Statistical Analysis of codon usage in extremely halophilic bacterium, *Salinibacter ruber* DSM 13855

Sanjukta RK[1], Farooqi MS[1*], Sharma N [1], Rai N[2], Mishra DC[1], Rai A[1], Singh DP[3] and Chaturvedi KK[1]

[1]Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Pusa, New Delhi – 110 012, India
[2]Central Drug Research Institute, Jhankipuram, Lucknow-226 021
[3]National Bureau of Agriculturally Important Microorganisms, Mau Nath Bhanjan, UP – 275 101, India
*E-mail: samir@iasri.res.in, Tel: 011-25847121-2 Ext 4305, Fax: 011-25841564

ABSTRACT

Synonymous codons are randomly distributed among genes, a phenomenon termed as codon usage bias. Understanding the extent and pattern of codon bias; the forces affecting codon usage are the key steps towards elucidating the adaptive choice of codon at the level of individual genes. Herein, trends in codon usage bias in a set of 1450 genes in *Salinibacter ruber*, an extremely halophilic bacterium have been evaluated. Notably, synonymous codon usage varies considerably among genes of this bacterium. Base composition (mutational bias) particularly C- and G-ending codons predominate with greater preference of 'C' at synonymously variable sites. The effect of natural selection acting at the level of translation has been observed. Certain genes with a high codon bias have been identified by multivariate statistical approach and investigations through various codon bias indices. These genes appear to be highly expressed, and their codon usage seems to have been shaped by selection favouring a limited number of translationally optimal codons. A subset of 27 optimal codons seems to be preferentially used in highly expressed genes. The frequency of these codons appears to be correlated with the level of gene expression, and may be a useful indicator in the case of genes (or open reading frames) whose expression levels are unknown.

Keywords: synonymous codon usage, mutational bias, multivariate statistical analysis, optimal codons.

INTRODUCTION

Microorganisms are adapted to unusual limits of abiotic factors such as temperature, pH, radiation, salinity etc. Salinity is an important deterrent to agriculture and its dynamic environment offers an excellent opportunity to enhance understanding of hyper saline physiology and genes related to salt tolerant. Halophilic microorganisms living in saline environments such as salt lakes, coastal lagoons, and man-made salterns (Pieper et al. 1998) are challenged by two stress factors (i) the high inorganic ion concentration and (ii) low water potential (Grammann et al. 2002). *Salinibacter ruber* is an extremely halophilic bacterium which requires at least 150g of salt/liter for growth. It grows optimally at NaCl concentrations between 200-300 g/liter (Anton et al. 2002; Corcelli et al. 2004).

Synonymous codon usage pattern is non-random and species-specific (Grantham et al. 1980; Gouy and Gautier 1982). The extent of this non-randomness is measured by Relative Synonymous Codon Usage (RSCU) (Sharp and Li. 1987). It has been reported that there is significant variation of synonymous codon usage bias among the different genes within the same organism (Ikemura 1985; Sharp and Li. 1986a). Studies show that genes with extremely biased codon usage are highly expressed. Further, highly expressed genes are enriched with specific codons. Apart from this, the pattern of codon usage in any gene reflects a complex balance among biases generated by mutation, selection and random genetic drift (Sharp and Cowe 1991; Gupta et al. 2004). In general, translational selection in nature and compositional constraints under the mutational pressure are considered to be two major factors accounting for codon usage variation among genes in various organisms (Muto and Osawa, 1987; Sharp et al., 1988; Andersson et al. 1996; Duret 2002; Sharp et al., 2005). The regulating mechanism of gene expression under salinity stress has little been examined in microorganisms. Moreover, the molecular basis of microbial resistance to salt stress is still not fully understood. Therefore, understanding of the molecular mechanisms involved in the halophilic adaptation of microorganism will not only provides insight into the factors responsible for genomic and proteomic stability under high salt conditions, but also, it is important for potential applications in agriculture. The availability of gene sequences in the public domain databases such as NCBI, EMBL, GenBank, DDBJ etc. enables to search for halophilic signatures of the microorganisms. In order to understand the basis of relevant mechanisms under halophilic condition for identification of gene expression, analysis of coodn usage pattern is very important (Ermolaeva 2001; Maria and Ermolaeva 2001; Lu et al. 2005).

In the present study, statistical analysis of codon usage of the genes of *S. ruber* were analysed to identify the highly expressed genes, their codon composition and various factors which are responsible for synonymous codon usage bias under salinity stress. The codon usage pattern was studied using several codon usage indices and their relationship with highly expressed genes has been obtained through various statistical methods such as correlation analysis and multivariate statistical analysis. The results were presented graphically for better understanding of whole phenomena. Further, the optimal codons were identified in highly expressed genes under extreme halophilic condition for this microorganism which may be used in identification of highly expressed genes of similar organism. This study may facilitate

the research on codon usage, ORF prediction and in understanding the mechanism of salt tolerance of microorganism.

MATERIALS AND METHODS

In order to perform synonymous codon usage analysis, the gene sequences (FASTA format) of *S. ruber* were retrieved from the Comprehensive Microbial Resource (http://www.tigr.org/CMR). The sampling errors were minimized by excluding sequence length less than 300bp and sequences with intermediate termination codons. Final dataset after exclusion consisted of 1450 genes (Table 1). Perl program has been developed for merging these gene sequences for further processing and analysis.

Table 1. List of number of genes of *S. ruber* on the basis of various functions

Sl. No.	Functions	No. of Genes
1	Amino Acid Biosynthesis	96
2	Biosynthesis Of Cofactor, Prosthetic Groups & Carriers	76
3	Cell Envelope	91
4	Cellular Processes	131
5	Central Intermediary Metabolism	12
6	Disrupted Reading Frame	7
7	DNA Metabolism	113
8	Energy Metabolism	232
9	Fatty Acid & Phospholipid Metabolism	52
10	Mobile and Extrachromosomal element functions	51
11	Protein Fate	135
12	Protein Synthesis	113
13	Purines, Pyrimidines, Nucleosides and Nucleotides	62
14	Regulatory Functions	67
15	Signal Transduction	27
16	Transcription	45
17	Transport & binding Protein	140
	Total	**1450**

The codon usage analysis of these 1450 gene sequences has been performed to study (i) base composition analysis of the codons (ii) codon usage bias.

Codon usage indices:

In order to study the base composition of the codons used by these genes, the different statistics have been calculated. The percentage of codons with different values of the nucleotides i.e. A, G, T and C at third position which is represented as A_{3s}, G_{3s}, T_{3s} and C_{3s} respectively were calculated for individual genes. Apart from this, values of total number of G and C nucleotides in gene i.e. GC content, frequency of codons with G or C at the third positions (GC_{3s}), GC skewness [(G-C)/(G+C)], AT skewness [(A-T)/(A+T)], GC_{3s} skewness [(G3s-C3s)/(G3s+C3s)], AT_{3s} skewness [(A3s-T3s)/(A3s+T3s)] were also calculated for each genes.

Measures of codon usage:

In order to investigate the characteristics of synonymous codon usage without the confounding influence of amino acid composition, the Relative Synonymous Codon Usage (RSCU) values among different codons in each gene was calculated. The RSCU value of the i^{th} codon for the j^{th} amino acid was calculated using following formula (Sharp and Li, 1986a).

$$\text{RSCU} = \frac{gij}{\sum_j^{nj} gij} . ni$$

Where, g_{ij} is the observed number of the i^{th} codon for j^{th} amino acid which has n_i type of synonymous codons. Here, RSCU values > 1.0 indicate that the corresponding codon is used more frequently than expected in a particular synonymous family, whereas, the reverse is true for RSCU values < 1.0 (Sharp and Li, 1986b). The effective number of codons of a gene (N_C) was also used to quantify the codon usage bias of a gene (Wright 1990). The value of N_C the overall estimate of absolute synonymous codon usage bias (Comeron and Aguade, 1998). The N_C value was calculated using following formula

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

where, s is the value of GC_{3s}

The N_c value ranges from 20 (when one codon is used per amino acid) to 61 (when all the codons are used with equal probability). The gene sequences in which N_c values are < 30 are considered to be highly expressed while those with > 55 are considered to be poorly expressed genes (Sharp et al. 1986b; Sharp and Cowe 1991; Sau et al. 2005).

Another measure for identification of gene expression is Codon Adaptation Index (CAI) (Sharp and Li, 1987); hydropathy (gravy) and aromaticity scores (Kyte and Doolittle, 1982; Lobry and Gautier 1994) of encoded proteins were also estimated.
The different properties of genes such as hydrophobicity, aromaticity and gene length were studied for further interpretation.

Statistical Analysis:

In order to study the linear relationship of base composition of codons with codon usage bias, correlation analysis has been done for the statistics obtained from base composition with N_c value as this will provide the measure of relationship between

the base composition of gene with codon usage bias. Further, in order to obtain the degree of relationship of gene expression with different statistics obtained from base composition, correlation analysis has been done with CAI and various statistical measures of base composition. In order to study the relationship of all 59 codons together with gene expression, multivariate statistical technique i.e. correspondence analysis (CA) has been applied. It has been found that CA has effectively been used to investigate the major trend in codon usage variation among genes (Abdi and Williams, 2010; Greenacre 1984). CA represents each gene as a 59 dimensional vector, and each dimension corresponds to the RSCU value of one sense codon (excluding AUG, UGG and three stop codons). In order to identify the difference between high and low expressed genes, the codon usage variation between 10% of the genes located at the extreme right of major axis and 10% of the genes located towards the extreme left produced by CA using RSCU values were compared. Chi squared contingency test ($P<0.01$) of the two groups were performed to estimate the optimal codons. Again, correlation analysis has been done using 10% of each highly expressed and lowly expressed genes to measure the degree of relationship between the value of major axis and various statistics such as base composition and Nc value. Correlation analysis was used for explanation of variation and association of gene feature values with axes scores generated through CA (Ewens and Grant 2001). This analysis was implemented based on the Pearson's correlation coefficient between various codon usage indices at $P<0.01$.

The values of the first major axis of CA will provide as an indicator of gene expression. In order to identify the highly expressed genes, the Chi squared contingency test ($P<0.01$) has been applied between top 10% genes having higher value of the major axis and 10% lowest genes having lowest value of the axis.

Software implementation:

CodonW 1.4.2 (Peden 1999) is employed for calculating the codon usage indices and CA data. SPSS 17.0 and SAS 9.2 are implemented for statistical analysis.

RESULTS AND DISCUSSION

Codon usage analysis:

The frequency of different codons in their respective synonymous family of 1450 genes along with its RSCU values within each family are shown in Table 1. From this table, the value of C_{3s} and G_{3s} can be derived as 44.6% and 37.2% respectively, whereas, it can be seen that value of T_{3s} and A_{3s} consist of 11.1% and 7.1% respectively. This suggests that the codons with G and/or C at the third position in all synonymous codon family are predominately used. Further, C ending codons are preferred over G ending codons. The result depicts maximum usage of acidic amino acids i.e. Asp, Glu, low proportion of hydrophobic amino acids and a high frequency of amino acids such as Gly and Ser. These results are in the line of work reported by Lanyi 1974; Oren and Mana, 1987 for halophilic protein. Furthermore, it is observed that the average percentage of GC content among genes is 65%, which is quite high and average value of GC skewness of the genes is low (0.0079) which is an indicative

of mutational bias. This clearly shows that, there is much greater preferential stability in the usages of codons with C and G nucleotides at the third position. RSCU value of all codons ending with C and G are >1.0 indicating the biased codon usage behaviour towards these codons in their respective synonymous codon family.

Further, the correlation analysis showed C and G at third codon position are negatively correlated (significantly, $P<0.01$) with N_c for correlation coefficient r=-0.74, -0.51 respectively while that of T and A are positively correlated (significantly, $P<0.01$) with Nc values, r=0.88, 0.89 respectively. Hence, it can be assumed that the influence of mutational bias of these genes is reflected in the choice of bases at the third codon position. However, this was expected since the optimal codons are, in general, chosen in accordance with the mutational bias of these genes. In other words, it is due to the translational selection that the mutational bias appears to be more prominent at the third codon position in highly expressed genes (Pan et al. 1998).

Table 2 Overall RSCU value for the genes of all functions

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	3642	0.58	Glu	GAG	17774	1.45
	UUC	9008	1.42	Ser	UCU	6012	0.63
Leu	UUA	1043	0.17		UCC	11095	1.17
	UUG	4289	0.71		UCA	7289	0.77
	CUU	4410	0.73		UCG	18538	1.95
	CUC	13142	2.17		AGU	3799	0.4
	CUA	2275	0.37		AGC	10404	1.09
	CUG	11254	1.85	Pro	CCU	8681	0.65
Ile	AUU	3295	0.75		CCC	15010	1.12
	AUC	9162	2.08		CCA	8183	0.61
	AUA	736	0.17		CCG	21749	1.62
Met	AUG	6241	1	Thr	ACU	4170	0.37
Val	GUU	3179	0.46		ACC	14511	1.29
	GUC	10211	1.47		ACA	6075	0.54
	GUA	2257	0.32		ACG	20205	1.8
	GUG	12217	1.75	Ala	GCU	7523	0.5
Tyr	UAU	1047	0.29		GCC	22967	1.53
	UAC	6278	1.71		GCA	9063	0.6
TER	UAA	602	0.26		GCG	20484	1.36
	UAG	742	0.32	Cys	UGU	3658	0.54
	UGA	5660	2.42		UGC	9831	1.46
His	CAU	4401	0.64	Trp	UGG	12693	1
	CAC	9333	1.36	Arg	CGU	9200	0.66
Gln	CAA	4209	0.61		CGC	19369	1.38
	CAG	9563	1.39		CGA	14869	1.06
Asn	AAU	2052	0.48		CGG	22871	1.63
	AAC	6450	1.52		AGA	6479	0.46
Lys	AAA	3090	0.61		AGG	11392	0.81
	AAG	7039	1.39	Gly	GGU	5912	0.45
Asp	GAU	5528	0.44		GGC	20183	1.52
	GAC	19546	1.56		GGA	10112	0.76

Glu	GAA	6763	0.55	GGG	16853	1.27

AA represents amino acid, N is the number of codons and RSCU represents relative synonymous codon usage.
Total number of genes=1450, total number of codons= 585618.

Heterogeneity of codon usage:

Two indices, *viz.* N_c and GC_{3s} are generally used to study the codon usage variation among the genes in any organism (Sahu et al. 2004). Figure 1 shows the N_c distribution of different genes in *S. ruber*. The N_c values range from 29.17 to 61 (with a mean of 37.85 and standard deviation of 5.73), indicating that there is a wide variation of codon usage bias among the genes. The heterogeneity of codon usage biases among the genes is further confirmed from the distributions of GC_{3s}, shown in Figure 2. It is obvious that GC_{3s} varies from 23 to 97% with a mean of 85% and standard deviation of 7.8%. These results indicate that apart from compositional constraints, other factors might influence the overall codon usage variation among the genes in *S. ruber.*

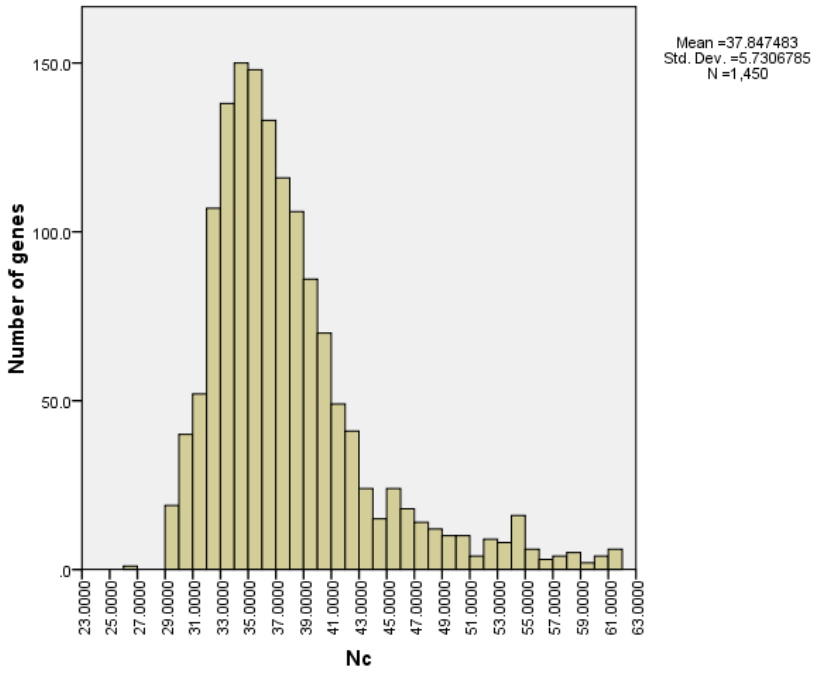

Figure 1: Distribution of effective number of codons (N_c) in *S. ruber* genes.

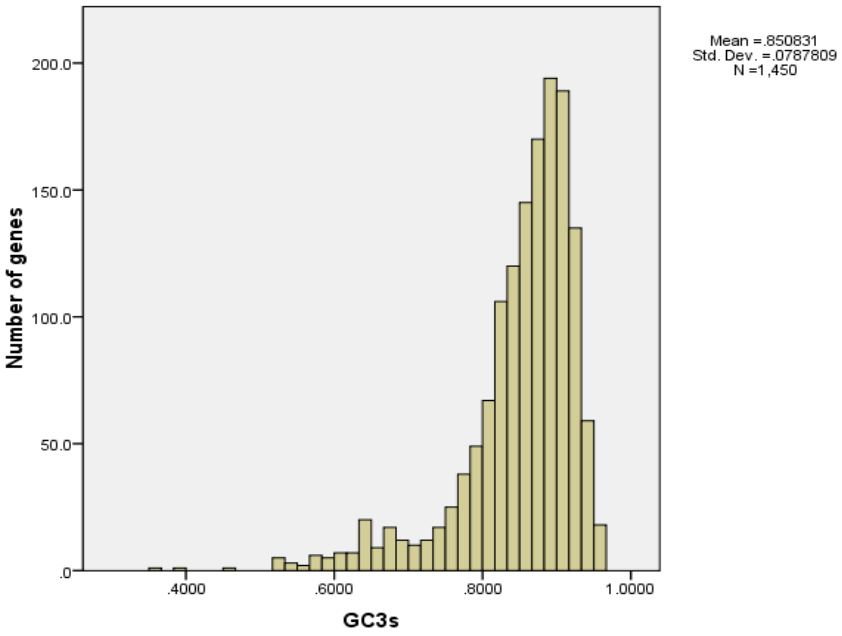

Figure 2: Compositional distribution of GC_{3s} in *S. ruber* genes.

Exploration of different factors in determining the codon usage variation

i. **The N_c plot:**

It was demonstrated by Wright 1990 that the comparison of the actual distribution of genes with the expected distribution under no selection could be indicative, if the codon usage bias of the genes had some influence other than the genomic GC composition. In order words, genes whose codon choice is constrained only by a G+C mutational bias, will lie on or just below the curve of the predicted value in the N_c plot (N_c versus GC_{3s}). From N_c plot (Figure 3), it is evident that few points lie on the expected curve towards GC rich regions, which certainly originates from the extreme compositional constraints. However, considerable number of points with low N_c values lie below the expected curve. This suggests that majority of genes have an additional codon usage bias apart from compositional bias. Moreover, it can be seen that almost all genes lie below the curve with low N_c value (30.60 to 42.74) and falls in the narrow range of higher GC_{3s} value (0.8 to 0.95). This suggests that translational selection is also responsible for codon bias among the genes. However, strong influence of compositional constraints on codon usage bias in the genes could be understood from the presence of significant negative correlation between GC_{3s} and N_c ($r = -0.94$, $P< 0.01$).

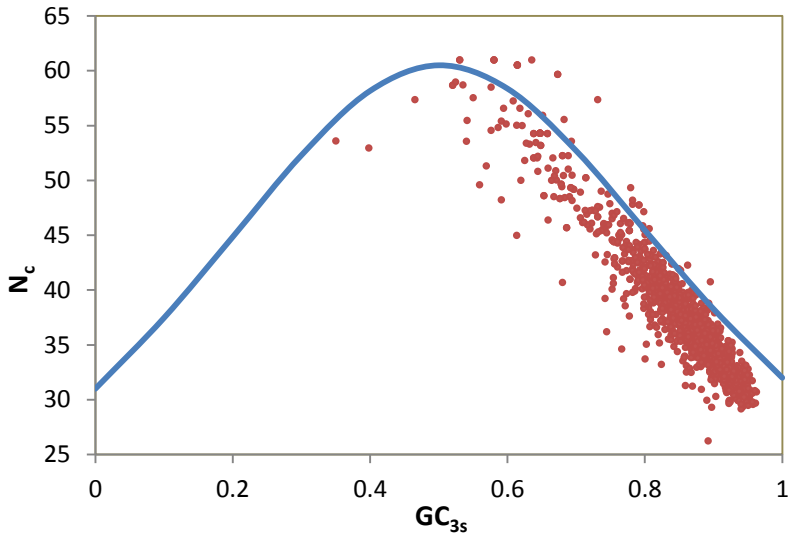

Figure 3: N_c plot of 1450 genes of *S. ruber*. The continuous curve represents the expected curve between GC_{3s} and N_c under random codon usage.

ii. **Multivariate Statistical Approach:**

The dataset of RSCU values of 1,450 genes is subjected to correspondence analysis (CA), a method of multivariate statistical analysis (MVA). For large multi-dimensional datasets, CA allows a reduction in the dimensionality of the data so that an efficient visualization that captures most of the variation can occur (Greenacre 1984). In this study, CA has been performed on RSCU values to minimize the effects of amino acid composition. The most prominent axes contributing to the codon usage variation among the genes are determined. It is seen that axis 1 has the largest fraction of the variation (12.61%) in the data. Axis 2 (6.79%) describes the second largest trend, and so on with each subsequent axis describing a progressively smaller amount of variation. It must be remembered that although the first axis explains a substantial amount of variation, its value is still lower than found in other organisms studied earlier (Eyre-Walker 1996). The low value might be due to the extreme genomic composition of this organism. It is also obvious from Figure 4 that the majority of the points are clustered around the origin of axes indicating that these genes have more or less similar codon usage biases. However, few points are widely scattered along the negative side of axis 1, which suggest that codon usage biases of these genes are not homogeneous. Genes with low N_c value indicate highly bias and vice-versa for the unbiased genes. It is noticeable from the figure that highly biased genes ($N_c \leq 35$) are scattered towards extreme positive side of axis 1 (highlighted with blue). Genes of N_c values (35 to 50) are dispersed in the middle while high N_c value (>50) genes are widely scattered towards the extreme negative side of the axis. Axis 1 is significantly positively correlated with G_{3s} ($r=0.55$, $P<0.01$) and C_{3s} ($r=0.75$, $P<0.01$) while significantly negatively correlated with A_{3s} ($r=-0.91$, $P<0.01$) and T_{3s} ($r=-0.92$, $P<0.01$).

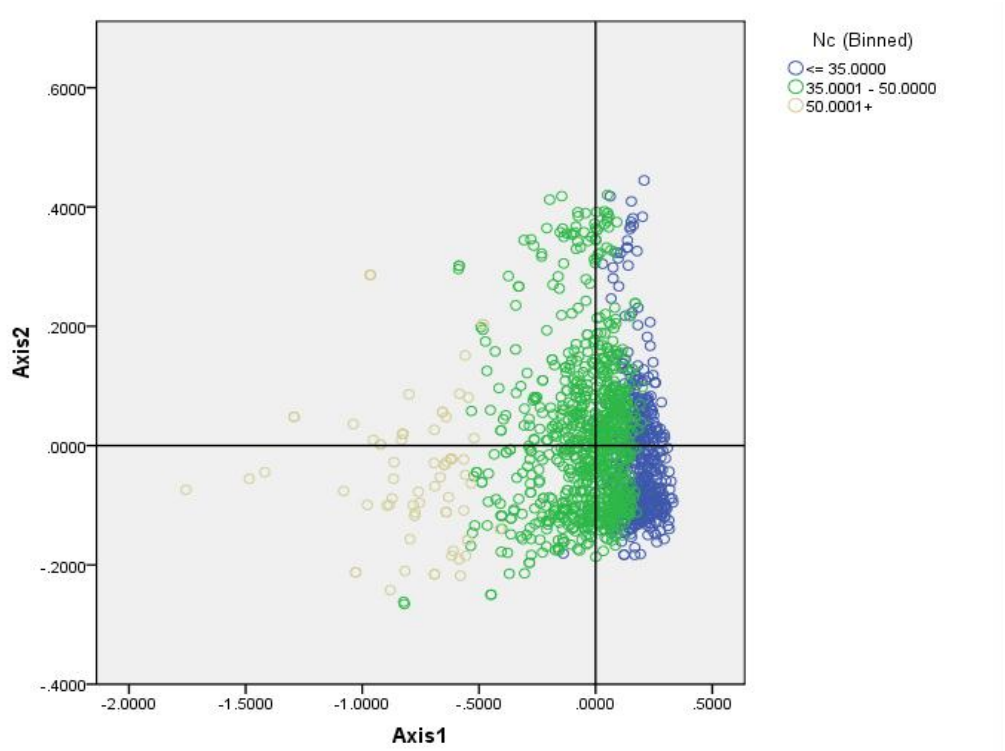

Figure 4: Positions of *S. ruber* genes along the two major axes of variation in the correspondence analysis on RSCU values.

Also, strong significant negative correlation exists between position of genes along the first axis with N_c (r=-0.95, P<0.01) and high degree of significant positive correlation with GC_{3s} (r=0.97, P<0.01). These findings suggest that highly biased genes, those ending with G and C, are clustered on the positive side, whereas those of A and T predominate on the negative side of the first major axis. Additionally, significant negative correlation is observed with N_c against GC_{3s} (r=-0.81, P<0.01) and GC (r=-0.43, P<0.01). Highly expressed genes tend to use C or G at the synonymous positions as compared to lowly expressed genes. It is also studied that C-ending codons are preferred over G-ending codons in highly expressed genes. Preference of C-ending codons in the highly expressed genes might be related to the translational efficiency of the genes as it has been reported that RNY (R-Purine, N-any nucleotide base, and Y-pyrimidine) codons are more advantageous for translation (Alvarez et al. 1994). Thus, compositional mutation bias possibly plays an important role in shaping the codon usage pattern of this organism.

iii. **Effect of gene expressivities on codon usage:**

It has been noted that in organisms with a highly skewed base composition, mutational bias is the main factor in shaping the codon usage variation among the genes whereas translational selection plays a minor role (Gupta and Ghosh 2001). Overall RSCU values (Table 1) and N_c plot (Figure 3) markedly indicate that mutational bias is key determinant of codon usage variation among the genes. However, correspondence analysis indicates that there is a single major trend in the codon usage among the genes in this bacterium. To assess the effect of expressivities of genes on codon usage biases, codon adaptation index (CAI) of *S. ruber* genes has been calculated. CAI has been considered as an

effective measure of gene expressivities (Gutierrez et al. 1996; Nakamura and Tabata 1997; Tiller and Collins 2000). The correlation coefficients are estimated for CAI values against the positions of genes along the first major axis, nucleotide compositions and N_c values.

Table 3 Correlation analysis data

	Axis1	N_c	GC_{3s}	G_{3s}	C_{3s}	A_{3s}	T_{3s}	GC
CAI	0.47*	-0.48*	0.42*	0.19*	0.68*	-0.36*	-0.36*	-0.11

* represents significantly correlated with probability, P<0.01

From Table 3, it is found that the gene expression level assessed by CAI value is significantly positively correlated with axis 1 and negative significant correlation with N_c values. A significant positive correlation between CAI and GC_{3s} content is noticed while CAI has negative correlation with GC, though lower negative value. From this analysis, it can be concluded that codon usage in genes of *S. ruber* is also affected by gene expression level. All the data suggest that genes with higher expression level, exhibiting a greater degree of codon usage bias and distributing at the right side of axis 1, are GC-rich and prefer to the codons with C or G at the synonymous variable site. As shown in Figure 5, it is interesting to note that there is a significant negative correlation between the positions of the genes along the first major axis and their corresponding CAI values, confirming that axis 1 is significantly correlated with the expression level of each gene of *S. ruber*.

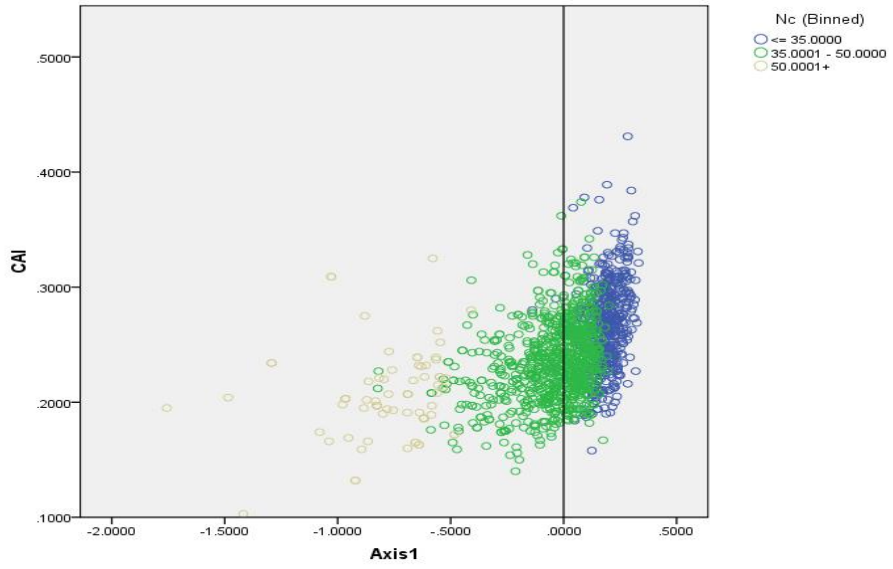

Figure 5: The Scatter diagram of gene position on axis 1 and CAI values

Correlation analysis of synonymous codon usage bias against hydrophobicity of each protein has also been investigated (r=0.22, P< 0.01). The findings indicate that genes, encoding more hydrophobic protein and bias to G/C bases at synonymous third codon positions, show a stronger codon bias which was also reported by Liu et al. 2010. Although the absolute value of this correlation

coefficient is low, it is statistically significant. Subsequently, it can be inferred that the hydrophobicity of the encoded protein play a minor role in affecting codon usage. However, no significant correlation has been observed between synonymous codon bias and aromaticity scores.

iv. **Relationship between codon bias and gene length:**

Selection for translational accuracy is predicted to have a positive correlation between codon bias and gene length. From the plot drawn with gene length against N_c (Figure 6), it is understood that shorter genes have a much wider variance in N_c values, and vice versa for longer genes. Lower N_c values in longer genes may be due to the direct effect of translation time or to the extra energy cost of proofreading associated with longer translating time. Correlation analysis of gene length against N_c, GC_{3s} and axis 1 was also examined. A significant negative correlation was observed with gene length against N_c ($r=-0.15$, $P<.01$). This revealed that gene length influences codon usage of these genes. Eyre-Walker 1996 has reported that the selection for accuracy in protein translation is likely to be greater in longer genes because the cost of producing a protein is proportional to its length. Therefore, selection of translational accuracy can be predicted to have positive correlation between codon usage bias (GC_{3s}) and gene length ($r=0.13$, $P<0.01$) as shown in Figure 7. In this study, the results of correlation analyses between gene length and the genes positions on axis 1 ($r=-0.10$, $P<0.01$) showed significant correlation. The findings indicated that more biased genes, with longer length, higher expression level and higher GC_{3s} values, are distributed at the left side of the first axis and vice versa for shorter genes. Subsequently, we supposed that gene length had an effect on codon bias. However, these correlation coefficients were far less than that of nucleotide composition. Therefore, nucleotide composition should be the major source of codon usage variation, while the gene length seemed to play a minor role in shaping codon usage in *S. ruber*.

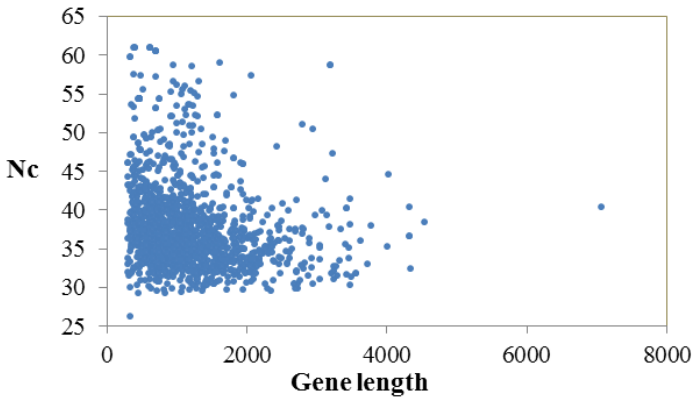

Figure 6: Plot of gene position on axis1 versus gene length

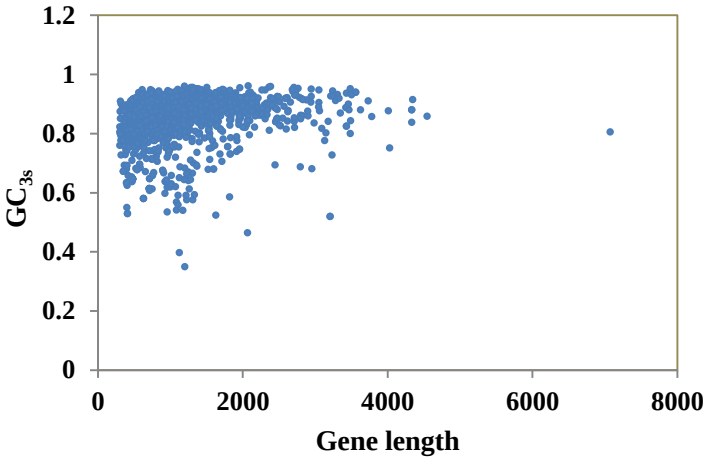

Figure 7: Plot of GC_{3s} versus gene length

v. **Translational optimal codons:**

In order to identify the optimal codons, 10% of genes each from both extremes of axis 1 were analysed (Table 4). A set of twenty-seven codons were determined as the optimal codons using χ^2 test at P<0.01. Out of 27 codons, 15 codons are ending with C nucleotide, whereas, 12 codons end with G accounting for 56% C ending while 44% G ending codons. Codons (UUC, UAC, AUC, AAC, GAC and GGU) are optimal in *E. coli, B. subtilis, S. cerevisiae, S. pombe, and D. melanogaster* (Sharp and Devine 1989) and are almost always preferentially used in highly expressed genes; similarly certain codons are commonly avoided in highly expressed prokaryotic genes (AGG, AGA). The predicted optimal codons corroborate with these observations. These optimal codons might be significant to introducing point mutation, and modifying heterologous genes in order to increase the product of specific protein. Ikemura 1981 showed that there is a match between these codons and the most abundant tRNAs. It has been reported that highly expressed genes have a strong selective preference for codons with a high concentration for the corresponding tRNA molecule (Moriyama et al. 1997; Duret 2000). This trend has been interpreted as the coadaptation between amino acid composition of protein and tRNA-pools to enhance the translational efficiency. Remarkably, in this study, there is a strong positive correlation (r = 0.84, P <0.01) between the frequency of optimal codon (F_{op}) in each gene and respective CAI value. This strongly suggests that translational selection influence the codon usage of *S. ruber* and the optional codons are more frequent in highly expressed genes.

Table 4 RSCU for the highly and lowly expressed genes highlighting translational optimal codons.

AA	Codon	RSCU[1]	N[1]	RSCU[2]	N[2]	AA	Codon	RSCU[1]	N[1]	RSCU[2]	N[2]
Phe	UUU	0.36	214	0.91	333	Glu	GAG*	1.77	3017	1.17	1020
	UUC*	1.64	975	1.09	395	Ser	UCU	0.07	20	0.86	250
Leu	UUA	0	0	0.21	68		UCC*	1.75	518	1.13	326
	UUG	0.07	40	0.77	245		UCA	0.03	9	0.71	206

	CUU	0.23	131	1.2	382		UCG*	2.11	624	1.23	356
	CUC*	3.02	1719	1.79	572	Ser	AGU	0.16	46	0.76	219
	CUA	0.05	29	0.49	158		AGC*	1.89	558	1.31	380
	CUG*	2.63	1496	1.54	491	Pro	CCU	0.03	14	0.98	283
Ile	AUU	0.47	269	0.98	315		CCC*	1.71	714	0.78	224
	AUC*	2.53	1439	1.66	531		CCA	0.04	16	0.88	253
	AUA	0	1	0.36	116		CCG*	2.22	926	1.36	393
Met	AUG	1	838	1	350	Thr	ACU	0.02	10	0.67	209
Val	GUU	0.09	63	0.72	274		ACC*	1.95	1101	1.28	400
	GUC*	1.58	1127	1.34	509		ACA	0.05	27	0.7	220
	GUA	0.06	41	0.64	245		ACG*	1.98	1117	1.35	423
	GUG*	2.27	1621	1.29	492	Ala	GCU	0.03	30	0.93	417
Tyr	UAU	0.07	35	0.74	243		GCC*	2.48	2366	1.23	551
	UAC*	1.93	953	1.26	414		GCA	0.07	70	0.79	351
TER	UAA	0.58	14	0.67	39		GCG*	1.41	1346	1.05	467
	UAG	1.5	36	0.79	46	Cys	UGU	0.13	17	0.79	125
	UGA	0.92	22	1.53	89		UGC*	1.87	251	1.21	190
His	CAU	0.05	23	0.85	270	Trp	UGG	1	354	1	404
	CAC*	1.95	830	1.15	364	Arg	CGU	0.17	73	0.8	327
Gln	CAA	0.08	56	0.72	326		CGC*	3.74	1611	1.2	487
	CAG*	1.92	1313	1.28	581		CGA	0.14	60	1.19	486
Asn	AAU	0.1	51	0.78	247		CGG*	1.92	828	1.55	631
	AAC*	1.9	937	1.22	383		AGA	0.01	3	0.64	262
Lys	AAA	0.18	97	0.75	402		AGG	0.02	9	0.61	250
	AAG*	1.82	983	1.25	676	Gly	GGU	0.05	32	0.66	326
Asp	GAU	0.12	190	0.8	489		GGC*	2.71	1910	1.37	677
	GAC*	1.88	2959	1.2	740		GGA	0.09	65	1.09	540
Glu	GAA	0.23	391	0.83	717		GGG*	1.15	808	0.88	433

∗Codons whose occurrences are significantly higher (P <0 .01) in the extreme left side of axis 1 than the genes present on the extreme right of the first major axis. AA: amino acid; N: number of codon; [1]: genes on extreme left of axis 1; [2]: genes on extreme right of axis 1.

In conclusion, high level of heterogeneity is seen within the genes of *S. ruber*. The findings reveal that there are large number of genes with high G+C content, and that the G+C content at the third codon position is higher than that of A+T. Accordingly, it is supposed that the usage frequency of codons ending with G or C bases is higher than that ending with A or T bases. In this study, the general association between codon usage bias and base composition suggests that mutational pressure, rather than natural selection, is mainly supported by the highly significant correlation between GC_{3s} and N_c. The (G+C) content is another factor which is found to play an important role in codon usage bias. The overall degree of synonymous codon usage bias is high as suggested by N_c value (mean N_c=37.85 and sd=5.73; less than 40). Apart from the two main factors, gene length also influences the codon usage while aromaticity and hydrophobicity of the encoded proteins play minor role in codon usage bias. The observation that genes expressed at high levels have increased frequencies of those codons that are expected to be translationally optimal is strongly suggestive that these codons are selectively favoured. Identification of the

codon usage patterns of halophilic bacterium may prove useful in the design of oligonucleotide probes, in deducing whether open reading frames are likely to be protein coding, determining the probable level of expression of genes, and indicating the codons to be used in synthetic genes are likely to be expressed in non-salt tolerant bacteria.

Acknowledgments

Financial assistance received by National Agricultural Innovative Project, Indian Council of Agricultural Research, New Delhi entitled 'Establishment of National Agricultural Bioinformatics Grid in ICAR' is gratefully acknowledged.

REFERENCES

Abdi H, Williams LJ (2010) Correspondence Analysis. Contrast analysis. In (Ed. Salkind N.J.), Encyclopedia of Research Design. Thousand Oaks: Sage. pp. 243-251.

Alvarez F, Robello C, Vignali M (1994) Evolution of codon usage and base contents in kinetoplastid protozoan. Mol Biol Evol 11:790-802.

Andersson GE, Sharp PM (1996) Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142:915–25.

Anton J, Oren A, Benlloch S, Valera FR, Amann R, Mora RR (2002) *Salinibacter ruber* gen. nov., sp. nov., a new species of extremely halophilic bacteria from saltern crystallizer ponds. Int J Syst Evol Microbiol 52:485–491.

Comeron JM, Aguade M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47: 268–274.

Corcelli A, Veronica MT, Lattanzio, Mascolo G, Babudri F, Oren A, Kates M (2004) Novel sulfonolipid in the extremely halophilic bacterium, *Salinibacter ruber*. Applied and Environmental Microbiology 70:6678–6685.

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 in Genetics 16:287–289.

Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640–49.

Ermolaeva MD (2001) Synonymous Codon Usage in Bacteria. Curr Issues Mol Biol 3:91-97.

Ewens WJ, Grant GR (2001) Statistical methods in Bioinformatics. Springer, Verlag Press, New York.

Eyre-Walker A (1996) Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy? Mol Biol Evol 13:864-872.

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055-7064.

Grammann K, Volke A, Kunte HJ (2002) New type of osmoregulated solute transporter identified in halophilic members of the bacteria domain: TRAP Transporter TeaABC mediates uptake of ectoine and hydroxyectoine in *Halomonas elongata* DSM 2581^{T}. J of Bacteriology 184:3078–3085.

Grantham RC, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8: 49-79.

Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic Press: London.

Gupta SK and Ghosh TC (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa.* Gene 273: 63-70.

Gupta SK, Bhattacharyya TK, Ghosh TC (2004) Synonymous codon usage in Lactococcus lactis: mutational bias versus translational selection. J Biomol Struc Dynam 21:527-536.

Gutierrez G, Marquez L, Mann A (1996) Preference for guanine at first codon position in highly expressed Escherichia coli genes. A relationship with translational efficiency. Nucleic Acids Res 24: 2525-2527.

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.

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132.

Lanyi JK (1974) Salt-dependent properties of proteins from extremely halophilic bacteria. Bacteriol Rev 38:272–290.

Liu H, He R, Zhang H, Huang Y, Tian M, Zhang J (2010) Analysis of synonymous codon usage in *Zea mays*. Mol Biol Rep 37:677–684.

Lobry JR, Gautier C (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acid Res 22:3174-80.

Lu H, Zhao WM, Zheng Y, Wang H, Qi M, Yu XP (2005) Analysis of synonymous codon usage bias in Chlamydia," Acta Biochimica et Biophysica Sinic 37:1–10

Maria D, Ermolaeva (2001) Synonymous codon usage in bacteria. Curr Issues Mol Biol 3:91–7.

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila.* J Mol Evol 45:514–523.

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

Nakamura Y, Tabata S (1997) Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes. Microbiol Comp Genomics 2:299-312.

Oren A, Mana L (2002) Amino acid composition of bulk protein and salt relationships of selected enzymes of *Salinibacter ruber*, an extremely halophilic bacterium. Extremophiles 6(3):217-23.

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of *Haemophilus influenza* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. Gene 215:405-413.

Peden JF (1999) Analysis of Codon Usage. Ph.D. Thesis, University of Nottingham.

Pieper U, Kapadia G, Mevarech M, Herzberg O (1998) Structural features of halophilicity derived from the crystal structure of dihydrofolate reductase from the Dead Sea halophilic archaeon, *Haloferax volcanii*. Structure 6:75-88.

Sahu K, Gupta SK, Ghosh TC, Sau S (2004) Synonymous Codon Usage Analysis of the Mycobacteriophage Bxz1 and its Plating Bacteria M. smegmatis: Identification of Highly and Lowly Expressed Genes of Bxz1 and the Possible Function of Its tRNA Species. J Biochem Mol Biol 37:487-492.

Sau K, Gupta SK, Sau S, Ghosh TC (2005) Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. Virus Res 113:123–131.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141–53.

Sharp PM, Cowe E (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast, 7(7): 657–678.

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon usage in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res 16:8207–8211.

Sharp PM, Devine KM (1989) Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do “prefer” optimal codons. Nucleic Acids Res. 17: 5029-5038.

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. Nucleic Acids Res 14:7737-7749.

Sharp PM, Li WH (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281- 1295.

Tiller ER, Collins RA (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50:249-257.

Wright F (1990) The ‘effective number of codons’ used in a gene. Gene 87:23-29.