The Journal of Microbiology, February 2007, p. 58-63
Copyright © 2007, The Microbiological Society of Korea

Vol. 45, No. 1

NOTE

Studies on Synonymous Codon and Amino Acid Usage Biases in the Broad-Host Range Bacteriophage KVP40

Keya Sau[3], Sanjib Kumar Gupta[1], Subrata Sau[2], Subhas Chandra Mandal[3,], and Tapash Chandra Ghosh[1,*]**

[1]Bioinformatics Centre, [2]Department of Biochemistry, Bose Institute, P1/12-CIT Scheme VII M, Calcutta 700 054, India
[3]Department of Mathematics, Jadavpur University Calcutta-700 032, India

(Received April 28, 2005 / Accepted August 24, 2006)

In this study, the relative synonymous codon and amino acid usage biases of the broad-host range phage, KVP40, were investigated in an attempt to understand the structure and function of its proteins/protein-coding genes, as well as the role of its tRNAs. Synonymous codons in KVP40 were determined to be AT-rich at the third codon positions, and their variations are dictated principally by both mutational bias and translational selection. Further analysis revealed that the RSCU of KVP40 is distinct from that of its Vibrio hosts, *V. cholerae* and *V. parahaemolyticus*. Interestingly, the expression of the putative highly expressed genes of KVP40 appear to be preferentially influenced by the abundant host tRNA species, whereas the tRNAs expressed by KVP40 may be required for the efficient synthesis of all its proteins in a diverse array of hosts. The data generated in this study also revealed that KVP40 proteins are rich in low molecular weight amino acid residues, and that these variations are influenced primarily by hydropathy, mean molecular weight, aromaticity, and cysteine content.

Keywords: relative synonymous codon usage (RSCU), amino acid usage, correspondence analysis, phage KVP40, GRAVY, mean molecular weight

Synonymous codon usage, studied in a number of living organisms, has basically proven to be non-random and species-specific. Several factors, including directional mutational bias (Levine and Whitemore, 2000; Jenkins and Holmes, 2003), translational selection (Ikemura, 1985; Sharp and Cowe 1991; Gupta and Ghosh, 2001), secondary protein structure (Oresisc and Shalloway, 1998; Gupta *et al.*, 2002; D'Onofrio *et al.*, 2002), replicational and transcriptional selection (McInerney, 1998; Romero *et al.*, 2000), and environmental factors (Lynn *et al.*, 2002; Basak *et al.*, 2004), have been shown to influence codon usage in a variety of organisms. By way of contrast, amino acid usage has been demonstrated to be influenced by hydropathy, aromaticity, cysteine content, mean molecular weight, growth temperature, etc. (Lobry and Gautier, 1994; Garat and Musto, 2000; Zavala *et al.*, 2002; Banerjee *et al.*, 2004; Basak *et al.*, 2004; Naya *et al.*, 2004).

Bacteriophage KVP40 was isolated from marine water, and has been reported to infect eight *Vibrio* and one *Photobacterium* species (Matsuzaki *et al.*, 1992). It is a double-stranded DNA phage with an overall G+C content of 42.6%. The genome size of this T4-like virus, a member of the *Myoviridae* family, is 244834 bp (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES). It harbors 386 protein-coding genes, 30 tRNA-encoding genes, and several promoters and transcription terminators (Miller *et al.*, 2003). Nearly 65% of its protein coding genes are unique, and have no known functions. Little, at present, is known regarding the structure and function of its proteins/protein-coding genes, as well as the function of its putative tRNAs. In this communication, both the synonymous codon and amino acid usage biases in KVP40 have been studied, in order to determine how the proteins or protein-coding genes in this broad-host range phage have been shaped, as well as its tRNAs contribute to the expression of its proteins.

Relative synonymous codon usage (RSCU) was assessed in 376 protein coding sequences (all encode for ≥50 amino acid residues) of phage KVP40, in accordance with the standard protocols (Sharp and Li, 1987). The RSCU values of the protein coding genes (each encoding for at least 100 amino acid residues) of *V. cholerae* and *V. parahaemolyticus* were also determined, and were compared with those of KVP40. All sequences were also downloaded from the GenBank database (NCBI, USA). A_3, T_3, G_3, and C_3 are the distributions of A, T, G, and C at the third codon position. GC_3 is the frequency of (G+C) at the third codon position. N_c designates the effective number of codons used by a gene, and is generally utilized to determine the bias of synonymous codons, independent of amino acid composition and codon number (Wright, 1990). The N_c values were calculated in accordance with the method of Banerjee *et al.* (Banerjee *et al.,* 2005). The putatively highly- and lowly-expressed genes were considered, respectively, on the basis of the lowest 10% and highest 10% of the genes, as determined by their

* To whom correspondence should be addressed.
(Tel) 91-33-2355-6626; (Fax) 91-33-2355-3886
(E-mail) tapash@bic.boseinst.ernet.in

Table 1. Relative synonymous codon usage analysis in KVP40

		RSCU			tRNA copy number	
		KVP40				
AA	Codon	Overall	HEG[1]	LEG[2]	VP*	KVP40
Phe	UUU	0.84	0.96	0.88		1
	UUC	1.16	1.04	1.12	4	1
Leu	UUA	0.86	0.59	1.13	3	1
	UUG	1.27	0.75	0.97	1	1
	CUU	1.49	1.82	1.25		
	CUC	0.36	0	0.72	2	
	CUA	1.25	2.04	1.13	9	1
	CUG	0.78	0.80	0.81		
Ile	AUU	1.48	0.99	1.28		
	AUC	1.28	2.01	1.17	2	2
	AUA	0.23	0	0.55		
Met	AUG	1	1	1	11	2
Val	GUU	1.67	1.48	1.37		
	GUC	0.57	0.29	0.48	2	
	GUA	1.01	1.36	1.41	4	1
	GUG	0.75	0.87	0.73		
Ser	UCU	1.26	1.78	1.67		
	UCC	0.10	0	0.15	1	
	UCA	1.75	2.51	1.06	4	
	UCG	0.87	0.49	0.76		
	AGU	1.03	0	0.99		
	AGC	0.99	1.22	1.37	1	1
Pro	CCU	1.00	1.33	1.05		
	CCC	0.26	0	0.57		
	CCA	1.51	1.92	1.24	3	2
	CCG	1.23	0.75	1.14		
Thr	ACU	1.58	3.15	1.14		
	ACC	0.26	0.09	0.43	2	
	ACA	1.29	0.52	1.43	5	1
	ACG	0.87	0.24	1		
Ala	GCU	1.30	1.85	1.08		
	GCC	0.23	0.02	0.47	1	
	GCA	1.64	1.59	1.42	4	
	GCG	0.83	0.54	1.03		
Tyr	UAU	0.87	0.26	0.87		
	UAC	1.13	1.74	1.13	7	
His	CAU	0.96	0.63	1.05		
	CAC	1.04	1.37	0.95	2	1
Gln	CAA	1.36	1.59	1.31	6	1
	CAG	0.64	0.41	0.69		
Asn	AAU	0.88	0.61	0.97		
	AAC	1.12	1.39	1.03	5	2
Lys	AAA	1.28	1.43	1.10	4	2
	AAG	0.72	0.57	0.90		
Asp	GAU	1.01	0.60	1.18		
	GAC	0.99	1.40	0.82	6	1
Glu	GAA	1.42	1.66	1.35	5	
	GAG	0.58	0.34	0.65		
Cys	UGU	1.37	1.62	1.29		
	UGC	0.63	0.38	0.71	4	1
Trp	UGG	1	1	1	2	1
Arg	CGU	2.57	3.46	1.54	8	1
	CGC	1.54	2.03	1.70		
	CGA	1.03	0.08	1.30		
	CGG	0.09	0	0.32	1	
	AGA	0.7	0.34	0.81	1	1
	AGG	0.07	0.08	0.32	1	
Gly	GGU	2.25	2.97	1.71		
	GGC	1	0.88	0.76	11	
	GGA	0.35	0	0.59	2	1
	GGG	0.40	0.15	0.94		

1 and 2; denote KP40-specific highly- expressed (having Nc<30) and lowly- expressed (having Nc>46) genes, respectively.

*; indicates *V. parahaemolyticus*. Codons of *V. parahaemolyticus* which are recognized by 2 or more number of its tRNAs were considered optimal here.

Table 2. Relative synonymous codon usage (RSCU) values for each codon for the two groups of genes. The asterisk denotes the codons whose occurrences are significantly (p<0.01) higher on the extreme positive side of axis 2 than in the genes present on the extreme negative sides of the second major axis. The circles denote the codons whose occurrences are significantly (p<0.01) higher on the extreme negative side of axis 2 than the genes present on the extreme positive sides of the second major axis. Superscript "a" denotes genes at the extreme positive side of axis 2, and "b" for genes at the extreme negative side of axis 2. Each group contains 10% of sequences at either extreme of the major axis generated via correspondence analysis. N is the number of codons, AA represents amino acid. See text for details

AA	Codon	N^a	$RSCU^a$	N^b	$RSCU^b$
Phe	TTT°	64	0.56	105	0.95
	TTC*	166	1.44	116	1.05
Tyr	TAT°	57	0.68	106	1.01
	TAC*	110	1.32	103	0.99
His	CAT°	33	0.62	75	1.14
	CAC*	73	1.38	57	0.86
Asn	AAT°	69	0.57	127	1.05
	AAC*	175	1.43	114	0.95
Asp	GAT°	137	0.78	164	1.09
	GAC*	213	1.22	136	0.91
Cys	TGT	35	1.19	47	1.25
	TGC	24	0.81	28	0.75
Gln	CAA*	149	1.54	91	1.31
	CAG°	45	0.46	48	0.69
Lys	AAA*	315	1.45	174	1.21
	AAG°	120	0.55	113	0.79
Glu	GAA	364	1.49	243	1.38
	GAG	125	0.51	110	0.62
Val	GTT	169	1.81	121	1.64
	GTC	41	0.44	48	0.65
	GTA	104	1.12	73	0.99
	GTG	59	0.63	53	0.72
Pro	CCT*	46	1.34	21	0.66
	CCC°	1	0.03	28	0.88
	CCA*	73	2.13	42	1.31
	CCG°	17	0.5	37	1.16
Thr	ACT*	210	2.59	86	1.37
	ACC	14	0.17	18	0.29
	ACA°	76	0.94	85	1.35
	ACG°	24	0.3	62	0.99
Ala	GCT*	212	1.71	67	1.02
	GCC°	13	0.1	23	0.35
	GCA	197	1.59	118	1.79
	GCG°	75	0.6	56	0.85
Gly	GGT*	212	2.65	122	2.02
	GGC	84	1.05	49	0.81
	GGA	17	0.21	20	0.33
	GGG°	7	0.09	50	0.83
Leu	TTA°	29	0.42	63	0.99
	TTG°	52	0.76	95	1.5
	CTT*	127	1.85	69	1.09
	CTC°	11	0.16	33	0.52
	CTA*	143	2.08	70	1.1
	CTG	50	0.73	51	0.8
Ser	TCT*	81	1.84	45	0.94
	TCC*	17	0.39	3	0.06
	TCA*	97	2.2	76	1.59
	TCG°	15	0.34	63	1.32
	AGT°	12	0.27	56	1.17
	AGC	42	0.95	44	0.92
Arg	CGT*	129	2.93	61	1.56
	CGC*	103	2.34	46	1.17
	CGA°	21	0.48	54	1.38
	CGG°	0	0	10	0.26
	AGA°	9	0.2	40	1.02
	AGG°	2	0.05	24	0.61
Ile	ATT°	108	0.97	185	1.59
	ATC*	211	1.9	116	1
	ATA°	14	0.13	47	0.41

N_c values. The tRNA genes in the KVP40 and *V. parahaemolyticus* genomes were identified via the 'tRNAscan-SE' program (http://www.genetics.wustl.edu/eddy/tRNAscan- SE). Chi-square tests were employed in the evaluation of the significance of pairwise differences in codon and amino acid compositions. The CodonW 1.3 program (http://www.molbio.ox.ac.uk/cu) was employed in the calculation of most of the parameters, including the correspondence analysis (CA) of relative synonymous codon and amino acid usage.

The overall RSCU values of the 376 protein encoding genes of phage KVP40 indicate that the majority of its codons end in A and T (Table 1). KVP40 is expected to harbor an AT-rich genome. In order to determine whether there is any variation in codon usage among the genes of KVP40, the effective number of codons used by genes (N_c) and the (G + C) percentages at the third positions of codons (GC_3) were calculated. It was noted that N_c in KVP40 ranges from 24.89 to 50.32 with a mean of 37.96 and a standard deviation (SD) of 5.35, whereas GC_3 ranged from 21.6 to 54.2, with a mean of 37.87 and an SD of 5.07. Data suggest marked codon usage variations among the genes of KVP40 and factors other than mutational bias might also be relevant to variations in codon usage among the genes.

In order to determine the factors that influence variations in codon usage among the genes of KVP40, correspondence analysis was conducted on the RSCU values of its 376 genes. Only the distributions of the genes along the first two major axes were shown, as these accounted for 7.68% and 6.49% of the total variation (Fig. 1A). The first major axis is correlated positively with A_3 ($r=0.194$, $p<0.01$) but correlated negatively with T_3 ($r=-0.212$, $p<0.01$). No correlation was determined to exist between the positions of the genes along the first major axis and N_c. By way of contrast, the positions of the genes along the second major axis were correlated positively with both A_3 ($r=0.117$, $p<0.05$) and C_3 ($r=0.201$, $p<0.01$), and correlated negatively with G_3 ($r=-0.416, p<0.01$), and GC_3 ($r=-0.142$, $p<0.01$). Interestingly, the positions of the genes along the second major axis were correlated negatively with N_c ($r=-0.172$, $p<0.01$). These findings suggest that the G-ending codons are clustered on the negative side, whereas the codons ending in A and C predominate on the positive side of the second major axis. In order to determine the differences between the two gene clusters, the degree of codon usage variation in the 10% of genes located at the extreme positive side of axis 2 was compared with that of the 10% of genes located at the extreme negative side of axis 2. In order to estimate the degree of codon usage variation between these two gene sets, we conducted chi square tests, with a p value of <0.05 considered to be significant. Table 2 shows the RSCU values for each codon for the two gene groups. The asterisk represents codons whose occurrences are significantly higher in the genes situated on the extreme positive side of axis 2, as compared to the genes that were present on the extreme negative side of the second major axis. It is worthy of note that among the 20 over-represented codons on the extreme positive side of axis 2, there are 8 C ending-, 5 A ending-, and 7 T ending- codons. Expressed in percentages, this would be 40% C ending-, 25% A ending-, and 35% T ending-codons. According to the results, it might be concluded that pyrimidine bases are preferred in the highly expressed genes.

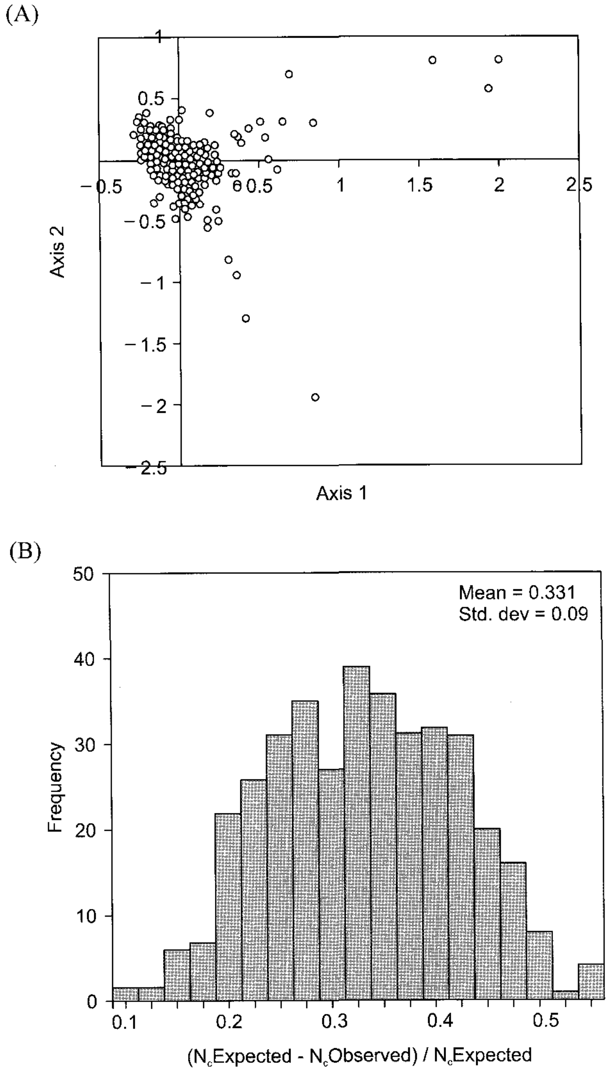

Fig. 1. (A) Positions of the KVP40 genes along the two major axes of variation in the correspondence analysis on RSCU values. The genes presented by the open circles. (B) Nc plot of phage KVP40 genes. See text for details.

Wright suggested that a plot of N_c vs GC_3 could be effectively used to determine codon usage variations among the genes. As was demonstrated by Wright (1990), comparisons of the actual distribution of genes, with the expected distribution under no selection, might be indicative of whether the codon usage bias of genes is influenced by factors other than mutational bias. If codon usage bias is dictated completely by GC_3, the values of N_c should fall on the expected curve between GC_3 and N_c. In other words, if codon usage bias is completely dictated by GC_3 composition, the difference between the observed and expected N_c values should be quite small in a majority of genes. In order to determine the possible influence of natural selection and mutational bias on synonymous codon usage in the KVP40 genome, we used the following calculation: $(N_{cExpected}-N_{cObserved})/N_{cExpected}$. The frequency distributions of $(N_{cExpected}-N_{cObserved})/N_{cExpected}$,

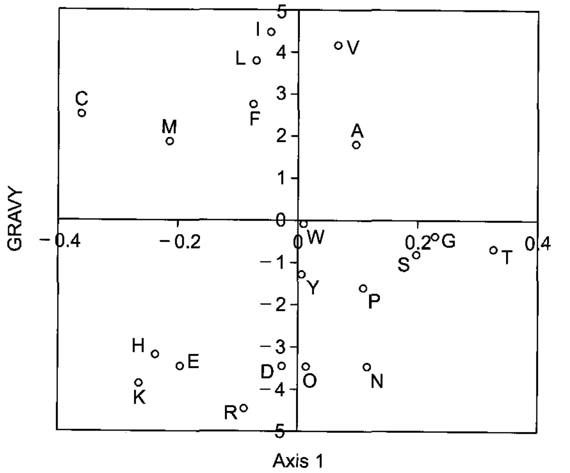

Fig. 2. Positions of amino acids along the first major axis in the correspondence analysis on amino acids. Single letter codes are used to show the positions of the 20 amino acid residues.

as shown in Fig. 1B, show that the majority of genes have substantial deviations between $N_{cObserved}$ and $N_{cExpected}$. These results corroborate the conclusion drawn from the above correspondence analysis, and affirm that the majority of genes in KVP40 exhibit additional codon usage bias, operating independently from mutational bias. The influence of mutational pressure on the evolution of synonymous codon usage variations had been previously demonstrated in studies conducted with certain bacterial viruses (Kunisawa *et al.*, 1998; Sahu *et al.*, 2004; Sahu *et al.*, 2005) and animal viruses from the order Nidovirales (Gu *et al.*, 2004).

In order to determine whether translation selection also influences codon usage variation in KVP40, we attempted to characterize the correlation, if any, between the synonymous codon usage of putatively highly expressed genes of KVP40 (Table 1), and the tRNA abundance of the host. Over-represented synonymous codons in the above gene types were initially detected via comparison of their RSCU values with those of the putatively low-expressed genes of KVP40. Next, the resulting data were compared with the copy number of tRNA *V. parahaemolyticus* species (Makino *et al.*, 2003), as cellular tRNA abundance in some organisms has been shown to be directly proportional to the tRNA copy number (Kanaya *et al.*, 2001). Among the 24 over-represented synonymous codons in the highly expressed genes of KVP40, 22 codons have been identified by the abundant tRNA species of *V. parahaemolyticus* (Table 1). Interestingly, 24 of the 35 over-represented codons of the low-expressed genes of KVP40 can also be identified by abundant tRNA species of *V. parahaemolyticus* (Table 1). By way of contrast, KVP40-specific tRNAs can be used to recognize 13 of 24 over-represented codons of the highly-expressed genes, and 20 of 35 over-represented codons in the low-expressed genes (Table 1). The data collected appear to suggest that the abundant tRNAs of *V. parahaemolyticus* are utilized preferentially in the expression of the putative highly-expressed genes of KVP40, whereas the tRNAs harbored by KVP40 may possibly augment the cellular levels of tRNA population, in order to synthesize all its proteins efficiently in a variety of hosts. Functions similar to those of the KVP40-specific tRNAs were also suggested to exist in the tRNAs of the mycobacteriophage, Bxz1 (Sahu *et al.*, 2004). The positive correlation between the abundant host tRNAs and synonymous codon usage had been previously demonstrated with the highly-expressed genes of bacteriophage T4, although its tRNAs had been suggested to preferentially express its lowly-expressed genes (Kunisawa, 1992).

Phage KVP50 and its *Vibrio* hosts, including *V. cholerae* and *V. parahaemolyticus*, are known to harbor AT-rich genomes (http://www.ncbi.nlm.nih.gov/ genomes). The RSCU values of the genes of the two previously mentioned Vibrio species indicate that the majority of their synonymous codons harbor either A or T at the third codon positions (data not shown), similarly to KVP40 (see above). Miller *et al.* (2003) reported that the codon usage of V. cholerae differs from that of KVP40 to a certain extent, although both harbor AT-rich genomes. In order to visualize this more clearly, we conducted correspondence analyses of the relative synonymous codon usage of all the *V. parahaemolyticus* and KVP40 genes together. We determined that the RSCU of KVP40 differs significantly from that of *V. parahaemolyticus* (data not shown).

To determine the factors influencing amino acid composition in KVP40, CA on relative amino acid usage of its 376 proteins had also been conducted. The data indicated that the first and second major axes of CA accounted for 17.03% and 10.46% of the total variations in the amino acid composition of KVP40 proteins, respectively. Further analysis showed that the first axis was correlated significantly with the GRAVY ($r=0.274$, $p<0.01$) and the mean molecular weight, MMWs ($r=-0.345$, $p<0.01$), whereas the second major axis is correlated significantly with aromaticity ($r=0.492$, $p<0.01$) and cysteine content ($r=0.470$, $p<0.01$) of KVP40 proteins. It was, in fact, determined that all of the charged amino acid residues of KVP40 proteins were distributed along the negative side of the first axis when the distributions of amino acids produced by the correspondence were plotted along the first major axis (Fig. 2). Similar correlations were also reported for *E. coli* (Lobry and Gautier, 1994) and *T. maritima* proteins (Zavala *et al.*, 2002).

The observed negative correlation between the first major axis and the MMWs appears to indicate that the KVP40 proteins located on the positive side of the first axis should preferentially harbor amino acid residues with lower MMWs. It was, indeed, determined that the first axis is correlated positively with each of the Gly, Ala, Ser, Pro, Val, and Thr residues (data not shown). Interestingly, cysteine residues, although cysteine is a low molecular weight amino acid, is absent in 84 of the KVP40 proteins. Aromatic amino acids that require excess energy for their biosynthesis are also rare in the KVP40 proteins. Smaller amino acid residues, which require comparatively less energy for their biosynthesis, have been shown to be prevalent in the highly expressed proteins of *G. lamblia* and *T. maritima* (Garat and Musto, 2000; Zavala *et al.*, 2002). By way of contrast, our analysis revealed that smaller amino acid residues are frequently utilized by both the putative highly-expressed and lowly-expressed proteins of KVP40 (data not shown). Phages depend utterly on their hosts for protein synthesis. Thus, the usage of smaller amino acid residues at higher frequency in KVP40 proteins may help this phage to economize the cost of its development in a diverse array of hosts. Recently, we reported that smaller amino acid residues are also incorporated preferentially into both the putative highly- and lowly-expressed proteins of *P. aeruginosa* phage PhiKZ (Krylov *et al.*, 2003; Sau *et al.*, 2005).

Acknowledgements

The authors would like to express their gratitude to the Department of Biotechnology, Government of India, for financial support.

References

Banerjee, T., S. Basak, S.K. Gupta, T.C. Ghosh. 2004. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. *J. Biomol. Struct. Dyn.* 22, 13-23.

Banerjee, T., S.K. Gupta, and T.C. Ghosh. 2005. Towards a resolution on the inherent methodological weakness of the "effective number of codons used by a gene". *Biochem. Biophys. Res. Commun.* 330, 1015-1018.

Basak, S.K., T. Banerjee, S.K. Gupta, and T.C. Ghosh. 2004. Investigation on the causes of codon and amino acid usages variation between thermophilic *Aquifex aeolicus* and mesophilic *Bacillus subtilis*. *J. Biomol. Struct. Dynamics.* 22, 205-214.

D'Onofrio, G., T.C. Ghosh, and G. Bernardi. 2002. The base composition of the genes is correlated with the secondary structures of the encoded proteins. *Gene* 300, 179-187.

Garat, B. and H. Musto. 2000 Trends of Amino acids usage in the proteins from the unicellular Parasite Giardia Lamblia. *Biochem. Biophys. Res. Commun.* 279, 996-1000.

Gu, W., T. Zhou, J. Ma, X. Sun, and Z. Lu. 2004. Analysis of synonymous codon usage in SARS *Coronavirus* and other viruses in the *Nidovirales*. *Virus Res.* 101, 155-161.

Gupta, S.K., S. Majumdar, T.K. Bhattacharya, and T.C. Ghosh. 2002. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem. Biophys. Res. Commun.* 269, 692-696.

Gupta, S.K. and T.C. Ghosh. 2001. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene* 273, 63-70.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2, 13-34.

Jenkins, G.M. and E.C. Holmes. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* 92, 1-7.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* 53, 290-298.

Kunisawa, T. 1992. Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and from its host *Escherichia coli*. *J. Theor. Biol.* 159, 287-298.

Kunisawa, T., S. Kanaya, and E. Kutter. 1998. Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res.* 5, 319-326.

Krylov, V., E. Pleteneva, M. Bourkaltseva, O. Shaburova, G. Volckaert, N. Sykilinda, L. Kurochkina, and V. Mesyanzhinov. 2003. *Myoviridae* bacteriophages of *Pseudomonas aeruginosa*: a long and complex evolutionary pathway. *Res. Microbiol.* 154, 269-275.

Levine, D.B. and B. Whitemore. 2000. Codon usage in nucleopolyhedroviruses. *J. Gen. Virol.* 81, 2313-2325.

Lobry, J.R. and C. Gautier. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* 22, 3174-3180.

Lynn, D.J., G.A. Singer, and D.A. Hickey. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* 30, 4272-4277.

Makino, K., K. Oshima, K. Kurokawa, K. Yokoyama, T. Uda, K. Tagomori, Y. Iijima, M. Najima, M. Nakano, A. Yamashita, Y. Kubota, S. Kimura, T. Yasunaga, T. Honda, H. Shinagawa, M. Hattori, and T. Iida, 2003. Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*. *Lancet* 361, 743-749.

Matsuzaki, S., S. Tanaka, T. Koga, and T. Kawata. 1992. A broad-host-range vibriophage, KVP40, isolated from sea water. *Microbiol. Immunol.* 36, 93-97.

Mclnerney, J.O. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA* 95, 10698-10703.

Miller, E.S., J.F. Heidelberg, J.A. Eisen, W.C. Nelson, A.S. Durkin, A. Ciecko, T.V. Feldblyum, O. White, I.T. Paulsen, W.C. Nierman, J. Lee, B. Szczypinski, and C.M. Fraser. 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriol.* 185, 5220-5233.

Naya, H., A. Zavala, H. Romero, H. Rodriguez-Maseda, and H. Musto. 2004. Correspondence analysis of amino acid usage within the family *Bacillaceae*. *Biochem. Biophys. Res. Commun.* 325, 1252-1257.

Oresisc, M. and D. Shalloway. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.* 281. 31-48.

Romero, H., A. Zavala, and H. Musto. 2000. Compositional pressure and translational selection determine codon usage in the extremely GC- poor unicellular eukaryote *Entamoeba histolytica*. *Nucleic Acids Res.* 28, 2084-2090.

Sahu, K., S.K. Gupta, S. Sau, and T.C. Ghosh. 2004. Synonymous Codon Usage Analysis of Mycobacteriophage Bxz1 and its plating bacteria *M. smegmatis*: identification of the Highly and Lowly Expressed Genes of Bxz1 and the possible function of its tRNA species. *J. Biochem. Mol. Biol.* 37, 487-492.

Sahu, K., S.K. Gupta, S. Sau, and T.C. Ghosh. 2005. Comparative analysis of the base composition and codon usages in fourteen mycobacteriophage genomes. *J. Biomol. Struct. Dyn.* 23, 63-71.

Sau, K., S. Sau, S.C. Mandal, and T.C. Ghosh. 2005. Factors influencing the synonymous codon and amino acid usage bias in an AT-rich *P. aeruginosa* phage PhiKZ. *Acta. Biochim. Biophys. Sin.* (Shanghai) 37, 625-633.

Sharp, P.M. and W.H. Li. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281-1295.

Sharp, P.M. and E. Cowe. 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* 7, 657-678.

Wright, F. 1990. The 'effective number of codons' used in a gene. *Gene* 87, 23-29.

Zavala, A., H. Naya, H. Romero, and H. Musto. 2002. Trends in codon and amino acid usage in *Thermotoga maritima*. *J. Mol. Evol.* 54, 563-568.