Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 22, Issue Number 1, (2004)

Evolutionary Forces in Shaping the Codon and Amino Acid Usages in *Blochmannia floridanus*

http://www.jbsdonline.com

T. Banerjee
S. Basak
S.K. Gupta
T.C. Ghosh*

Bioinformatics Centre
Bose Institute
P 1/12
C.I.T. Scheme VII M
Kolkata 700 054, India

Abstract

Endosymbiotic relationship has great effect on ecological system. Codon and amino acid usages bias of endosymbiotic bacteria *Blochmannia floridanus* (whose host is an ant *Camponotus floridanus*) was investigated using experimentally known genes of this organism. Correspondence Analysis on RSCU values show that there exists only one single explanatory major axis that is linked to the strand specific mutational biases. Majority of the genes have a tendency to concentrate on the leading strand, which may be related to the adaptive property related to the replication mechanisms. Amino acid usages were markedly different between the highly and lowly expressed genes in this organism and in particular, GC rich amino acids were found to occur significantly higher in highly expressed genes than the lowly expressed genes. Comparative analyses of the orthologous genes of *Escherichia coli* and *Blochmannia floridanus* show that highly expressed genes are significantly more conserved than lowly expressed genes. Based on our results we concluded that strand specific mutational bias is strongly operational in selecting the codon usage in this organism. Replicational-transcriptional selection can be invoked from the presence of majority of highly expressed genes in the leading strand. Conservation of GC rich amino acids in the highly expressed genes to its ancestor is the major source of variation in amino acid usages in the organism. Hydrophobicity of the genes is the second major source in differentiating the genes according to their amino acid usages in this organism.

Introduction

Codon usage is non-random and species specific (1). Moreover, codon usage bias differs significantly among the genes within the same organism (2). Biased codon usage may results from various factors. In extremely compositionally skewed unicellular organisms it has been reported that compositional biases are the only dictator in shaping the codon usage variation among the genes in those organisms (3-5). It has also been suggested that translational selection affects the codon usage in highly expressed genes and subsequently it has been advocated that preferred codons in highly expressed genes are recognized by most abundant tRNA genes (6-8). Since lowly expressed genes are less constrained by translational pressures, codon usage biases of these genes are mostly governed by mutational bias (9, 10). It has been shown that codon usage biases of *Drosophila* genes are governed by translational accuracy (11). Comparing the orthologous genes of *Mycobacterium tuberculosis* and *Mycobacterium leprae* it has been reported that gene expression, amino acid conservation and hydrophobicity are the main factors in determining the codon usage variation among the genes in those two *Mycobacterium* species (12). In some unicellular organisms it was reported that both translational and compositional constraints are operational in detecting the codon usage variation among the genes in those organisms (13-18). Codon usage of *Pseudomonas aeruginosa* is mainly dictated by translational selection rather than mutational biases though it is a high GC rich organism (19).

Phone: +91-33-2334 6626
Fax: +91-33-2334 3886
Email: tapash@bic.boseinst.ernet.in

In contrast, it has been reported that gene expression does not have any effect in influencing the codon usages in endosymbiotic bacteria *Wigglesworthia glossinidia brevipalpis* (20). A strong effect on codon usages has been reported due to the strand specific mutational biases in different organisms (21-24). Very recently it has been reported that organism's optimal growth temperature influences the codon bias of its genes and it has been argued that growth temperature exerts strong selection on codon usage (25).

Symbiotic relationship has great effect on our ecological system as well as on evolution of life on the earth (26). Parasitic relationship is well studied, but the reasons that insist bacteria to build the symbiotic relationship, is not clearly known (27). Bacterial symbioses widespread among insects and it have been estimated that at least 15-20% of insects live in this manner (28). This relationship may play great role for evolutionary success of insects, because it may help the insects to take new imbalanced food, which also helpful for maintain ecological balance (29). The bacterial transmission occurs vertically from generation to generation, known as vertical transmission. Vertically transmitted, obligate endosymbionts may have relatively small effective population size (N_e) caused by recurrent bottleneck on transmission between host generations and limited genetic recombination between endosymbiont of different hosts (30-32). As a result the efficacy of selection may be reduced in intracellular genomes compared to free living, recombining organisms, which have large effective population size. The probability of genetic drift is more in endosymbiotic bacteria whose effective population size is less. Most parasitic and symbiotic obligate intracellular bacteria show same genomic features i.e., bias toward a high A+T content (33) and massive genome size reduction with respect to their free-living ancestor (34).

Other than aphids and tsetse flies, ants are another social insects are particularly interested for investigating mutual relationship, because they have many interactions with different species of animals, plants, and microorganisms. Ants belong to a different insect order than tsetse and aphids. The biological function of this symbiosis is to control the nutritional system of ants (35).

Though the symbiotic relationship has great effect on ecology and also on evolution of life on the earth, the numbers of symbiotic bacteria whose complete genomes are available are very few, as there exist practical difficulties in cultivating these bacteria in laboratory (36). Recent availability of complete genome of *Blochmannia floridanus* (27) gives us an opportunity to analyse the genome wide amino acid and codon bias in this organism. Here we analyse the codon and amino acid usages in *B. floridanus,* an endosymbiont of the ant *Camponotus floridanus* (27) with an aim to understand the genetic organization of this organism. We find that strand asymmetry is the major cause of codon usage variation in this organism. Majority of the highly expressed genes are concentrated on the leading strand genes, suggesting replicational-transcriptional selection is operative in selecting the codon usage in this organism. In addition to that, it has also been observed that gene expression level has a profound effect in influencing their amino acid usages. Conservation of GC rich amino acids in the highly expressed genes to its ancestor has been observed to be the main source of variation in amino acid usages in the organism as observed in *Wigglesworthia glossinidia brevipalpis* (20).

Materials and Methods

The complete genome of *B. floridanus* has been downloaded from ftp.ncbi.nlm.nih.gov/genbank/genomes. Our own program developed in C was used to retrieve the coding sequences from the complete genome. To minimize sampling errors we have chosen only those sequences that are greater than or equal to 300 bp and have correct initial and termination codons. We also excluded hypothetical proteins. Finally 375 sequences were selected for data analysis.

Relative synonymous codon usage (RSCU) is defined as the ratio of the observed frequency of a codon to the expected frequency if all the synonymous codons for those amino acids are used equally (37). RSCU values greater than 1.0 indicate that the corresponding codons are used more frequently than the expected frequency whereas the reverse is true for RSCU value less than 1.0.

GC_{3s} is the frequency of (G+C) at the synonymous third positions of codons.

The effective number of codons used by a gene (N_c) is generally used to measure the bias of synonymous codons (38). The values of N_c range from 20 (when only one codon is used per amino acid) to 61 (when all codons are used in equal probability). The expected value of N_c under random codon usage is given by the following formula:

$$N_c = 2 + s + \{29/[s^2 + (1-s)^2]\};$$

Where $s = GC_{3s}$

All the parameters were calculated by using the programme CodonW 1.3 (available at www.molbiol.ox.ac.uk/cu). Correspondence analysis (CA) available in the CodonW program was used to investigate the major trend in codon usage variation among the genes (39). ORILOC program (available at http://pbil.univ-lyon1.fr/software/oriloc.html) was used to find the origin of replication (40). Orthologous sequences between *E. coli* and *B. floridanus* were identified by gapped BLASTP searches (41) using cutoff of $E = 10^{-3}$ and to avoid any ambiguity regarding orthology we have taken only those sequences whose gene name match in each of the two organism. Information on gene expression data on *B. floridanus* is scarce. For extracting highly and lowly expressed genes of *B. floridanus* we calculated the Codon Adaptation Index (CAI) of *B. floridanus* genes taking ribosomal proteins (which are known to be highly expressed in most of the bacterial species (42)) as a reference set. We then took 10% of genes at the two extreme ends of the CAI scales as highly expressed and lowly expressed genes respectively. In this way thirty-eight highly and thirty-eight lowly expressed genes were extracted for our final analysis. In *E. coli* it has been observed that gene expression levels are highly correlated with Codon Adaptation Index (CAI) (43). The CAI of *E. coli* was also calculated taking experimentally known highly expressed genes like ribosomal protein, heat shock protein and elongation factor as reference set. Pair-wise non-synonymous distance (d_N) between the genes of *B. floridanus* and its *E. coli* orthologs was calculated by using the method of Yang and Nielsen (44). The Students's *t* test was used to evaluate the significance of the pairwise differences in amino acid composition.

Results and Discussion

Overall Synonymous Codon Usage

RSCU values of 375 genes of *B. floridanus* shown in Table I indicate that A and/or T ending codons are predominant in this organism as observed in other endosymbiotic bacteria. Preferred A and/or T ending codons in their synonymous third codon positions are due to AT enrichment in this organism. However, overall codon usage analysis may hide some heterogeneity in codon usage among the genes that might be superimposed on extreme genomic composition of this genome. To find if there is some heterogeneity in codon usage between the genes two parameters namely, effective number of codons (N_c) and (G+C) percentage at the synonymous third codon positions (GC_{3s}) were used. The effective numbers of codons (N_c) range from 26.98 to 51.89 with a mean of 36.54 and standard deviation 2.74. The (G+C) percentage at the synonymous third codon positions (GC_{3s}) was found to vary from 6.70 to 22.60 with a mean of 13.00 and standard deviation

2.50. These results indicate that there exists a wide variation of codon usage among the genes in this organism.

Table I

Overall codon usage data of *B. floridanus* genes. RSCU represents relative synonymous codon usage values, calculated by summing over all the genes together. N is the number of codons, AA represents amino acid.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	4946	(1.84)	Ser	UCU	3519	(2.43)
	UUC	427	(0.16)		UCC	419	(0.29)
Leu	UUA	8306	(4.08)		UCA	2041	(1.41)
	UUG	1947	(0.96)		UCG	392	(0.27)
Tyr	UAU	4540	(1.79)	Cys	UGU	1738	(1.73)
	UAC	524	(0.21)		UGC	270	(0.27)
ter	UAA	284	(2.27)	ter	UGA	53	(0.42)
ter	UAG	38	(0.30)	Trp	UGG	1230	(1.00)
Leu	CUU	855	(0.42)	Pro	CCU	1805	(1.88)
	CUC	154	(0.08)		CCC	167	(0.17)
	CUA	754	(0.37)		CCA	1543	(1.61)
	CUG	185	(0.09)		CCG	327	(0.34)
His	CAU	2763	(1.77)	Arg	CGU	1618	(1.96)
	CAC	356	(0.23)		CGC	181	(0.22)
Gln	CAA	3930	(1.68)		CGA	1065	(1.29)
	CAG	743	(0.32)		CGG	197	(0.24)
Ile	AUU	7471	(1.62)	Thr	ACU	2906	(1.99)
	AUC	1017	(0.22)		ACC	447	(0.31)
	AUA	5374	(1.16)		ACA	2067	(1.42)
Met	AUG	2965	(1.00)		ACG	416	(0.29)
Asn	AAU	7431	(1.78)	Ser	AGU	1978	(1.37)
	AAC	932	(0.22)		AGC	325	(0.22)
Lys	AAA	7452	(1.71)	Arg	AGA	1566	(1.89)
	AAG	1260	(0.29)		AGG	344	(0.40)
Val	GUU	2815	(1.58)	Ala	GCU	2823	(2.00)
	GUC	232	(0.13)		GCC	265	(0.19)
	GUA	3104	(1.74)		GCA	2083	(1.48)
	GUG	973	(0.55)		GCG	475	(0.34)
Asp	GAU	16508	(1.46)	Gly	GGU	2462	(1.39)
	GAC	6060	(0.54)		GGC	265	(0.15)
Glu	GAA	4449	(1.68)		GGA	3643	(2.06)
	GAG	841	(0.32)		GGG	704	(0.40)

Various Factors in Determining the Codon Usage Variation

N_c Plot: Wright (38) in his classic paper demonstrated that a plot of N_c versus GC_{3s} could be effectively used to explore the codon usage variation among the genes. He suggested that the comparison of actual distributions of genes, with the expected distribution under no selection could be indicative, if codon usage bias of genes had some influence other than genomic GC composition. If codon usage bias is completely dictated by GC_{3s} the values of N_c should fall on the expected curve between

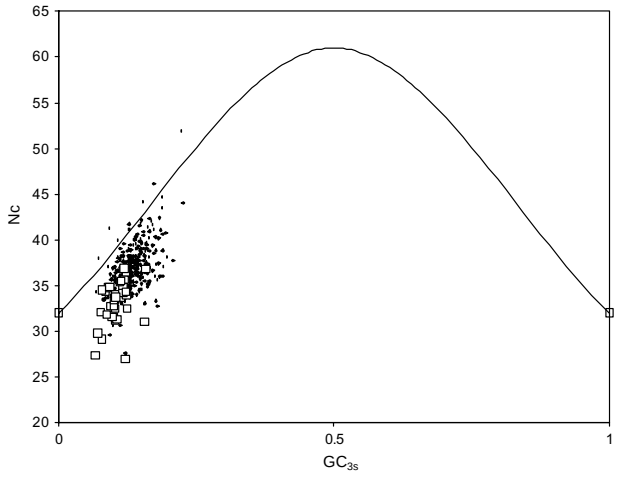

Figure 1: N_c plot of *B. floridanus* genes. The continuous curve represents the expected curve between GC_{3s} and N_c under random codon usage. Highly expressed genes are represented by squares.

GC_{3s} and N_c. However, N_c plot of *B. floridanus* shown in Figure 1 shows that a large number of points are lying well below the expected curve with low N_c values. It is also interesting to note that majority of highly expressed genes (represented by square symbols in Fig. 1) are lying well below the expected curve. These results indicate that a large number of genes including the highly expressed genes have additional codon usage bias that are independent of overall base composition.

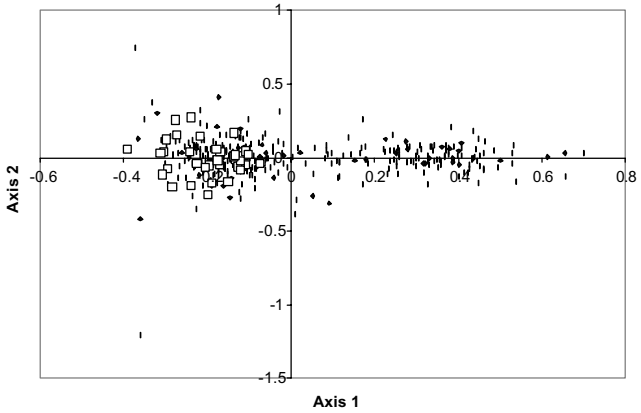

Figure 2: Positions of *B. floridanus* genes along the two major axes of variation in the correspondence analysis on RSCU values. Highly expressed genes are represented by squares.

Multivariate Statistical Analysis: Multivariate statistical analysis has been widely used to study the codon usage variation among the genes in different organisms. Correspondence analysis is one of the multivariate statistical technique in which the data are plotted in a multidimensional space of 59 axes (excluding Met, Trp and stop codons) and then it determines the most prominent axes contributing the codon usage variation among the genes. Correspondence analysis has been performed on RSCU values to minimize the effects of amino acid composition. Figure 2 shows the positions of the genes along the first and second major axes. The first axis accounted for 19.2% of the total variation, while none of other axis accounted for more than 5.2% of the total variation. Thus it can be said that in *B. floridanus* genome there is a single major explanatory axis on the synonymous codon usage variation among the genes. It is interesting to note that genes are separated along the first major axis according to the location of the genes on the leading or lagging strands. It was observed that 67% of the leading strand genes (a total of 239 genes transcribed on the leading strand; 160 genes are located on the left side of first major axis) and 56% of the lagging strand genes (a total of 136 genes transcribed on the lagging strand; 76 genes are on the left side of axis 1) are located on the negative side of axis 1. The positions of the genes on the first major axis are negatively correlated with T_3 ($r = -0.708$; $P < 0.0001$) and G_3 ($r = -0.841$; $P<0.0001$) and positively correlated with C_3 ($r = 0.950$; $P<0.0001$) and A_3 ($r = 0.727$; $P<0.01$). We got significant negative correlation between the positions of the genes along the first major axis with GT_3 ($r = -0.926$; $P<0.0001$) but we have not found any significant correlation with GC_{3s} ($r = 0.016$). These results suggest that local variation in synonymous GC_3 composition does not have any effect in variation of codon usage among the genes in this organism. Table II shows the cumulative codon usage of the genes of leading and lagging strand located at the extreme ends of axis 1 produced by CA on RSCU values. Simple chi square test was performed to assess the differences in codon usage between these two sets of gene taking $p<0.01$ as a significant criterion. The asterisk represents the codons whose occurrences are significantly higher in the leading strand genes. It is important to note that out of 25 statistically over represented codons in the leading strand genes, there are 12 G ending codons and 13 T ending codons. This actually represents 48% of G ending and 52% of T ending codons. G and T ending codons in the leading strand of replication has been explained by strand specific mutational biases in different organisms (45, 46) and such strand specific biases has been observed to take a dominant role in shaping the codon usage variation in different organisms (21, 22). Separation of genes along the first major axis according to their location on leading or lagging strands and a strong correlation between the positions of the

genes along the first major axis with GT_3 we can conclude that strand specific mutational bias has a profound effect in differentiating the genes along the first major axis in this organism. It is to be noted that all the highly expressed genes are lying on the negative side of the first major axis (denoted by squares). We also found significant negative correlation with the positions of the genes along the first major axis with expressivities ($r = -0.869$; $P<0.0001$). It is interesting to note though majority of the highly expressed genes are found on the leading strand they do not cluster together on the first major axis produced by CA on RSCU values. These results indicate that the translational selection may not be so strong to overcome strand specific mutational bias in this organism. In order to confirm our assumption we further performed CA on RSCU values only on the leading strand genes. When the positions of the genes along the first two major axes are plotted (data not shown) it was again observed that the highly expressed genes are not clustered together, though all the highly expressed genes are located on the negative side of the first major axis. The chi square test performed on codon usage frequencies between the highly expressed genes and the rest of the other leading strand genes shows that there are only three T ending codons which were found to be significantly higher in the highly expressed genes than the other genes (data not shown). These results further reinforce that translational selection is less operational than the strand specific mutational bias in selecting the codon usage in this organism. In some of the bacterial genomes it has been reported that majority of the genes are located on the leading strand and it was argued that selection at the

Table II

Cumulative codon usage of leading and lagging strand genes located at the extreme ends of axis 1 produced by CA on RSCU values. Superscript "a" denotes for leading strand genes and "b" for lagging strand genes. Asterisk represents the codons occurring significantly more often in the leading strand genes than that of lagging strand genes.

AA	Codon	RSCU[a]	N[a]	RSCU[b]	N[b]	AA	Codon	RSCU[a]	N[a]	RSCU[b]	N[b]
Phe	UUU*	1.95	(409)	1.65	(435)	Ser	UCU*	2.66	(310)	2.08	(305)
	UUC	0.05	(11)	0.35	(91)		UCC	0.06	(7)	0.70	(102)
Leu	UUA	3.98	(659)	3.86	(841)		UCA	1.21	(141)	1.77	(259)
	UUG*	1.56	(258)	0.22	(49)		UCG*	0.39	(45)	0.09	(13)
	CUU	0.30	(50)	0.56	(122)	Pro	CCU*	2.36	(157)	1.47	(153)
	CUC	0.01	(1)	0.29	(63)		CCC	0.02	(1)	0.47	(49)
	CUA	0.11	(18)	1.00	(217)		CCA	1.14	(76)	1.90	(198)
	CUG	0.04	(7)	0.07	(16)		CCG*	0.48	(32)	0.16	(17)
Ile	AUU*	1.72	(580)	1.33	(693)	Thr	ACU*	2.30	(212)	1.49	(285)
	AUC	0.08	(27)	0.45	(236)		ACC	0.11	(10)	0.67	(128)
	AUA	1.20	(405)	1.21	(632)		ACA	1.14	(105)	1.74	(334)
Met	AUG	1.00	(244)	1.00	(245)		ACG*	0.45	(41)	0.10	(19)
Val	GUU	1.63	(294)	1.47	(170)	Ala	GCU	2.00	(219)	1.80	(267)
	GUC	0.04	(8)	0.34	(39)		GCC	0.07	(8)	0.44	(65)
	GUA	1.58	(284)	2.07	(239)		GCA	1.43	(156)	1.63	(242)
	GUG*	0.75	(135)	0.12	(14)		GCG*	0.49	(54)	0.13	(20)
Tyr	UAU*	1.94	(350)	1.50	(403)	Cys	UGU*	1.94	(160)	1.31	(115)
	UAC	0.06	(11)	0.50	(136)		UGC	0.06	(5)	0.69	(60)
TER	UAA	2.11	(26)	2.68	(33)	TER	UGA	0.57	(7)	0.24	(3)
	UAG	0.32	(4)	0.08	(1)	Trp	UGG	1.00	(121)	1.00	(106)
His	CAU*	1.96	(157)	1.48	(344)	Arg	CGU*	2.33	(167)	1.25	(77)
	CAC	0.04	(3)	0.52	(120)		CGC	0.06	(4)	0.46	(28)
Gln	CAA	1.46	(219)	1.93	(549)		CGA	0.81	(58)	2.00	(123)
	CAG*	0.54	(80)	0.07	(19)		CGG*	0.43	(31)	0.05	(3)
Asn	AAU*	1.92	(491)	1.53	(760)	Ser	AGU*	1.60	(187)	0.90	(132)
	AAC	0.08	(21)	0.47	(233)		AGC	0.09	(10)	0.47	(69)
Lys	AAA	1.49	(475)	1.92	(903)	Arg	AGA	1.47	(105)	2.11	(130)
	AAG*	0.51	(163)	0.08	(40)		AGG*	0.91	(65)	0.13	(8)
Asp	GAU*	1.97	(421)	1.66	(400)	Gly	GGU*	1.65	(246)	0.93	(149)
	GAC	0.03	(7)	0.34	(83)		GGC	0.03	(5)	0.36	(57)
Glu	GAA	1.48	(304)	1.94	(463)		GGA	1.74	(259)	2.56	(410)
	GAG*	0.52	(106)	0.06	(15)		GGG*	0.58	(87)	0.16	(25)

level of replication is responsible for maintaining more genes in the leading strand (21, 22). The biased distribution of genes between leading and lagging strands is not a new observation. It was argued that co-oriented collisions of DNA and RNA polymerases happened in the leading strand, whereas head-on collisions happened in the lagging strand (47). Therefore, genes are preferentially positioned in the leading strand to avoid head-on collisions. It was also observed by French (48) that in *E. coli,* replication proceeded more slowly through a gene that transcribed in the opposite direction of replication. In other words it can be said that maintenance of majority of genes in the leading strand is due to the selective advantage to an organism at the level of replication. In this organism we also found majority of the genes (65%) are located in the leading strand of replication which can be described due to the differential replication rates between the leading and lagging strand genes. These results suggest that replicational selection is also operational in maintaining the majority of the genes on the leading strand of replication in *B. floridanus*. It was also observed that about 66% of the highly expressed genes are located on the leading strand and selective advantages for the transcription of these genes on the leading strand has been ascribed to overcome genetic drift and these genotypes may be advantageous to fixing in the population level.

To examine if amino acid composition exerts any constraint on synonymous codon usage we performed CA on simple codon count. Figure 3 shows the positions of the genes along the first two major axes. The first axis accounted for 24.56% of the total variation, while none of other axis accounted for more than 4.62% of the total variation. Thus it can be said that in *B. floridanus* genome there is a single major explanatory axis on the codon usage variation among the genes in this organism. The genes are separated along the first major axis according to their positions on the leading or lagging strand as observed on CA on RSCU values. The positions of the genes are strongly positively correlated with GT_3 ($r = 0.906$; $P<0.0001$) and but we have not found any correlation with GC_3. These results suggest that strand specific mutational bias is strong enough to discriminate the genes according to their codon usage along the first major explanatory axis and amino acid composition does not exert any constraint on this axis.

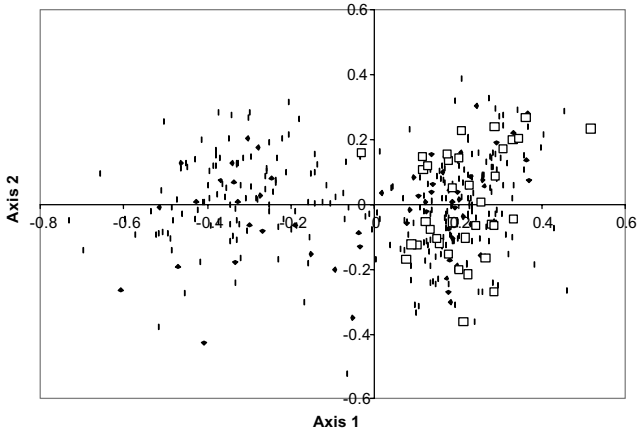

Figure 3: Positions of *B. floridanus* genes along the two major axes of variation in the correspondence analysis on codon count. Highly expressed genes are represented by squares.

Correspondence Analysis on Amino Acid Usages

Correspondence analysis on amino acid usages has been carried out to understand the possible evolutionary forces in defining the functional adaptations of the encoded proteins in this organism. Correspondence analysis on relative amino acid usages actually plot the data in a multidimensional space of 20 axes and then it determines the most prominent axes contributing the relative amino acid usage variation among the genes. CA on relative amino acid usages accounted for 22% and 9% of the total variation in protein amino acid content respectively. When the positions of the genes against the first two major axes are plotted (Fig. 4) it was observed that the majority of the highly and lowly expressed genes form distinct

groups along the horizontal axis. The positions of the genes along the first major axis are significantly correlated with GC_{12} (r = -0.796; P<0.0001) and GT_3 (r = -0.486; P<0.0001). We also got significant but weak correlation between the positions of the genes along the first major axis and the expressivities of genes (r = -0.517; P<0.0001). Since the majority of the highly expressed genes are located on the negative side of first major axis and there exists a negative correlation between positions of the genes along the first major axis and GC_{12} it is assumed that occurrences of GC rich amino acids will be higher in highly expressed genes than the lowly expressed genes. Moderate but significant negative correlation between the positions of the genes along the first major axis and GT_3 indicate that genes are separated along this axis according to their locations either on the leading or lagging strand of replication and in fact it was observed that 54% (130 out of 239) of the leading strand genes are located on the negative side of first major axis. The positions of the genes along the second major axis are highly correlated with the hydrophobicity of the genes (r = 0.803; P<0.0001).

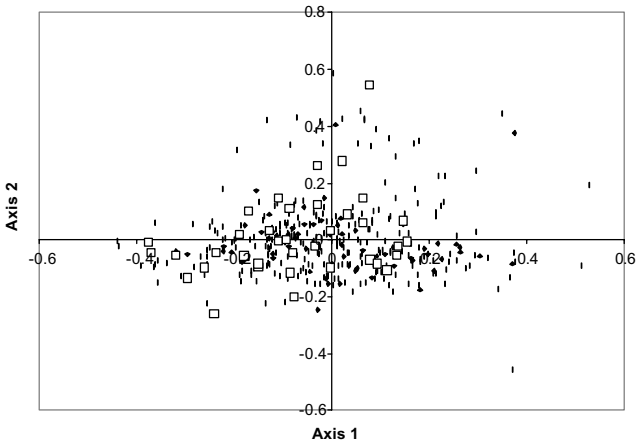

Figure 4: Positions of *B. floridanus* genes along the two major axes of variation in the correspondence analysis on amino acid usages. Highly expressed genes are represented by squares.

To see how the different amino acids are contributing towards amino acid usage variation among the genes along the first major axis we have plotted the distribution of amino acids on the first two major axes (Fig. 5). From the Figure 5 it is evident that GC rich amino acids namely Arg, Gly, and Ala are located on the extreme left of axis 1 whereas GC poor amino acids are located on the extreme right of axis 1. These results further confirm that GC rich amino acids are relatively more abundant in the highly expressed genes than the lowly expressed genes. The pair-wise analyses of the averages amino acid frequencies of highly and lowly expressed genes (Table III) show that there are significant increases in the frequencies of Arg and Gly in highly expressed genes. The similar trends were also observed in *Buchnera* and *Wigglesworthia* genomes. *B. floridanus* is an extremely AT rich organism and therefore it is expected that AT rich amino acids will be predominant

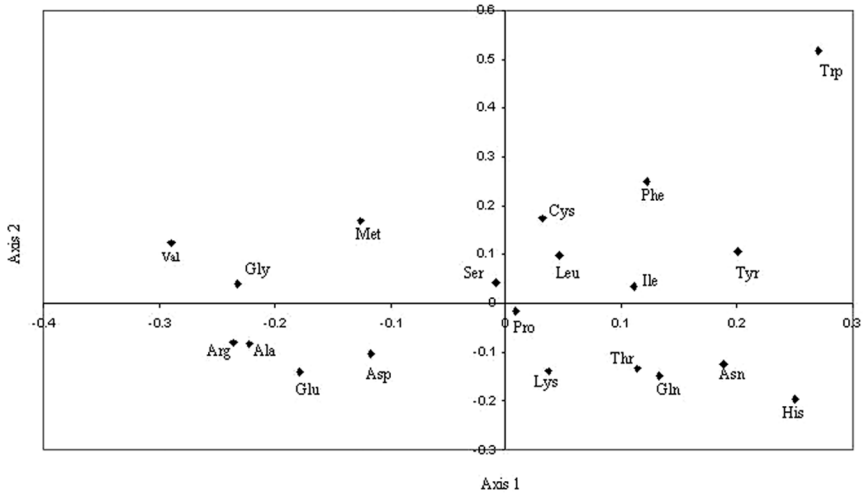

Figure 5: Distribution of synonymous amino acids along the first and second major axes of the correspondence analysis on amino acid usages.

in this organism and in fact it was also observed that AT rich amino acids are used more frequently in this organism. But comparison of amino acid frequency differences between the highly and lowly expressed shows that GC rich amino acids are used more frequently in highly expressed genes than the lowly expressed genes. This phenomenon can be explained by any of the following two ways: (i) Selection against AT rich amino acids at high expression genes and/or (ii) Overall conservation of GC rich amino acids in highly expressed genes between *B. floridanus* and its ancestor. In order to test the hypotheses we calculated the amino acid divergences among the orthologous sequences of *B. floridanus* and its close relative *E. coli*. Figure 6 shows the scattered plot between the amino acid divergences and gene expression levels taking CAI of *E. coli* genes. From the Figure 6 it is evident that gene expression level is significantly negatively correlated with amino acid divergence ($r = -0.386$; $P<0.001$), indicating highly expressed genes are more conserved than the lowly expressed genes. It is also interesting to note that there are a considerable number of genes having low CAI values also have relatively lower non-synonymous distances (d_N) (Fig 6). Recently micro-array data analyses of *E. coli* genome reveal that gene expression levels are highly positively correlated with the CAI, except with some relatively AT-rich genes where the values were found to be negatively correlated (49). In fact, the correlation value between amino acid divergence and gene expression levels increases substantially when we did the analysis excluding relatively high AT genes. These results further confirm that in *E. coli* AT rich genes, having extremely low CAI values are also highly expressed as observed earlier (49). We also calculated the average non-synonymous distance for the highly and lowly genes separately and it was observed that average non-synonymous distance for the lowly expressed genes (d_N=0.52) is significantly higher ($P < 1.8 \times 10^{-2}$) than the highly expressed genes (d_N=0.39). Frequent usages of GC rich amino acids in highly expressed genes of *B. floridanus* can be explained as the maintenance of same amino acid composition of ancestor. If GC rich amino acids are over-represented in highly expressed genes due to selection against AT rich amino acids one would expect changes of GC rich amino acids to AT rich amino acids will be higher in lowly expressed genes than the highly expressed genes. However, amino acid substitution matrix based on pair-wise alignment of orthologous sequences in both highly and lowly expressed genes shows that conversion of GC rich amino acids to AT rich amino acids do not differ significantly in any of the two categories of genes. These results suggest that variation in amino acid usages in *B. floridanus* is mainly influenced by the overall conservation of amino acid in highly expressed genes but not for selection against the use of AT rich amino acids at high expression *B. floridanus* genes.

Table III
Average amino acid (aa) frequencies of highly and lowly expressed genes of *B. floridanus* and p values of pair-wise comparisons.

Amino Acid	Highly Expressed Gene	Lowly Expressed Gene	p values
Phe	3.70	4.37	ns
Leu	8.22	11.90	4.3×10^{-13}
Ile	10.66	11.61	ns
Met	2.60	2.41	ns
Val	7.17	5.52	3.1×10^{-3}
Ser	7.48	6.63	ns
Pro	2.93	3.00	ns
Thr	4.50	4.86	ns
Ala	5.17	3.65	5.5×10^{-4}
Tyr	3.53	4.99	1.4×10^{-4}
His	2.28	2.55	ns
Gln	3.39	4.26	1.0×10^{-2}
Asn	6.38	6.78	ns
Lys	7.41	8.40	4.2×10^{-2}
Asp	4.78	4.24	ns
Glu	3.92	4.21	6.8×10^{-6}
Cys	1.16	2.02	1.3×10^{-4}
Trp	0.56	1.52	2.4×10^{-7}
Arg	5.23	4.15	4.8×10^{-2}
Gly	6.92	4.91	2.5×10^{-4}

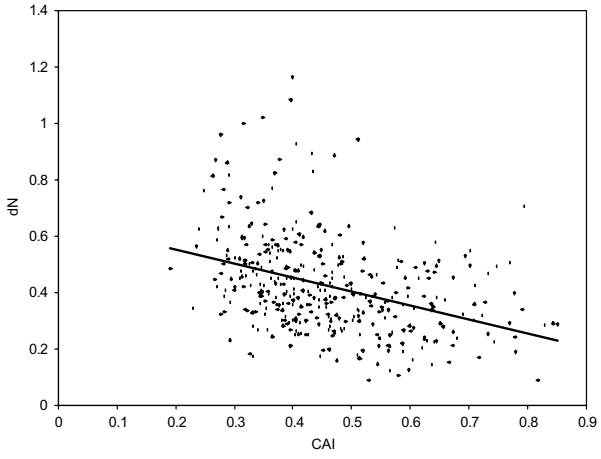

Figure 6: The scatter diagram between amino acid divergence and gene expression level of *B.floridanus* genes.

In conclusion it can be said that strand specific mutational bias is the major source of variation of codon usage among the genes of *B. floridanus* genes. Replicational and transcriptional selection on codon usage can be inferred from the presence of the majority of the genes in the leading strand as well as from the presence of the

majority of the highly expressed genes on the leading strand. GC_3 composition does not have any effect in codon usage variation among the genes in this organism. Amino acid composition does not have any influence in selecting the codon usage in this organism. Abundance of GC rich amino acids in the highly expressed genes can be explained by the overall conservation of amino acid at high expression genes. Hydrophobicity of the genes are the second major source in differentiating the genes according to their amino acid usages in this organism. The evolutionary forces in selecting the codon and amino acid usages described here has a lot of practical implications in understanding the common features generally found in endosymbiotic bacteria.

Acknowledgements

Authors are thankful to the Department of Biotechnology, Government of India, for the financial help. We are also thankful to the anonymous referees for their helpful comments in improving the manuscript.

References and Footnotes

1. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. *Nucl. Acids Res. 9*, r43-r74 (1981).
2. K. Wada, S. Aota, R. Tsuchiya, F. Ishibashi, T. Gojobori, and T. Ikemura. *Nucl. Acids Res. 18* (Suppl.), 2367-2411 (1990).
3. S. Ohkubo, A. Muto, Y. Kawauchi, F. Yamao, and S. Osawa. *Mol. Gen. Genet. 210*, 314-322 (1987).
4. F. Wright and M. J. Bibb. *Gene 113*, 55-65 (1992).
5. S. K.Gupta. T. K. Bhattacharyya, and T. C. Ghosh. *Ind. J. Biochem. & Biophys. 39*, 35-48 (2002).
6. T. Ikemura. *J. Mol. Biol. 146*, 1-21 (1981).
7. T. Ikemura. *J. Mol. Biol. 158*, 573-587 (1982).
8. J. L. Bennetzen and B. D. Hal. *J. Biol. Chem. 257*, 3026-3031 (1982).
9. P. M. Sharp and W.-H. Li. *Nucl. Acids Res. 19*, 7737-7749 (1986).
10. D. C. Shields and P. M. Sharp. *Nucl. Acids Res. 15*, 8023-8040 (1981).
11. H. Akashi. *Genetics 136*, 927-935 (1994).
12. A. B. de Miranda, F. A. Valin, K. Jabbari, W. M. Degrave, and G. Bernardi. *J. Mol. Evol. 50*, 45-55 (2000).
13. T. Ikemura. *J. Mol. Biol. 146*, 1-21 (1981).
14. M. Gouy and C. Gautier. *Nucl. Acids Res. 10*, 7055-7074 (1982).
15. S. G. Andersson and P. M. Sharp. *Microbiology 142*, 915-925 (1996).
16. H. Romero, A. Zavala, and H. Musto. *Gene 242*, 307-31 (2000).
17. T. C. Ghosh, S. K. Gupta, and S. Majumdar. *Int. J. Parasitol. 30*, 715-722 (2000).
18. S. K. Gupta, T. K. Bhattacharyya, and T. C. Ghosh. *J. Biomol. Str. and Dyn. 21*, 527-535 (2004).
19. S. K. Gupta and T. C. Ghosh. *Gene 273*, 63-70 (2001).
20. J. T. Herbeck, D. P. Wall, and J. J. Wernegreen. *Microbiology 149*, 2585-2596 (2003).
21. J. O. McInerney. *Proc. Natl. Acad. Sci. 95*, 10698-10703 (1998).
22. H. Romero, A. Zavala, and H. Musto. *Nucl. Acids Res. 28*, 2084-2090 (2000).
23. C. Palacios and J. J. Wernegreen. *Mol. Biol. Evol. 19*, 1575-1584 (2002).
24. C. Rispe, F. Delmotte, R. C. van Ham, and A. Moya. *Genome Res. 14*, 44-53 (2004).
25. D. J. Lynn, G. A. C. Singer, D. A. Hickey. *Nucleic Acids Res. 30*, 4272-4277 (2002).
26. L. Margulis and R. Fester. *Symbiosis as a Source of Evolutionary Innovation*. MIT Press, Cambridge, MA (1991).
27. R. Gil, F. J. Silva, E. Zientz, F. Delmotte, F.. Gonzalez-Candelas, A. Latorre, C. Rausell, J. Kamerbeek, J. Gadau, B. Holldobler, R. C. H. J. van Ham, R. Gross, and A. Moya. *Proc. Natl. Acad. Sci. 100*, 9388-9393 (2003).
28. P. Buchner. *Endosymbiosis of Animals with Plant Microorganisms*. Interscience, New York (1965).
29. P. Baumann, N. A. Moran, and L. Baumann. In *The Prokaryotes*. Ed., M. Dworkin. Springer, New York (2000).
30. N. Moran and P. Baumann. *Trends Ecol. Evol. 9*, 15-20 (1994).
31. D. J. Funk, L. Helbling, J. J. Wernegreen and N. A. Moran. *Proc. R. Soc. Lond. B 267*, 2517-2521 (2000).
32. J. J. Wernegreen and N. A. Moran. *J. Bacteriol. 183*, 785-790 (2001).
33. N. Moran. *Proc. Natl. Acad. Sci. 93*, 2873-2878 (1996).
34. F. J. Silva, A. Latorre, and A. Moya. *Trends Genet. 17*, 615-618 (2001).
35. M. Pfeiffer and K. E. Linsenmair. *Insectes Soe 47*, 123-132 (2000).
36. D. L. Mclean and E. J. Houk. *J. Insect Physiol. 19*, 625-633 (1973).
37. P. M. Sharp and W.-H. Li. *Nucl. Acids Res. 14*, 7737-7749 (1986).

38. F. Wright. *Gene 87*, 23-29 (1990).
39. M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London (1984).
40. C. A. Frank and J. R. Lobry. *Bioinformatics 16*, 560-561 (2000).
41. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. *Nucleic Acids Res. 25*, 3389-3402 (1997).
42. A. K. Srivastava and D. Schlessinger. *Annu. Rev. Microbiol. 44*, 105-129 (1990).
43. P. M. Sharp and W. H. Li. *Nucleic Acids Res. 15*, 1281-1295 (1987).
44. Z. Yang and R. Nielsen. *Mol. Biol. and Evol. 17*, 32-43 (2000).
45. M. P. Francino and H. Ochman. *Ann. N. Y. Acad. Sci. 870*, 428-431 (1999).
46. E. P. Rocha and A. Danchin. *Mol. Biol. Evol. 18*, 1789-1799 (2001).
47. B. J. Brewer. *Cell 53*, 679-686 (1988).
48. S. French. *Science 258*, 1362-1365 (1992).
49. M. dos Reis, L. Wernisch, and R. Savva. *Nucleic Acids Res. 31*, 6976-6985 (2003).

Date Received: March 28 2004

Communicated by the Editor Ramaswamy H Sarma