a gene is considered as lowly expressed if its CAI < 0.318 and highly expressed if it's CAI > 0.502. The transfer RNA gene copy numbers were taken from the tRNA scan database (http://lowelab.cse.ucsc.edu/GtRNAdb/Vibr_chol/). The Student's t-test was used to evaluate the significance of all the pairwise differences. The correlation coefficients were determined using SPSS (13.0) to assess the statistical significance of the correlation, if any.

Results and Discussion:

Codon usage bias and gene expression level

Many studies have demonstrated a positive correlation between degree of codon bias and level of gene expression **[25, 29, 30]**. As a result, it is generally expected that lowly expressed genes should have lower codon bias and highly expressed genes should have higher codon bias. When analysis was performed taking all the genes in *V. cholerae* genome, we also observed a negative correlation (r= -0.2994, P< 0.01) between Nc and CAI which indicates that degree of codon bias increases with the increase in gene expression level. However, careful inspection of the plot between Nc and CAI reveals (Figure 1) that though lowly biased genes are lowly expressed but not all the highly biased genes are highly expressed. Finally, we have identified 138 genes which show unusual pattern of codon usage (i.e., high codon usage bias yet lowly expressed).

Correspondence analysis on RSCU: Identification of translationally non-optimal codons

We have performed correspondence analysis (CoA) on the set of highly and lowly expressed genes. Since codon usage by its very nature is multivariate, one of the most popular multivariate methods for studying codon usage variation is correspondence analysis **[23]**. Correspondence analysis identifies the major trends in the variation of the synonymous codon usage data and distributes genes along continuous axes in accordance with these trends. Correspondence analysis on relative synonymous codon usage (RSCU) detected one major trends of codon usage variation on first axis of inertia. The first axis accounted for 17.48% of the total variation and no other axis accounted for more than 6.97% of the total variation. As expected, the position of the genes along the first major axis is significantly correlated with the corresponding CAI values (r= -0.9648, P< 0.01). By looking at the distribution of codons along the first two major axes (Figure 2), we have identified five most preferred codons in lowly expressed genes situated at the most extreme positions in the positive side of the Axis 1. The names of these codons are: AGG, CGA, CGG, AUA, and AGA. Non-optimal codons are defined by their low usage in the genome and the low abundance of their corresponding tRNA **[8, 31]**. If we compare the RSCU values of the above five codons among other synonymous alternatives (Table 1 in supplementary material), we can see that they are used less frequently among other synonyms. Moreover, we have used tRNA gene copy number data (Table 1 in supplementary material) to assess the abundance of their corresponding tRNA. Table 1 also suggests that the five most preferred codons in lowly expressed genes has either nil or lowest abundance of the corresponding tRNA gene copy number. Thus, these five codons can also be considered as the non-optimal codons.

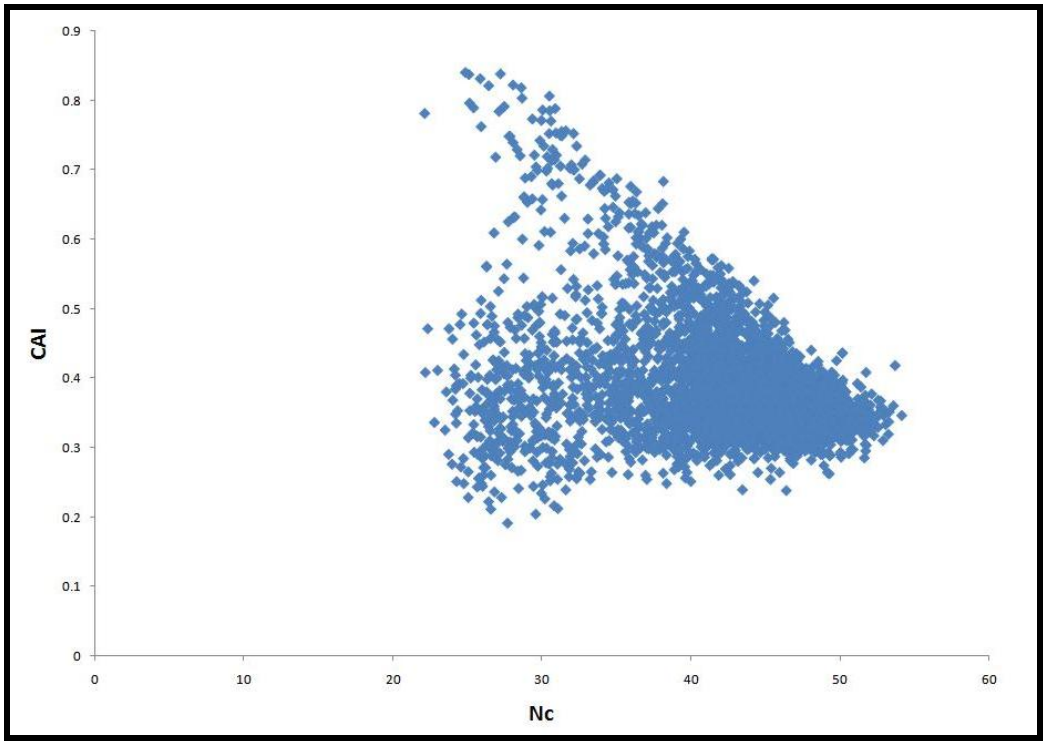

Figure 1: Variation of Effective Number of Codons (Nc) against Codon Adaptation Index (CAI).

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 213-217 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

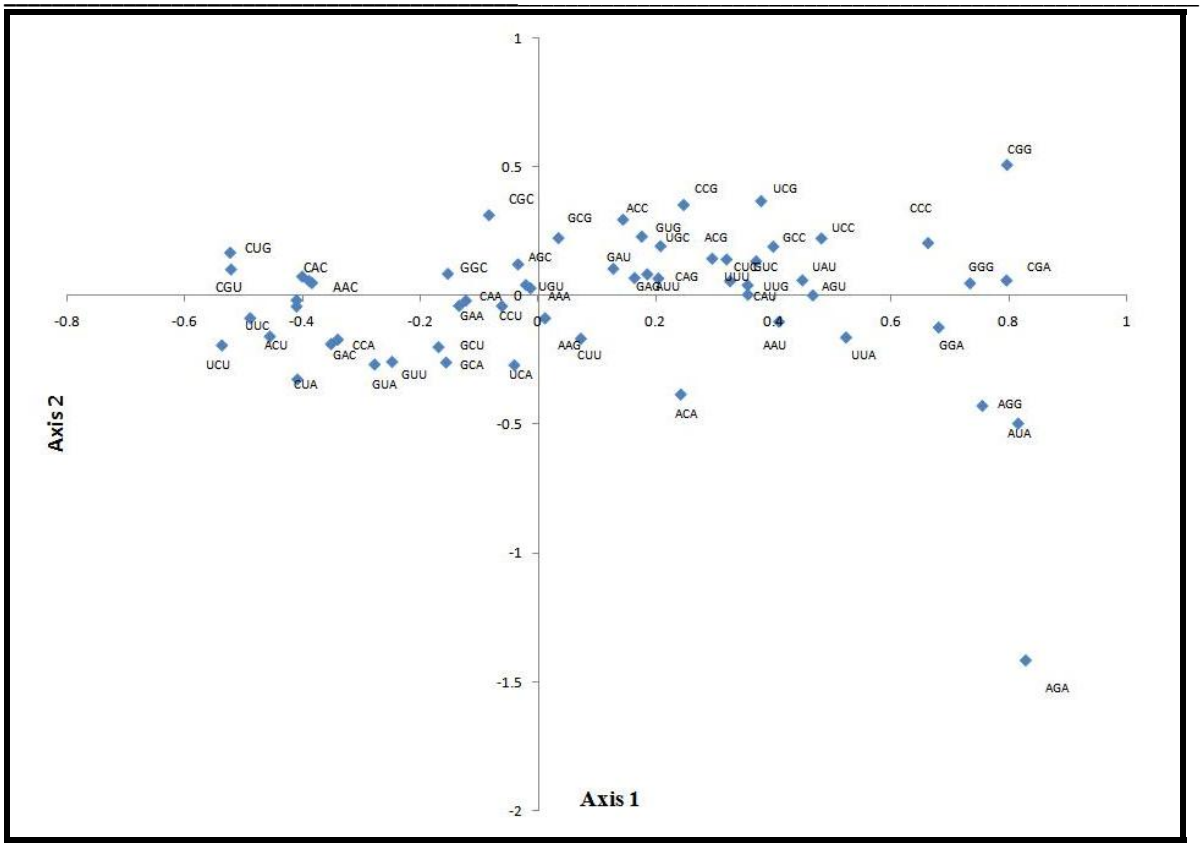

Figure 2: The distribution of codons of all *Vibrio cholerae* genes along the first and second axes of the correspondence analysis.

Correspondence analysis on RSCU of low expression genes

We have identified a set of 138 genes showing unusual pattern of synonymous codon usage, i.e., they are highly biased (low Nc) but lowly expressed (low CAI). Therefore, correspondence analysis was performed on RSCU of lowly expressed genes to analyze the differential nature of selective constraints acting on synonymous codon usage pattern of lowly expressed genes. CoA detected one single explanatory axis of major synonymous codon usage variation. The first major axis accounted for 9.58% of the total variation in codon usage and no other axis accounted for more than 5.34% of the total variation. More importantly, we have noted that position of the genes along the first major axis is significantly correlated with Nc (r= -0.3395, P< 0.01) and gene length (r= -0.2818, P<0.01). Thus genes placed at positive side of Axis 1 are highly biased and of smaller length. We have also compared the average length of lowly biased and highly biased groups of lowly expressed genes. The average length of highly biased genes is 50.15 and the average length of lowly biased genes is 372.08. This difference in gene length between lowly biased and highly biased groups of lowly expressed genes is statistically significant at P<0.001.

We have also analyzed the codon distribution along first major axis generated from CoA on RSCU of lowly expressed genes (data not shown). The name of the most preferred codons at the extreme of the positive side of Axis 1 is: AGG, AUA, UCA, ACA, AGA and those at the extreme of the negative side of Axis 1 are: CGC, CGG, CCG, CGU, CUG. One interesting observation is that among the five most preferred codons at the positive side of Axis 1, three are non-optimal codons (please see section *Correspondence analysis on RSCU: Identification of translationally non-optimal codons*). On the other hand, there is no non-optimal codon present among the most preferred codons at the negative side of Axis 1.

Non-optimal codon usage: Shorter gene length and greater codon biasness of lowly expressed genes

From the above results it is clear that the frequencies of non-optimal codons are greater in the highly biased group of lowly expressed genes and the average length of these group of genes are significantly smaller than the average length of lowly biased lowly expressed genes. The presence of non-optimal codons has been postulated to reduce translation rate **[10]**, probably due to a relative scarcity of cognate tRNA species. Considering these facts it is reasonable to argue that selective constraints on the usage of non-optimal codons are greater in the highly biased groups of lowly expressed genes than lowly biased lowly expressed genes.

If synonymous codon usage pattern among the lowly expressed genes is explained by selection to reduce translational rate, is this consistent with the length effect? Several earlier studies have also documented the influence of gene length on codon bias in a variety of organisms **[32. 33]**. Powell and Moriyama **[34]** hypothesized that length

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 213-217 (2008)

effect could be explained by selection for translation rate, e.g., in a short gene with 100 codons, such mutation would increase translation time by 1%, whereas the same mutation in a gene with 1000 codons would increase translation time by only 0.1%. In the present study, among the lowly expressed genes, mutations of non-optimal codon will have greater relative effect in smaller genes compared with larger genes. Thus such mutations are likely to be counter selected in short genes than in long genes.

Conclusion:

In summary, the present study attempts to focus on the unusual trends in synonymous codon usage pattern in lowly expressed genes of *V. cholerae* genome. Selection forces governing the synonymous codon usage in bacterial genes usually vary across or within the genomes. One might, therefore, expect to observe species- specific and/or gene- specific trends in synonymous codon usage pattern. This study finds that selective preference of the non-optimal codons in shorter lowly expressed genes has made them highly bias and might have a greater role in translational pausing to allow correct folding of proteins. The unusual pattern of synonymous codon usage observed in a subset of lowly expressed genes of *Vibrio cholerae* genome may provide a new starting point for the study of the organism's environmental and pathobiological characteristics. It will be interesting to see if the synonymous codon usage pattern could influence to determine the gene expressions that are unique to its survival and replication during human infection **[35]** as well as in the environment **[17, 36]**.

Acknowledgment:

This work is supported by an extramural project (IRIS ID: 2001-0496D) funded by ICMR, Government of India.

References:

[01] M. Bulmer, *Genetics*, 129: 897 (1991) [PMID: 1752426]

[02] P. M. Sharp *et al.*, *Biochem Soc Trans*., 21: 835 (1993) [PMID: 8132077]

[03] B. Kahalii *et al.*, *Biochem Biophys.Res Commun.*, 354: 693 (2007) [PMID: 17258174]

[04] S. Basak and T. C. Ghosh , *Biochem Biophys Res Commun*., 330: 629 (2005) [PMID: 15809043]

[05] S. Basak *et al.*, *FEBS Letters*, 581: 5825 (2007) [PMID: 18047839]

[06] C. G. Kurland, *FEBS Letters*, 285: 165 (1991) [PMID: 1855585]

[07] T. Ikemura, *J Mol Biol*., 146: 1 [PMID: 6167728]

[08] T. Ikemura., *J Mol Biol*., 151: 389 (1981) [PMID: 6175758]

[09] P. M. Sharp *et al.*, *Nucleic Acids Res*., 14: 5125 (1986) [PMID: 3526280]

[10] J. H. Kinnaird *et al.*, *J Mol Biol.*, 221: 733 (1991) [PMID: 1834852]

[11] M. Kimura., *The neutral theory of molecular evolution, Cambridge University Press, Cambridge*, U.K (1983)

[12] P. M. Power *et al.*, *Biochem. Biophys Res Commun*., 322: 1038 (2004) [PMID: 15336569]

[13] Y. D. Li *et al.*, *Biosystems*, 85: 225 (2006) [PMID: 16644095]

[14] A. Eyre-Walker, *Mol Biol Evol.*, 13: 864 (1996) [PMID: 8754221]

[15] J. M. Comeron *et al.*, *Genetics*, 151: 239 (1999) [PMID: 9872963]

[16] Z. C. Hou and N. Yang., *Acta Genet Sin*., 29: 747 (2000) [PMID: 12200868]

[17] R. R. Colwell, *Science*, 274: 2025 (1996) [PMID: 8953025]

[18] M. K. Waldor and D. R. Chaudhuri, *Nature*, 406: 469 (2000) [PMID: 10952295]

[19] S. Shinoda., *Biocontrol Sci*., 13: 1 (2008) [PMID: 18432110]

[20] G. B. Nair, *Indian J Med Res*., 125: 612 (2007) [PMID: 17642495]

[21] J. F. Heidelberg *et al*, *Nature*: 406: 477 (2000) [PMID: 10952301]

[22] F. Wright, *Gene*, 87: 23 (1990) [PMID: 2110097]

[23] M. J. Greenacre, *Theory and Application of Correspondence Analysis*, *Academic Press*, *London* (1984)

[24] T. Banerjee *et al.*, *Biochem Biophys Res Commun*., 330: 1015 (2005) [PMID: 15823544]

[25] P. M. Sharp and W. H. Li, *Nucleic Acids Res*., 15: 1281 (1987) [PMID: 3547335]

[26] S. Karlin and J. Mrázek., *J. Bacteriol*., 182: 5238 (2000) [PMID: 10960111]

[27] S. Karlin and J. Mrázek, *Proc Natl Acad Sci U S A*, 98: 5240 (2001) [PMID: 11296249]

[28] S. Karlin *et al.*, *J Bacteriol.*, 183: 5025 (2001) [PMID: 11489855]

[29] M. Gouy and C. Gautier, *Nucleic Acids Res*., 10: 7055 (1982) [PMID: 6760125]

[30] J. L. Bennetzen and B. D. Hall, *J Biol Chem.*, 257: 3026 (1982) [PMID: 7037777]

[31] Y. M. Zalucki *et al.*, *Biochem Biophys Res Commun.*, 366: 135 (2008). [PMID: 18053805]

[32] L. Duret and D. Mouchiroud, *Proc Natl Acad Sci U S A*, 96: 4482 (1999) [PMID:18053805]

[33] B. Lemos *et al.*, *Mol Biol Evol*, 22: 1345 (2005) [PMID: 15746013]

[34] J. R. Powell and E. N. Moriyama, *Proc Natl Acad Sci U S A*, 94: 7784 (1997) [PMID: 9223264]

[35] S. H. Lee *et al.*, *Cell*, 99: 625 (1999) [PMID: 10612398]

[36] R. R. Colwell and A. Huq, *Ann. N.Y Acad Sci*, 740: 44 (1994) [PMID: 7840478]

Edited by P. Kangueane

Citation: Basak *et al.*, Bioinformation 3(5): 213-217 (2008)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 213-217 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Amino acid	Codon	RSCU values	tRNA copy number	Amino acid	Codon	RSCU values	tRNA copy number
PHE	UUU	1.32		TYR	UAU	1.07	
	UUC	0.68	3		UAC	0.93	5
LEU	UUA	1.10	2	HIS	CAU	1.10	
	UUG	1.30	1		CAC	0.90	2
	CUU	0.71		GLN	CAA	1.30	5
	CUC	0.82	1		CAG	0.70	
	CUA	0.49	5	ASN	AAU	1.01	
	CUG	1.59	3		AAC	0.99	4
ILE	AUU	1.55		LYS	AAA	1.46	4
	AUC	1.24	3		AAG	0.54	
	AUA	0.21		ASP	GAU	1.46	
MET	AUG	1.00	9		GAC	0.54	5
VAL	GUU	0.91		GLU	GAA	1.23	4
	GUC	0.82	2		GAG	0.77	
	GUA	0.63	2	CYS	UGU	1.14	
	GUG	1.65			UGC	0.86	3
SER	UCU	1.07		TRP	UGG	1.00	1
	UCC	0.57	1	ARG	CGU	2.41	6
	UCA	1.01	2		CGC	2.14	
	UCG	0.89			CGA	0.63	
	AGU	1.13			CGG	0.34	1
	AGC	1.34	2		AGA	0.35	1
PRO	CCU	1.11			AGG	0.14	
	CCC	0.60	1	GLY	GGU	1.55	
	CCA	1.22	3		GGC	1.45	6
	CCG	1.07			GGA	0.47	2
THR	ACU	1.00			GGG	0.53	
	ACC	1.55	2				
	ACA	0.60	4				
	ACG	0.85					
ALA	GCU	0.88					
	GCC	0.95	1				
	GCA	0.83	4				
	GCG	1.34					

Table 1: RSCU values and corresponding tRNA copy number of *Vibrio cholerae* genome.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 213-217 (2008)