J Mol Evol (2012) 75:34–42
DOI 10.1007/s00239-012-9524-1

Transfer RNA Gene Numbers may not be Completely Responsible for the Codon Usage Bias in Asparagine, Isoleucine, Phenylalanine, and Tyrosine in the High Expression Genes in Bacteria

Siddhartha Sankar Satapathy · Malay Dutta · Alak Kumar Buragohain · Suvendra Kumar Ray

Received: 19 August 2012 / Accepted: 24 September 2012 / Published online: 2 October 2012

Abstract It is generally believed that the effect of translational selection on codon usage bias is related to the number of transfer RNA genes in bacteria, which is more with respect to the high expression genes than the whole genome. Keeping this in the background, we analyzed codon usage bias with respect to asparagine, isoleucine, phenylalanine, and tyrosine amino acids. Analysis was done in seventeen bacteria with the available gene expression data and information about the tRNA gene number. In most of the bacteria, it was observed that codon usage bias and tRNA gene number were not in agreement, which was unexpected. We extended the study further to 199 bacteria, limiting to the codon usage bias in the two highly expressed genes *rpoB* and *rpoC* which encode the RNA polymerase subunits β and β', respectively. In concordance with the result in the high expression genes, codon usage bias in *rpoB* and *rpoC* genes was also found to not be in agreement with tRNA gene number in many of these bacteria. Our study indicates that tRNA gene numbers may not be the sole determining factor for translational selection of codon usage bias in bacterial genomes.

Keywords Codon usage bias · Gene expression · Translational selection · tRNA gene number

Electronic supplementary material The online version of this article (doi:10.1007/s00239-012-9524-1) contains supplementary material, which is available to authorized users.

S. S. Satapathy · M. Dutta
Departments of Computer Science and Engineering, Tezpur University, Tezpur 784 028, Assam, India

A. K. Buragohain · S. K. Ray (✉)
Departments of Molecular Biology and Biotechnology, Tezpur University, Tezpur 784 028, Assam, India
e-mail: suven@tezu.ernet.in

Introduction

The twenty amino acids used for making proteins in the biologic system are generally coded by sixty one codons. The overabundance of codons allows multiple codons to code for the same amino acid which are called synonymous codons. The unequal occurrence of synonymous codons, termed as "*codon usage bias*," is a general feature in all genomes. Interestingly, the extent of codon usage bias is different among species as well as among the different genes within a genome (Ermolaeva 2001). Because it is a general feature in all genomes and each species is different from another with respect to its codon usage bias, there has been enormous interest among evolutionary biologists to understand the underlying forces. Nucleotide composition (G+C %) is a major force responsible for codon usage bias that brings a difference among genomes as well as within a genome—in A+T-rich DNA, synonymous codons ending with A/T predominate, whereas the reverse is also true for G+C-rich DNA. Regarding the evolution of nucleotide composition in DNA, theories in support of both selection and nucleotide substitution have been proposed (Muto and Osawa 1987; Chen et al. 2004; Hershberg and Petrov 2009). The recent findings in support of the selection view of genome composition (Hershberg and Petrov 2010; Hildebrand et al. 2010) argue against the presence of any neutral sites in bacterial genomes (Rocha and Feil 2010).

In the early 1980s, it was observed that high and low expression genes of *Escherichia coli* are different with respect to their codon usage bias (Gouy and Gautier 1982; Sharp and Li 1986a, b). It was also discovered that tRNA molecules corresponding to frequently used codons are more abundant in the cytosol of *E. coli* and yeast (Ikemura 1981; Ikemura 1985). From these observations, it was concluded that the evolution of codon usage bias in

organisms is a consequence of selection for high gene expression (Ikemura 1985). The greater extent of codon usage bias in high expression genes in comparison to that in low expression genes is a clear vindication of this (Bulmer 1991; Hershberg and Petrov 2009). This phenomenon known as translational selection of codon usage bias is an evolutionary response to higher and more accurate translation rates (Ran and Higgs 2010).

Transfer RNA influences translational selection of codon usage bias in two ways (Ran and Higgs 2010). Firstly, the cytosolic abundance values of isoacceptor tRNAs are not same. A synonymous codon with high cognate tRNA abundance is preferred to the other synonymous codons with low cognate tRNA abundance. Secondly, a tRNA molecule does not decode two or more synonymous codons with equal efficiency due to difference at the wobble position in codon-anticodon pairing. In addition, the codon-anticodon pairing is also influenced by tRNA base modifications (Ran and Higgs 2010). The codon-anticodon pairing rule for the amino acids with two set codons and four set codons is different. The exact quantitative contribution of each phenomenon toward codon usage bias is not known, which is evident from our inability to determine the exact codon usage bias given the expression of a gene, tRNA gene copy number, and codon-anticodon pairing. Due to these complexities, the exact contribution of tRNAs at the level of translational selection is difficult to quantify.

In this article, we have tried to assess the role of tRNA gene number on codon usage bias by considering the codons of the four amino acids asparagine (Asn), isoleucine (Ile), phenylalanine (Phe), and tyrosine (Tyr), which are encoded by two set codons (AUA is ignored in the case of Ile because of its low abundance) ending with pyrimidine (C/U). There is no isoacceptor tRNA for any of these amino acids because a single tRNA with G at the first anticodon position decodes both the synonymous codons as G can pair with U as well as with C at the wobble position. The other advantage of analyzing these amino acid codons is that G at the wobble position is very specific for its pairing with either C or U. This specificity of G limits its modification only to queuosine (Harada and Nishimura 1972) or 2′-*O*-methylguanosine (Arnold and Keith 1977) at the wobble position (Osawa et al. 1992). The C-ending codons are preferred to the U-ending codons in Asn, Ile, Phe, and Tyr due to more stable base pairing between G and C at the wobble position. Though there are other amino acids with two set codons ending with pyrimidine such as cysteine and glutamine, the first two nucleotides of these are not entirely made up of W nucleotides unlike the above four amino acids. W nucleotides in the first two positions of a codon favor a wobble pairing between G and C, to make the pairing stronger during translation (Grosjean and Fiers 1982). So, in Asn, Ile, Phe, and Tyr, the selected codons for translation are C-ending. This idea has been exploited by Sharp et al. (2005) to measure the strength of selected codon usage bias (S) in bacteria. Out of the eighty bacterial genomes studied by them, 30 % of the genomes exhibited weak selected codon usage bias. Using the approach of Sharp et al. (2005), dos Reis and Wernisch (2009) found out the strength of selected codon usage bias in several eukaryotes. This suggested that the selection for C-ending codons in these four amino acids is a general feature in different genomes.

Data regarding tRNA abundance values in cytosol are not available for many organisms. However, in several genomes, it has been shown that tRNA gene number and its abundance in cytosol correlate positively (Dong et al. 1996; Percudani et al. 1997; Kanaya et al. 1999). In addition, it is known that more frequently used codons (optimal codons) correspond to the highest tRNA gene numbers in a multicellular eukaryote *Caenorhabditis elegans* (Duret 2000). Combining the information about tRNA gene numbers from different bacterial genomes and growth rate of different bacteria, Rocha (2004) established the fact that growth rate in bacteria is a positive selection factor for higher tRNA gene number, which in turn is required for a high translation rate (Rocha 2004). In a different approach, by measuring the strength of selected codon usage bias in bacteria, Sharp et al. (2005) also demonstrated that growth rate is an important selection factor for codon usage bias among genes in bacterial genomes. The growth rate also correlated positively with tRNA gene numbers. This finding was extended by the observation of Higgs and Ran (2008) by quantifying the selection from tRNA gene number and showing its positive correlation with the bacterial growth rate. All these studies support the view that tRNA gene number can be considered as a selection factor for codon usage bias in bacteria.

In this study, we have compared the tRNA gene number with the extent of codon usage bias of Asn, Ile, Phe, and Tyr in high expression genes and in whole genomes in 199 bacteria. We proceed with the hypothesis that the extent of codon usage bias will be more for the amino acid with a higher number of tRNA genes. Surprisingly, in many bacteria, we do not find this correlation. Our results indicate that there may be some additional selection mechanisms contributing toward codon usage bias in genomes.

Materials and Methods

Genome sequences were taken from the DDBJ site (www.gib.genes.nig.ac.jp). Proteome data of *E. coli* were taken from Ishihama et al. (2008). Transcriptomic data for the sixteen different bacteria were downloaded from the

NCBI GEO website (ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/). Total tRNA gene numbers were collected from the Genomic tRNA Database (http://gtrnadb.ucsc.edu), which uses tRNAscan-SE to classify tRNA into different groups on the basis of their anticodon sequences (Lowe and Eddy 1997). Codon frequencies in the coding regions and the amino acid usage % were calculated using a computer program written in C language.

To compare codon usage bias with tRNA gene numbers, ranks were allotted to both tRNA gene number as well as the ratio between the C-ending and the U-ending codons in each genome. As we have considered only four amino acids in this study, the ranks given were in the order of 1, 2, 3, and 4. The minimum tRNA gene numbers were given the rank 1 and the next higher rank 2 and so on. The same tRNA gene number values were given the same rank. Ranks for the ratio of the C-ending and the U-ending codons were assigned in a similar way after rounding the values to the nearest integers. Using the above logic, a program was written in C language to calculate rank values.

Results

Similarity Between High Expression Genes and the Whole Genomes with Respect to their Amino Acid Composition

We analyzed the codon usage bias of Asn, Ile, Phe, and Tyr codons in seventeen bacterial genomes (Table 1). Codon usage bias was compared between the high expression genes (genes with the top 100 expression values) and the whole genomes of these bacteria. Prior to comparing the codon usage bias, the usage percentage of the four amino acids in the two groups of genes, i.e., the top 100 high expression genes and the whole genomes, was compared (Table 1). In general, abundance values of the four amino acids were in the order Ile > Asn > Phe > Tyr in bacteria. The average value for Ile was ~6 %, while the same for Tyr was ~3 % in a genome, which is close to the expected amino acid usage of 5 %, provided all amino acids were used equally. The amino acid usage was found to vary from less than one and half times between the two categories of genes in all bacteria except in two. In these two bacteria, *Pseudomonas aeruginosa* and *Streptomyces coelicolor*, the four amino acids were used more than threefold in the high expression genes than in the whole genome. On the other hand, while the amino acid usage in the whole genome in these two bacteria is different from that in the other genomes, the amino acid usage in the high expression genes in these species is similar to that in the other bacterial genomes (Table 1). It may be pointed out that the genomic G+C content of *P. aeruginosa* and *S. coelicolor* is high and the four amino acids considered in this study are encoded by AT-rich codons, which may be attributed to their lower usage in the whole genome. We therefore studied the usage of alanine (Ala), arginine (Arg), glycine (Gly), and proline (Pro) that are encoded by GC-rich codons in the two categories of genes in these genomes. In general, the usage of these four amino acids was observed to be very low in the high expression genes than in the whole genome in *P. aeruginosa* and *S. coelicolor* (Supplementary Table 1A). However, in the other bacteria with G+C-rich genomes (Table 1), the amino acid usage was found to be similar with respect to both the high expression genes and the whole genomes, thereby indicating that perhaps the abundance of GC in the genomes cannot be attributed as wholly responsible for the differential amino acid usage in these two bacterial species. This indicates that the high expression genes and the low expression genes in these two species have different amino acid usage—an observation which needs further investigation to find out the underlying reason(s).

Codon Usage Bias in High Expression Genes is not in Concordance with their Corresponding tRNA Gene Numbers in Bacteria

The tRNA gene number and synonymous codon frequencies in 100 high expression genes in seventeen bacteria are given in Table 2. Although the level of transcription might vary among the tRNA genes, the cellular tRNA abundance is expected to correlate with the tRNA gene number (Dong et al. 1996; Kanaya et al. 1999; Percudani et al. 1997; Duret 2000).

To study the extent of codon usage bias, we used a simple measure: ratio of abundance values of the C-ending codon by the U-ending codon (C/U ratio) for each of the four amino acids (Table 2). The C/U ratio gives an idea of the extent of preference for the C-ending codons over the U-ending codons in the genome. If the ratio is more than one, the C-ending codons are preferred, whereas the U-ending codons are preferred when the ratio is less than one. The data in Table 2 suggested positive correlation between C/U ratios and genome G+C %, which was expected. For example, the Pearson correlation coefficient between genome G+C % and Phe codon C/U ratio was significant (r value 0.65; p value < 0.01) and similar correlation results were found between genome G+C % and C/U ratio with respect to the other three amino acids. The other observation was in agreement with the theory of higher selection for C-ending codons in these amino acids as we observed a higher C/U ratio in the high expression genes than the corresponding ratio in the whole genome. There were some exceptions such as *Nitrosomonas europaea*, *Pseudomonas syringae*, and *Bradyrhizobium japonicum*, where the C/U ratio was lower in the high

Table 1 Usage (in %) of phenylalanine (Phe), asparagine (Asn), isoleucine (Ile), and tyrosine (Tyr) amino acids

S. no.	Strain	Group	Genome size	GC (%)	Top 100 HE genes					Whole genome				
					Total	Phe	Asn	Ile	Tyr	Total	Phe	Asn	Ile	Tyr
1	*S. aureus*	Firmicutes	2903636	32.84	33828	3.50	4.70	7.42	2.82	801462	4.48	5.69	8.59	3.93
2	*S. mutans*	Firmicutes	2030921	36.83	39508	3.60	5.06	6.85	3.26	579702	4.76	4.85	7.74	3.85
3	*L. monocytogenes*	Firmicutes	2944528	37.98	31671	3.72	4.58	6.96	3.2	870878	4.53	4.62	7.84	3.45
4	*H. influenzae*	γ Proteobacteria	1830069	38.15	25071	4.38	5.23	6.75	3.28	521077	4.48	4.88	7.11	3.14
5	*B. subtilis*	Firmicutes	4214630	43.52	32519	3.65	4.06	6.45	3.02	1228408	4.50	3.95	7.37	3.49
6	*L. plantarum*	Firmicutes	3348625	44.42	32119	3.63	4.58	6.54	3.50	920243	3.97	4.41	6.56	3.52
7	*E. coli*	γ Proteobacteria	4639675	50.00	22219	3.34	3.94	5.99	2.58	1313473	3.90	3.89	6.01	2.83
8	*N. europaea*	β Proteobacteria	2812094	50.72	22536	4.05	3.70	6.17	2.95	800071	3.92	3.58	6.39	2.84
9	*P. syringae*	γ Proteobacteria	6538260	58.34	26326	3.18	3.26	5.10	2.87	1814263	3.63	3.19	4.98	2.55
10	*B. longum*	Actinobacteria	2260266	60.13	30848	3.55	3.90	6.04	2.91	640513	3.41	3.43	5.37	2.68
11	*D. vulgaris*	δ Proteobacteria	3773159	63.28	26270	3.49	3.03	5.09	2.55	1020841	3.59	2.41	4.45	2.27
12	*B. japonicum*	α Proteobacteria	9105828	64.06	23090	3.55	3.61	5.29	2.70	2634346	3.73	2.76	5.26	2.21
13	*R. palustris*	α Proteobacteria	5467640	65.03	30075	3.91	3.96	5.46	3.20	1580833	3.64	2.58	5.26	2.22
14	***P. aeruginosa***	γ Pro teobacteria	6264404	66.56	21744	**3.59**	**3.68**	**5.05**	**2.72**	1855396	**1.35**	**0.96**	**1.1**	**0.93**
15	*R. sphaeroides*	α Proteobacteria	4603060	68.79	23416	3.74	3.25	5.56	2.72	943868	3.49	2.02	4.52	1.91
16	*T. thermophilus*	Deinococcus-thermus	2116056	69.50	23854	3.45	2.84	4.87	3.11	589163	3.77	1.55	2.67	2.87
17	***S. coelicolor***	Actinobacteria	9054847	72.00	27997	**2.68**	**2.61**	**4.02**	**2.33**	2533573	**0.87**	**0.91**	**0.97**	**0.71**

Table presents total number of codons and percentage of codons encoding four amino acids, phenylalanine (Phe), asparagine (Asn), isoleucine (Ile), and tyrosine(Tyr) in top 100 high expression (HE) genes as well as in the whole genome of seventeen bacteria (*Bacillus subtilis*, *Bifidobacterium longum*, *Bradyrhizobium japonicum*, *Desulfovibrio vulgaris Hildenborough*, *Escherichia coli*, *Haemophilus influenzae*, *Lactobacillus plantarum*, *Listeria monocytogenes*, *Nitrosomonas europaea*, *Pseudomonas aeruginosa*, *Pseudomonas syringae*, *Rhodobacter sphaeroides*, *Rhodopseudomonas palustris*, *Staphylococcus aureus*, *Streptococcus mutans*, *Streptomyces coelicolor*, *Thermus thermophilus*). *P. aeruginosa* and *S. coelicolor* are shown in bold because in these two bacteria, amino acid usage is very much different between the two groups of genes

expression genes than in the corresponding ratio in whole genomes for most of the cases. These genomes need further analyses to address this contradiction.

Our interest in this study was also to compare the codon usage bias with the tRNA gene numbers. We first compared the C/U ratio among the four amino acids both in the high expression genes as well as in the whole genomes and then compared the high C/U ratio in the high expression genes with the tRNA gene numbers in the genomes. We started with the bacteria *Streptococcus mutans*, *Nitrosomonas europaea*, and *Thermus thermophilus*, where the tRNA gene numbers were the same for all the four amino acids (Table 2). In *S. mutans* as well as in *N. europaea*, the C/U ratios among the four amino acids were found to be similar, as also their tRNA gene numbers. This is in agreement with the hypothesis that tRNA gene number is influencing codon usage bias. In *S. mutans*, the C/U ratio is more in the high expression genes than in the whole genome, indicating selection for the C-ending codons in the former. But, in *N. europaea*, the C/U ratio between the high expression genes and the whole genome was found to deviate from this, i.e., the C/U ratio for Phe, Tyr, and Ile codons was more in the whole genome than in the high expression genes. This suggested that either this bacterium has insignificant tRNA-mediated selection, which has already been reported by Sharp et al. (2005), or the U-ending codons are more favorably selected than the C-ending codons in this bacterium, which is somewhat unusual. A critical analyses of the C/U ratio in the high expression genes of Phe, Asn, Ile, and Tyr *vis-à-vis* the tRNA gene numbers together with the C/U values for the whole genomes reveal that the tRNA gene numbers probably do not have any role in determining the C/U ratios in both the high expression genes as well as in the whole genome. This is explicitly seen in the case of Ile and Asn, where there is sharp difference in the C/U values with respect to the two categories of genes. The scenario is witnessed with greater clarity in the case of *T. thermophilus*, where the C/U ratio in the high expression genes differs greatly [5.65 (Phe) to 134.4 (Asn)] in spite of the same tRNA gene numbers for the four amino acids. These analyses indicate that perhaps it is not the tRNA gene numbers alone that influences the C/U ratio in the two categories of genes as is believed to be the case.

We made similar analyses in *P. syringae* and *Desulfovibrio vulgaris*, where Phe and Tyr have one tRNA gene

Table 2 Transfer RNA gene number and synonymous codon usage in phenylalanine (Phe), asparagine (Asn), isoleucine (Ile), and tyrosine (Tyr) amino acids

S. no.	Name	tRNA gene number				Top 100 HE genes[$]				Whole genome[$]			
		Phe GAA	Asn GTT	Ile GAT	Tyr GTA	Phe UUC/ UUU	Asn AAC/ AAU	Ile AUC/ AUU	Tyr UAC/ UAU	Phe UUC/ UUU	Asn AAC/ AAU	Ile AUC/ AUU	Tyr UAC/ UAU
1	*S. aureus*	2	3	2	2	1.02	0.75	0.51	0.51	0.37	0.31	0.28	0.28
2	*S. mutans*	2	2	2	2	0.38	0.33	0.35	0.35	0.25	0.24	0.30	0.27
3	*L. monocytogenes*	2	4	3	2	1.11	1.18	0.67	0.79	0.47	0.45	0.36	0.46
4	*H. influenzae*	1	2	3	1	0.45	0.36	0.36	0.33	0.38	0.33	0.28	0.28
5	*B. subtilis*	3	4	3	2	0.78	1.36	1.06	0.81	0.46	0.77	0.73	0.53
6	*L. plantarum*	2	5	3	2	0.59	0.72	0.58	0.73	0.53	0.66	0.52	0.63
7	*E. coli*	2	4	3	3	2.42	4.54	1.99	1.88	0.74	1.24	0.83	0.76
8	*N. europaea*	1	1	1	1	0.70	0.72	0.74	0.62	0.98	0.68	1.49	0.68
9	*P. syringae*	1	2	5	1	1.38	3.03	2.30	0.94	1.81	2.36	2.25	1.80
10	*B. longum*	1	3	1	1	25.05	9.95	4.85	8.07	6.89	2.81	2.81	2.24
11	*D. vulgaris*	1	2	5	1	12.10	5.91	7.84	4.32	7.04	4.37	7.49	2.43
12	*B. japonicum*	2	2	1	1	2.55	1.77	3.12	1.51	4.70	2.21	7.16	1.15
13	*R. palustris*	1	1	2	1	12.21	4.54	10.27	2.46	6.24	2.40	8.19	1.33
14	*P. aeruginosa*	1	2	4	1	40.05	10.13	11.66	5.09	2.35	1.20	2.32	2.01
15	*R. sphaeroides*	1	1	3	0*	37.09	11.06	31.50	1.86	11.96	3.63	18.90	1.12
16	*T. thermophilus*	1	1	1	1	5.65	134.40	7.30	56.15	4.68	40.06	9.30	23.38
17	*S. coelicolor*	1	2	1	1	73.90	59.83	64.41	35.22	17.81	9.83	18.06	15.71

* No tRNA gene according the Genomic tRNA Database (http://gtrnadb.ucsc.edu)

[$] The ratio between abundance values of the two synonymous codons of an amino acid in the high expression (HE) genes and in the whole genome

each, while the number of tRNA genes for Asn and Ile is 2 and 5, respectively. The analyses revealed that in each of the cases with tRNA gene number 1, 2, and 5, there is very little variation in the C/U ratios in the high expression genes and the whole genomes. This again indicates that the tRNA gene number has a very small role in determining the C/U ratios in the high expression genes and in the whole genome.

Several other observations (Table 2) support our above analyses which indicated that the tRNA gene number was not the only selection mechanism to cause codon usage bias in these four amino acids. In the genome of *Bifidobacterium longum*, all the C/U ratios in the high expression gene were higher than that of the whole genome, which suggested translational selection. However, the highest C/U ratio 25.05 for Phe did not correspond to the highest tRNA gene numbers among the four amino acids. In this organism, there were three tRNA genes for Asn, which was higher than the tRNA gene number for the other three amino acids. The difference of C/U ratio between the high expression gene and whole genome was not reflected by the tRNA gene numbers because the fold difference for Phe was similar to the fold difference for Asn, whereas the former was with one tRNA gene and the latter was with three tRNA genes. The highest C/U ratio for Phe codon which is 25.05 is in contrast with the C/U ratios for the other three amino acids in high expression genes in *B. longum*. This raises the interesting question as to why such a variation occurs in codon usage bias among the high expression genes in this bacterium.

The observation in high expression genes is summarized in Table 3. We assigned a rank for each of the tRNA gene numbers among the four amino acids in a genome as discussed in the "Materials and Methods" section. We also worked out the ranks for each of the C/U ratios for the four amino acids in a genome. However, in genomes with low genome G+C % ($\leq$45.0), most of the C/U ratios were less than 1.00. When rounded to nearest integer, these values would converge into similar values which precluded realistic ranking in terms of C/U ratios. In order to avoid this difficulty in the ranking of C/U ratios in these genomes with low G+C %, we considered values within the ranges (0.00–0.25), (0.26–0.50), (0.51–0.75), and (0.76–1.00) to be similar and calculated the ranks accordingly. Then, we found the differences in the respective ranks, and the mean of the four difference values was considered. Theoretically, the maximum and minimum possible mean rank differences were 0.0 and 2.0, respectively. The mean rank difference of 0.0 indicated that preference for the C-ending codon over the U-ending codon falls in the same line as the tRNA gene numbers for all the four amino acids, whereas values larger than 0.0 indicated that preference for the C-ending codon does not match the tRNA gene number. In twelve out of seventeen genomes, the mean rank difference

Table 3 Absolute difference between the rank of the C/U ratio ratios in high expression genes and the rank of the tRNA gene number

S. no.	Name	Rank (tRNA gene number)				Rank (C/U ratio in top 100 HE genes[$])				Absolute rank difference				Mean rank difference
		Phe GAA	Asn GTT	Ile GAT	Tyr GTA	Phe UUC/UUU	Asn AAC/AAU	Ile AUC/AUU	Tyr UAC/UAU					
1	*S. aureus*	1	2	1	1	2	1	1	1	1	1	0	0	0.50
2	*S. mutans*	1	1	1	1	1	1	1	1	0	0	0	0	0.00
3	*L. monocytogenes*	1	3	2	1	3	3	1	2	2	0	1	1	1.00
4	*H. influenzae*	1	2	3	1	1	1	1	1	0	1	2	0	0.75
5	*B. subtilis*	2	3	2	1	1	2	2	1	1	1	0	0	0.50
6	*L. plantarum*	1	3	2	1	1	1	1	1	0	2	1	0	0.75
7	*E. coli*	1	3	2	2	1	2	1	1	0	1	1	1	0.75
8	*N. europaea*	1	1	1	1	1	1	1	1	0	0	0	0	0.00
9	*P. syringae*	1	2	3	1	2	4	3	1	1	2	0	0	0.75
10	*B. longum*	1	2	1	1	4	3	1	2	3	1	0	1	1.25
11	*D. vulgaris*	1	2	3	1	4	2	3	1	3	0	0	0	0.75
12	*B. japonicum*	2	2	1	1	2	1	2	1	0	1	1	0	0.50
13	*R. palustris*	1	1	2	1	4	2	3	1	3	1	1	0	1.25
14	*P. aeruginosa*	1	2	3	1	4	2	3	1	3	0	0	0	0.75
15	*R. sphaeroides*	1	1	2	*	4	2	3	1	3	1	1	*	1.67
16	*T. thermophilus*	1	1	1	1	1	4	2	3	0	3	1	2	1.50
17	*S. coelicolor*	1	2	1	1	4	2	3	1	3	0	2	0	1.25

[$] Ratio of the codon frequencies are rounded to nearest integer

* Rank is not calculated as there is no tRNA gene according the tRNA database

found was 0.75 or more, and those include *E. coli* and *B. longum*, which are known to have strong selection for codon usage bias. This indicated that the C/U ratio in these amino acids was not a clear reflection of the tRNA gene numbers in most of the genomes.

Codon Usage Bias in *rpoB* and *rpoC* Genes is not in Agreement with tRNA Gene Numbers in Many Bacterial Genomes

Finally, we made an attempt in validating our observation with other bacterial genomes. However, in view of the unavailability of expression data for such genomes, we chose two such genes which are universally highly expressed and are reasonably large in size. Accordingly, our analyses for codon usage of the above four amino acids in concatenated sequences of the two genes *rpoB* and *rpoC* encoding the β and β' subunits of RNA polymerase in bacteria, respectively, in 199 bacterial genomes were carried out (Supplementary Table 2). It is pertinent to point out that the findings derived from the analyses in the 17 bacterial genomes pertaining to the high expression genes are also valid with respect to *rpoB* and *rpoC* genes, justifying our use of these two genes in the analysis. In concordance with the above result with the top 100 high expression genes in the 17 bacterial genomes, the C/U ratios for the four amino acids were found to be higher in the *rpoB* and the *rpoC* genes than the whole genome. This suggested that the two genes are indeed under translational

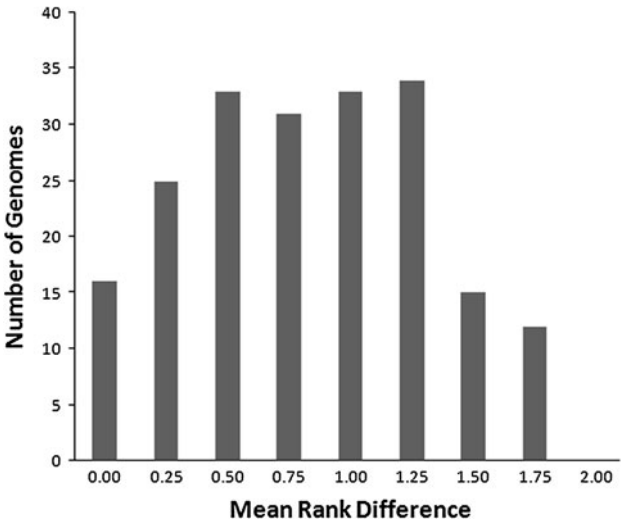

Fig. 1 A histogram presenting frequency distribution of mean rank difference For each of the four amino acids Phe, Asn, Ile, and Tyr, difference between the rank of tRNA gene numbers and rank for C/U ratios was calculated, and then mean of the four difference values was calculated for each genome. Figure presents the frequency distribution of the mean rank differences. Height of the *vertical bars* shows number of genomes with mean rank difference value shown in the *x*-axis

selection. Our analysis of codon usage in the *rpoB* and *rpoC* genes is summarized in Fig. 1. In most of the genomes, the mean rank differences are larger than 0.0 (Fig. 1). This is in agreement with the result obtained from the analysis using the high expression genes that the C/U ratio between the *rpoB* and *rpoC* genes and the whole genome is unrelated to the tRNA gene numbers in bacteria.

Discussion

In this study, we performed a comparative analysis of tRNA gene number with codon usage bias with respect to the amino acids Asn, Ile, Phe, and Tyr. Codon usage bias of these amino acids were studied earlier by Sharp et al. (2005) to calculate the strength of selected codon usage bias (S) in bacterial genomes and later by dos Reis and Wernisch (2009) to calculate S in eukaryotes. The assumption in their study was that the C-ending codons are favorably selected over the U-ending codons of these amino acids in the high expression genes. We also proceeded with the same assumption in our study here.

Out of the seventeen bacteria, in general, the C/U ratio was found to be more in the high expression genes than in the whole genomes except in *B. japonicum*, *N. europaea*, and *P. syringae*. The observation of higher C/U ratio in the high expression genes than in the whole genome in most of the bacteria suggested that the C-ending codons are favorably selected in the high expression genes in these bacteria. In the above three bacteria, the low C/U ratio in the high expression genes in comparison to the whole genome might be attributed to low translational selection (Sharp et al. 2005) as well as to the nucleotide compositional bias between the strands created due to the effect of replication and transcription (Francino and Ochman 1997; Ran and Higgs 2010). In addition, horizontal gene transfer might also be an important factor for the low C/U ratio in the high expression genes as has been observed in the case of *N. europaea* (Sharp et al. 2005).

The C/U ratios for all the four amino acids are high in *P. aeruginosa* as well as *S. coelicolor*, indicating high selection in these two organisms. But, according to Sharp et al. (2005), translation selection is moderately high for *S. coelicolor*, whereas the same is low for *P. aeruginosa*. The set of high expression genes used in our study (Supplementary Table 1A) is taken from the transcriptomic analysis, which is not all that similar to what was used by Sharp et al. (2005). Further, the genes used in the study consisted primarily of ribosomal protein encoding genes. We do not have the *S* value calculated for these two bacteria using the gene set we used here and therefore, we will not be able to comment conclusively on the results of Sharp et al. (2005).

The extent of codon usage bias difference between the high expression gene and whole genome varies considerably among the four amino acids and unexpectedly, in most of the cases, this variation could not be explained on the basis of tRNA gene numbers. This difference is also difficult to explain on the basis of tRNA modification of G at the wobble position because G-modification of the bases at wobble position has not been reported to affect the pairing with synonymous codons. The influence of nucleotide compositional bias and genome G+C causing this variation looks far from practical as their influence will affect equally codons of all the four amino acids within a gene.

According to our present understanding, translational selection is low in the low expression genes that constitute a major part of any genome. But, the C/U value of 40.06 in the case of Asn codons in comparison to 4.68 in the case of Phe codons of *T. thermophilus* is difficult to explain under the low selection bias. The selection for Asn codons looks strong even in the case of the whole genome/low expression genes. In the case of high expression genes, the value is 134.40, which is quite high. The role of tRNA gene copy numbers causing so much codon usage bias in the whole genome is unexpected because tRNA gene copy numbers are the same for Asn and Phe.

In general, we observed that the higher the codon usage bias for an amino acid codon in the whole genome, the greater the difference between the codon usage bias of the high expression gene and the whole genome for that amino acid. This indicated that the selection may be strong across the board in all genes within a genome, but in the high expression genes, it is stronger. The results indicate that an additional selection factor apart from tRNA genes influences translation and the selection mechanism is likely to be a contributing factor toward codon usage bias in the low expression genes as well. In support of our hypothesis, we argue that the four *S* values obtained for a single genome are going to be highly variable if the *S* value (Sharp et al. 2005) is calculated considering single amino acid codon usage bias. Had the codon usage difference between the high expression genes and the low expression genes been solely dependent on tRNA-mediated translational selection, we would not have expected the anomaly in the *S* values within a genome.

In support of the result obtained with gene expression data analyses in seventeen bacteria, the study made in 199 bacteria considering *rpoB* and *rpoC* genes also yielded similar result. The result obtained here is of significant interest considering the range of bacteria in which the analysis has been carried out. In addition, the result that was obtained with gene expression data could also be reproduced when *rpoB* and *rpoC* genes were used from the same bacteria. We therefore are of the view that the observation obtained with *rpoB* and *rpoC* genes in 199 bacteria also support that codon usage bias is perhaps not related to the tRNA genes' numbers in many bacteria.

The present study points out at some unknown selection forces which might be operating at the DNA structural level that are likely to be different in different genomes. For example, TA dinucleotide has been shown to be avoided universally in many genomes, but the extent of this avoidance is variable among the genomes (Karlin et al. 1998). The relative dinucleotide frequency in a genome remains similar in different parts of the genome, but the frequency is different among different genomes. So, dinucleotide frequency has been proposed to be used as the genome signature in organisms (Karlin et al. 1998). The significance of dinucleotide frequency as genome signature is not known. It is not reported whether TA avoidance is different between the genic and the non-genic regions within a strand, and similarly it is also not known if TA avoidance is different for high expression genes and the whole genome. But, because of its palindromic nature, TA abundance remains the same both in the leading and the lagging DNA strands. As a consequence of the differential TA avoidance within genomes at different contexts, if any, the avoidance of TA in genomes is likely to influence the C/U ratio in the four amino acid codons. This is because TA avoidance would positively select C at the third position of these four amino acids if the succeeding amino acid encoded by a codon begins with the A nucleotide. So, whether it is the selection on codon usage bias that determines dinucleotide signature or the reverse in the genome is an interesting question yet to be solved. As proposed for dinucleotides, trinucleotide or tetranucleotide might also be effecting codon usage variation. It has been shown that some palindromes are avoided in bacterial genomes because these are the sites for some restriction enzymes (Gelfand and Koonin 1997), which might be indirectly affecting the synonymous codon usage. It has already been reported that during translation, reading of a given codon is influenced by mRNA sequences external to the codon (Salser 1969). Bossi and Roth (1980) investigated this aspect of the decoding process and provided direct evidence in support of the influence of mRNA sequence outside the codon in determining translational efficiency. This influence of neighboring nucleotide surrounding a codon on choice of a codon from a synonymous group is known as context-dependent codon bias (Shpaer 1986; Gouy 1987). In this connection, the first nucleotide of the following codon is most important in determining context-dependent codon bias that can bring variation in dinucleotide frequency at the third codon position (Fedorov et al. 2002).

The present work is an attempt toward understanding codon usage bias in a fundamental perspective at the level of amino acids in contrast to the contemporary approach for identifying factors for such bias at a more global level. The important observations from this study can be summarized as follows: (i) The observations revealed by the present analysis are not fully in agreement with the generally accepted view that transfer RNA is responsible for the codon usage bias in the high expression genes; (ii) we hypothesize that there could be greater contribution of dinucleotides or other oligonucleotides in causing codon usage bias; and (iii) the results are not in concordance with the concept of coevolution of codon usage bias and tRNA gene number (Bulmer 1987) because our inferences drawn from our analysis in the present work are that the direction of evolution of tRNA gene numbers is a consequence of codon usage bias, but may not be the reverse. Although the study is limited to only four amino acids, the result of the analyses is of considerable significance. Also, the findings in this study is interesting considering the fact that the common factors such as genome G+C, strand compositional bias, and translational selection, which is known to effect codon usage in organisms, are unable to explain the high codon usage bias observed at the amino acid level.

Acknowledgments We thank Dr. Supek, the Rudjer Boskovic Institute, Croatia, for providing information for retrieving transcriptome data from the NCBI. We extend our thanks to Mrs. Madhusmita Dash, Qr. No. C-83, Tezpur University, for writing the C program for rank calculation. We also thank Dr. B.R. Powdel, Darang College, Tezpur, for his comments on the manuscript. We thank the two anonymous reviewers for their helpful comments on the previous version of this manuscript.

References

Arnold HH, Keith G (1977) The nucleotide sequence of phenylalanine tRNA from *Bacillus subtilis*. Nucl Acids Res 4:2821–2829

Bossi L, Roth JR (1980) The influence of codon context on genetic code translation. Nature 286:123–127

Bulmer M (1987) Coevolution of codon usage and tRNA abundance. Nature 325:728–730

Bulmer M (1991) The selection–mutation–drift theory of synonymous codon usage. Genetics 129:897–907

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained by genome wide mutational processes. Proc Natl Acad Sci USA 101:3480–3485

Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J Mol Biol 260:649–663

dos Reis M, Wernisch L (2009) Estimating translational selection in eukaryotic genomes. Mol Biol Evol 26:451–461

Duret L (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16:287–289

Ermolaeva MD (2001) Synonymous codon usage in bacteria. Curr Issues Mol Biol 3:91–97

Fedorov A, Saxonov S, Gilbert W (2002) Regularities of context-dependent codon bias in eukaryotic genes. Nucl Acids Res 30:1192–1197

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends Genet 13:240–245

Gelfand MS, Koonin EV (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucl Acids Res 25:2430–2439

Gouy M (1987) Codon contexts in enterobacterial and coliphage genes. Mol Biol Evol 4:426–444

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucl Acids Res 10:7055–7074

Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199–209

Harada F, Nishimura S (1972) Possible anticodon sequences of $tRNA^{his}$, $tRNA^{asn}$, and $tRNA^{asp}$ from *Escherichia coli* B. Universal presence of nucleoside Q in the first position of the anticodon of these transfer ribonucleic acids. Biochemistry 11:301–308

Hershberg R, Petrov DA (2009) General rules for optimal codon choice. PLoS Genet 5:e1000556

Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115

Higgs PG, Ran W (2008) Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25:2279–2291

Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6:e1001107

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D (2008) Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 9:102

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species specific diversity of codon usage based on multivariate analysis. Gene 238:143–155

Karlin S, Campbell AM, Mrázek J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185–225

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genome sequences. Nucl Acid Res 25:955–964

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Percudani R, Pavesi A, Ottonello S (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J Mol Biol 268:322–330

Ran W, Higgs PG (2010) The influence of anticodon–codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol 27:2129–2140

Rocha EPC (2004) Codon usage bias from tRNA's point of view, redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14:2279–2286

Rocha EPC, Feil EJ (2010) Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? PLoS Genet 6:e1001104

Salser W (1969) The influence of the reading context upon the suppression of nonsense codons. Mol Gen Genet 105:125–130

Sharp PM, Li WH (1986a) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Sharp PM, Li WH (1986b) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucl Acids Res 14:7737–7749

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucl Acids Res 33:1141–1153

Shpaer EG (1986) Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. J Mol Biol 188:555–564