Biochemistry & Analytical Biochemistry
ISSN: 2161-1009

Farooqi et al., Biochem Anal Biochem 2016, 5:1
DOI: 10.4172/2161-1009.1000257

Research Article Open Access

Genome-Wide Relative Analysis of Codon Usage Bias and Codon Context Pattern in the Bacteria *Salinibacter Ruber, Chromohalobacter Salexigens* and *Rhizobium Etli*

Mohammad Samir Farooqi[1]*, DC Mishra[1], Niyati Rai[1], DP Singh[2], Anil Rai[1], KK Chaturvedi[1], Ratna Prabha[2] and Manjeet Kaur[1]

[1]*Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi-110012, India*
[2]*National Bureau of Agriculturally Important Microorganisms, Mau, U.P. - 275101, India*

Abstract

Codon is the basic unit for biological message transmission during synthesis of proteins in an organism. Codon Usage Bias is preferential usage among synonymous codons, in an organisms. This preferential use of a synonymous codon was found not only among species but also occurs among genes within the same genome of a species. This variation of codon usage patterns are controlled by natural processes such as mutation, drift and pressure. In this study, we have used computational as well as statistical techniques for finding codon usage bias and codon context pattern of *Salinibacter ruber* (extreme halophilic), *Chromohalobacter salexigens* (moderate halophilic) and *Rhizobium etli* (non-halophilic). In addition to this, compositional variation in translated amino acid frequency, effective number of codons and optimal codons were also studied. A plot of EN_c versus GC_{3s} suggests that both mutation bias and translational selection contribute to these differences of codon bias. However, mutation bias is the driving force of the synonymous codon usage patterns in halophilic bacteria (*Salinibacter ruber* and *Chromohalobacter salexigens) and* translational selection seems to affect codon usage pattern in non-halophilic bacteria (*Rhizobium etli*). Correspondence analysis of Relative Synonymous Codon Usage revealed different clusters of genes varying in numbers in the bacteria under study. Moreover, codon context pattern was also seen variable in these bacteria. These results clearly indicate the variation in the codon usage pattern in these bacterial genomes.

Keywords: Codon Usage Bias (CUB); Halophilic bacteria; Relative Synonymous Codon Usage (RSCU); Optimal codon; Correspondence analysis; Codon context pattern

Abbreviations: CUB: Codon Usage Bias; RSCU: Relative Synonymous Codon Usage; EN_c: Effective Number of Codons; *S. ruber: Salinibacter ruber*; *C. salexigens: Chromohalobacter salexigens; R. etli: Rhizobium etli*.

Introduction

The study of organisms from extreme environments is an important field of research for enhancing knowledge in context of the molecular and biological approaches in agriculture. It helps in better and deeper understanding in multiple scientific areas for developing new varieties /breeds and biological materials. The ability of organism to survive under high salt conditions offers an excellent opportunity to increase understanding of hyper saline physiology and in identifications of genes which are responsible for salt tolerant. Halophilic organisms which thrive in saline environments such as salt lakes, coastal lagoons and man-made salterns characterized by two stress factors, the high ion concentration and low water potential [1,2]. It can be seen that extensive information on the taxonomy, physiology and ecology of halophilic microorganisms has been reported but relative codon usage patterns in these organisms have little been studied. There is a wide range of halophilic microorganisms which comprise domains of Archaea and Bacteria. The saline cytoplasm of these bacteria requires enzymes which are rich in acidic amino acids and dependent on K+ or Na+ for their biological activity [3]. Oren and Mana 2002 [4] have reported that these organisms include: (i) the extremely halophilic Archaea of the family Halobacteriaceae, which comprises *Halobacterium*, *Haloarcula*, *Haloquadratum*, *Halorhabdus*, *Natronobacterium* and *Natrononococcus* (ii) the halophilic Bacteria of the order Haloanaerobiales and (iii) the bacterium *S. ruber*. Extremely halophile, *S. ruber* is a red, aerobic bacterium, requires at least 150g of salt/liter for growth and grows optimally at NaCl concentrations between 200-300 g/litre [5]. *C. salexigens*, a gram-negative aerobic bacterium, is moderately halophilic in nature. It grows at NaCl concentrations ranging between 0.5M and 4M, with an optimum growth at 2.0-2.5M and at an optimum temperature of 37°C [6,7]. *Rhizobium etli* (*R. etli*) is gram-negative soil bacteria, which fixes nitrogen and forms an endosymbiosis nitrogen fixing association with roots of legumes. *R. etli* inoculants are useful as bio fertilizers. These inoculants promote plant growth, productivity and are internationally accepted as an alternative source of N-fertilizer [8].

The genetic code is the sequence of nucleotides in DNA or RNA that determines specific amino acid sequence in synthesis of proteins. It employs 64 codons, which can be grouped into 20 disjoint families, one family for each of the standard amino acid, and 21st family for translation/ termination signal. Different codons that encode the same amino acid are called synonymous codons and they usually differ by nucleotide at the third codon position. According to the number of synonymous codons related to each amino acid, there are two amino acids with one codon choice, nine with two, one with three, five with four and three with six. These represent five synonymous families types (SF), designated as SF types 1, 2, 3, 4 and 6 [9]. The unequal or

***Corresponding author:** Mohammad Samir Farooqi, Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi-110012, India, Tel: 9899038037; Fax: 011-25841564; Email: samir@iasri.res.in

Received: January 22, 2016; **Accepted:** March 10, 2016; **Published** March 14, 2016

Citation: Farooqi MS, Mishra DC, Rai N, Singh DP, Rai A, et al. (2016) Genome-Wide Relative Analysis of Codon Usage Bias and Codon Context Pattern in the Bacteria *Salinibacter Ruber, Chromohalobacter Salexigens* and *Rhizobium Etli*. Biochem Anal Biochem 5: 257. doi:10.4172/2161-1009.1000257

preferred usage of a particular codon by an amino acid among the SF family is termed as Synonymous Codon Usages (SCU). The specific SCU patterns may be due to mutational bias [10], bias in G+C content, natural selection etc. However, SCU pattern is non-random and species-specific [11,12] frequency of synonymous codons usage varies among species [13]. It has also been reported that there is significant variation of codon usage bias (CUB) among different genes within the same organism [14,15]. In some organisms codon bias is very strong, whereas, in others, different synonymous codons are used with similar frequencies [16-18]. Similarly, the strength of codon bias varies across genes within each genome, with some genes using a highly biased set of codons and others using the different synonymous codons with similar frequencies. The degree of codon usage bias is also shown to depend on the level of gene expression, with highly expressed genes exhibiting greater codon bias than infrequently expressed genes [14,15]. This correlation was used to predict highly expressed genes of an organism especially in case of prokaryotes. Among bacteria, genomic G+C content varies over a wide range, presumably reflecting variation in mutation biases [19] with a major impact on codon usage [20]. Analysis of codon usage pattern can provide a basis for understanding the relevant mechanism for biased usage of synonymous codons [15,21-22]. The codon context pattern may affect the translation selection of genes that is suitable for studying codon bias patterns in understanding the genetic diversity. Studies have already shown that set of preferred codons are used by each genome and that codon context is not a random event [23-25].

In this study, the genomes of bacteria *S. ruber* and *C. salexigens* and *R. etli* bacteria have been analyzed in terms of synonymous codon usage bias and codon context pattern for understanding of molecular mechanism under salinity stress and also to have a comparative analysis of codon usage in these bacteria.

Materials and Methods

Nucleotide sequence data

In our study we have included complete coding sequences (CDSs) of three bacteria *viz. S. ruber, C. Salexigens and R. etli*, i.e, extreme, moderate and non-halophilic bacteria respectively. The nucleotide sequence in FASTA format was retrieved from http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi. In order to minimize the sampling errors, gene sequences less than 300bp length and those with intermediate termination codons were removed [26]. Final dataset after exclusion of these sequences consisted of 1450, 2147 and 3703 genes of *S. ruber, C. salexigens and R. etli* respectively. Perl program has been developed for merging these gene sequences for further processing and analysis [27,28].

Calculation of codon usage indices

The frequency of codons (excluding stop codons) corresponding to each amino acid in the CDSs is used for codon usage analysis. Relative Synonymous Codon Usage (RSCU) [29], Effective Number of Codons (EN_c) [9] and Codon Adaptation Index (CAI) [30] were calculated. Highly and lowly expressed genes, and frequency of optimal codons were identified in all the three bacterial species using CodonW software (http://codonw.sourceforge.net/).

Statistical analysis

Statistical analysis was carried out using SAS 9.2. In order to derive valid biological conclusions, multivariate statistical analysis using Correspondence Analysis (CA) was applied. 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 [31,32]. The CA was also used for determining highly expressed genes and optimal codons. Pearson correlation was calculated to identify the relationship between CAI and EN_c values. In order to know the effect of base composition of third position of codons on their effective number, Poisson regression analysis was performed taking EN_c as dependent variable and A3s, T3s, C3s and G3s frequencies as independent variables.

Codon context pattern analysis

Codon context generally refers to sequential pair of codons in a gene. Codon context pattern analysis was performed using the Anaconda 2.0 software [33]. The amino acid pairs and the residual values of each codon pair were calculated from these coding sequences. Cluster tree was generated to compare the genomes through codon context pattern analysis. The cluster pattern is based on average matrix of residuals of each codon context among the species [34]. Codon context patterns reveal that the specific codons are frequently used as the 3′- and 5′-context of start and stop codons.

Results

Codon usage pattern

Over all RSCU value for the 59 codons (Table 1) provides ample evidence of codon usage bias in the studied bacterial genomes. It can be seen that codon ending with C and G nucleotides in all synonymous codon family are the most preferred as compared to A and T ending codons in these bacteria. Moreover, bias is more towards codons ending with C as compared to G, this clearly shows that genes of these bacteria are highly dominated by codons ending with C. These results indicate that the codon usage pattern in these bacterial species is mostly contributed by compositional constraints.

The list of identified optimal codons for each species is summarized in Table 2. The optimal codons are summarized in Figure 1. Venn diagram indicates that 23 (82.1%) are common codons in *S. ruber, C. salexigens* and *R. etli.* These codons are UUC, UCC, AGC, UAC, UGC, CUC, CCC, CAC, CGC, CAG, AUC, ACC, AAC, GUC, GCC, GAC, GGC, UCG, CUG, CCG, AAG, GUG and GAG. Two (7.1%) common codons are found in *S. ruber* and *R. etli* i.e., GCG and ACG. One (3%) codon i.e., AGG is found exclusively in *R. etli* and two (7.1%) codons i.e., CGG and GGG are found exclusively in *S. ruber*.

Heterogeneity of codon usage

In order to study the heterogeneity of codon usage, two different indices, namely, EN_c and GC_{3s} were used (Figure 2). *S. ruber* (green color) shows extreme GC_{3s} content from 23 to 97% (mean: 85% and standard deviation: 7.8%) and a wide range of EN_c variation from 29.17 to 61 (mean: 37.85 and standard deviation: 5.73). In *C. salexigens* (blue color), GC_{3s} values vary from 33 to 94% (mean: 81% and standard deviation: 6.1%) and their corresponding EN_c values varies from 26.32 to 61 (mean: 37.9 and standard deviation: 4.462). In *R. etli* (red color), GC_{3s} values vary from 29 to 91% (mean: 61% and standard deviation: 14.6%) and their corresponding EN_c values varies from 25.04 to 61 (mean: 46.2 and standard deviation: 7.004). The heterogeneity of means was also supported by independent sample t-test, which was performed between the means of *S. ruber* and *C. salexigens*, *S. ruber* and *R. etli* and *R. etli and C. salexigens* for GC_{3s} and EN_c (Table 3) distribution. All the means were found significant with $P < 0.05$ except the means of EN_c in case of *S. ruber* and *C. salexigens.*

AA	Codon	N			RSCU		
		S. ruber	*C. salexigens*	*R. etli*	*S. ruber*	*C. salexigens*	*R. etli*
Phe	UUU	3642	3989	10984	0.58	0.29	0.5
	UUC	9008	23346	32907	1.42	1.71	1.5
Leu	UUA	1043	545	2269	0.17	0.04	0.13
	UUG	4289	9973	13261	0.71	0.67	0.75
	CUU	4410	3265	14594	0.73	0.22	0.82
	CUC	13142	21829	31960	2.17	1.47	1.8
	CUA	2275	1280	3089	0.37	0.09	0.17
	CUG	11254	52246	41611	1.85	3.52	2.34
Ile	AUU	3295	5722	12813	0.75	0.45	0.56
	AUC	9162	31012	49742	2.08	2.46	2.17
	AUA	736	1032	6323	0.17	0.08	0.28
Met	AUG	6241	20118	32074	1	1	1
Val	GUU	3179	2311	9983	0.46	0.16	0.56
	GUC	10211	25639	36679	1.47	1.73	2.06
	GUA	2257	2881	3046	0.32	0.19	0.17
	GUG	12217	28379	21344	1.75	1.92	1.2
Tyr	UAU	1047	7301	11785	0.29	0.77	1.14
	UAC	6278	11775	8930	1.71	1.23	0.86
TER	UAA	602	321	14842	0.26	0.45	2.56
	UAG	742	280	1418	0.32	0.39	0.24
	UGA	5660	1546	1148	2.42	2.16	0.2
His	CAU	4401	8634	12006	0.64	0.87	1.07
	CAC	9333	11311	10333	1.36	1.13	0.93
Gln	CAA	4209	5091	9882	0.61	0.34	0.56
	CAG	9563	25158	25443	1.39	1.66	1.44
Asn	AAU	2052	5684	13284	0.48	0.57	0.83
	AAC	6450	14280	18639	1.52	1.43	1.17
Lys	AAA	3090	2679	11866	0.61	0.26	0.57
	AAG	7039	17678	29793	1.39	1.74	1.43
Asp	GAU	5528	17438	21366	0.44	0.71	0.88
	GAC	19546	31573	27399	1.56	1.29	1.12
Glu	GAA	6763	18269	23833	0.55	0.73	0.99
	GAG	17774	31987	24153	1.45	1.27	1.01
Ser	UCU	6012	1167	13183	0.63	0.16	0.58
	UCC	11095	9092	22161	1.17	1.28	0.98
	UCA	7289	1406	18449	0.77	0.2	0.82
	UCG	18538	13627	52657	1.95	1.92	2.33
	AGU	3799	3163	4704	0.4	0.45	0.21
	AGC	10404	14019	24557	1.09	1.98	1.09
Pro	CCU	8681	2377	15045	0.65	0.25	1.64
	CCC	15010	15410	16164	1.12	1.59	0.48
	CCA	8183	1823	13412	0.61	0.19	1.5
	CCG	21749	19154	50131	1.62	1.98	0.64
Thr	ACU	4170	1681	5848	0.37	0.16	1.38
	ACC	14511	23742	27398	1.29	2.26	0.64
	ACA	6075	1921	11668	0.54	0.18	0.68
	ACG	20205	14669	31232	1.8	1.4	0.57
Ala	GCU	7523	3899	19082	0.5	0.17	2.12
	GCC	22967	49504	59219	1.53	2.16	0.31
	GCA	9063	7171	25169	0.6	0.31	1.44
	GCG	20484	31220	54361	1.36	1.36	0.61
Cys	UGU	3658	1372	5441	0.54	0.38	0.39
	UGC	9831	5936	22350	1.46	1.62	1.61
Trp	UGG	12693	11103	22974	1	1	1
Arg	CGU	9200	11852	11434	0.66	1.2	0.46
	CGC	19369	33421	49439	1.38	3.4	2
	CGA	14869	2934	21764	1.06	0.3	0.88
	CGG	22871	9394	32600	1.63	0.95	1.32
	AGA	6479	471	12482	0.46	0.05	0.51
	AGG	11392	953	20368	0.81	0.1	0.83
Gly	GGU	5912	8201	12221	0.45	0.5	0.53
	GGC	20183	40390	56191	1.52	2.48	2.42
	GGA	10112	4506	10922	0.76	0.28	0.47
	GGG	16853	12063	13485	1.27	0.74	0.58

*AA represents amino acid, N is the number of codons and RSCU represents relative synonymous codon usage.

Table 1: Overall codon usage data of the genes in *S. ruber*, *C. salexigens* and *R. etli*.

Type	Bacteria	Optimal codons	
Extreme halophile	*S. ruber*	**17 C-ending:** UUC (Phe), UCC (Ser), AGC (Ser), UAC (Tyr), UGC (Cys), CUC (Leu), CCC (Pro), CAC (His), CGC (Arg),**CAG (Gln)**, AUC (Ile), ACC (Thr), AAC (Asn), GUC (Val), GCC (Ala), GAC (Asp), GGC (Gly)	**10 G-ending:** UCG (Ser), CUG (Leu), CCG (Pro), **CGG (Arg), GGG (Gly), GCG (Ala), ACG (Thr),** AAG (Lys), GUG (Val), GAG (Glu)
Moderate halophile	*C. salexigens*	**16 C-ending:** UUC (Phe), UCC (Ser), AGC (Ser), CUC (Leu), UAC (Tyr), UGC (Cys), CCC (Pro), CAC (His), CGC (Arg), AUC (Ile), ACC (Thr), AAC (Asn), GUC (Val), GCC (Ala), GAC (Asp), GGC (Gly)	**7 G-ending:** UCG (Ser), CUG (Leu), CCG (Pro), CAG (Gln), AAG (Lys), GUG (Val), GAG (Glu)
Non-halophile	*R. etli*	**16 C-ending:** UUC (Phe), UCC (Ser), AGC (Ser), UAC (Tyr), UGC (Cys), CUC (Leu), CCC (Pro), CAC (His), CGC (Arg), AUC (Ile), ACC (Thr), AAC (Asn), GUC (Val), GCC (Ala), GAC (Asp), GGC (Gly)	**10 G-ending:** UCG (Ser), CUG (Leu), CCG (Pro), CAG (Gln), **ACG (Thr), AGG (Arg),GCG (Ala),**AAG (Lys), GUG (Val),GAG (Glu)

Table 2: Identified optimal codons for each species.

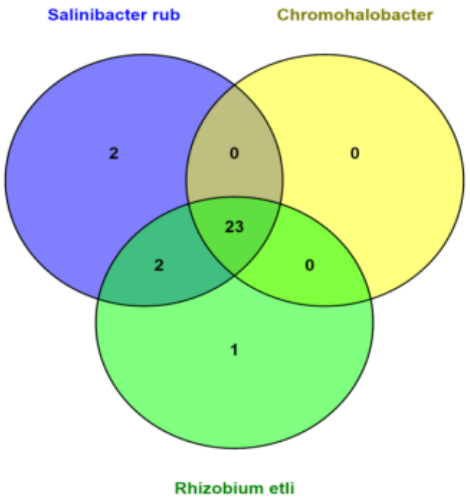

Figure 1: Venn diagram illustrating optimal codon comparison between all three species *S. ruber* (Blue circle) *C. salexigens* (Yellow circle) and *R. etli* (Green circle). 23 common elements are included in *"Salinibacter rub", "Chromohalobacter" and "Rhizobium etli"*, 2 common elements are included in "*Salinibacter rub*" and "*Rhizobium etli*", 1 element included exclusively in "*Rhizobium etli*" and 2 elements included exclusively in "*Salinibacter rub*".

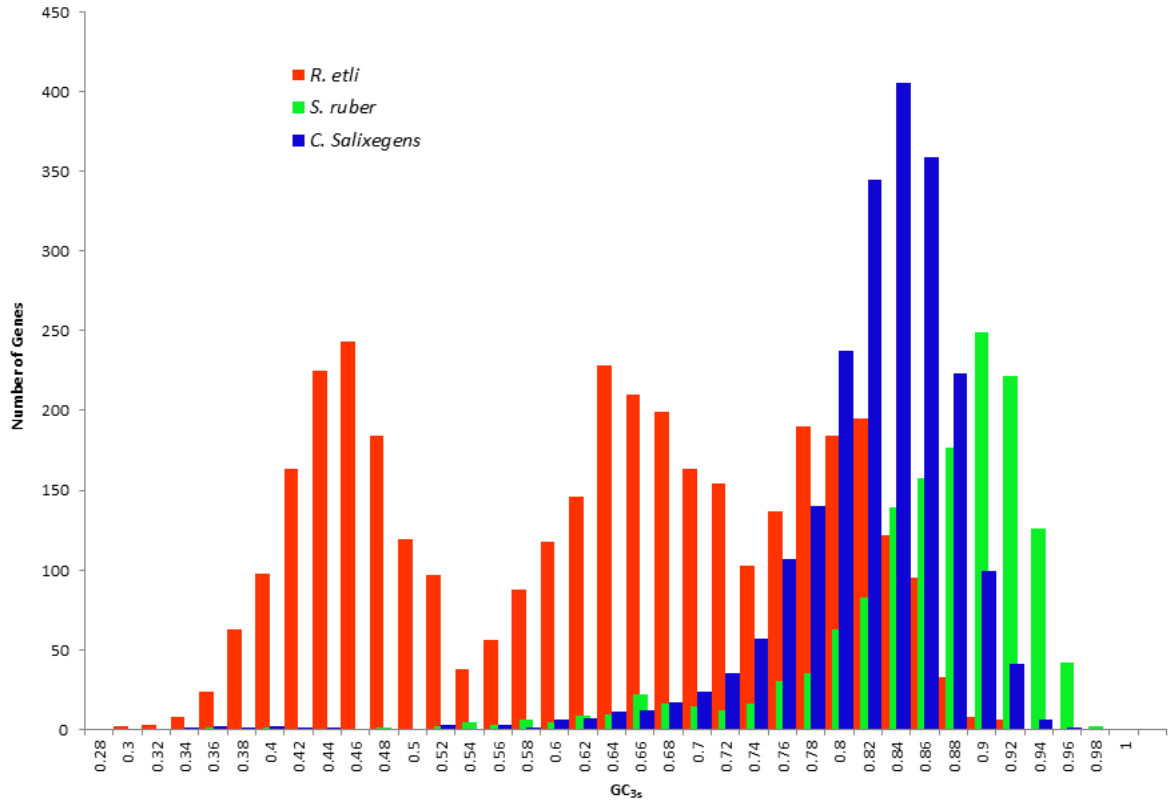

Figure 2: The GC_{3s} distribution of different genes in *S. ruber, C. salexigens and R. etli* genes.

Between Organism	t value for GC_3	t value for EN_C
R. etli and C. salexigens	70.03166	-55.0366
S. ruber and R. etli	-73.3905	44.09617
S. ruber and C. salexigens	16.65059	-0.53152

Table 3: Independent sample t test for GC_3 and EN_C.

It can be clearly seen from the Figure 2 that *R. etli* is tri-modal, whereas other two are uni-modal. This indicates that high variation in codon usage occurs in these bacteria. This large range of variation in codon usage is probably due to differential mutational pressure acting on different coding regions of a genome.

Relationship between EN_c and GC_{3s}

Based on the codon homozygosity, EN_c is the most useful concept reflecting codon usage bias pattern in different organisms. EN_c values were calculated for coding sequences of these three bacterial species. In order to shape codon usage bias, EN_c and their corresponding GC_{3s} values are required to demonstrate the role of dominant factors in bacteria. EN_c and GC_{3s} plot was made in absence of selection pressure (Figure 3). Generally, if GC-composition bias is not responsible for any codon usage bias, all genes must lie on normal curve but this actually doesn't occur in this study.

The EN_c plot in Figure 3 for *S. ruber* (in green colour) and *C. salexigens* (in blue colour) shows maximum negative correlation, whereas, in *R. etli* (red in colour), two distinguished groups are observed, where, one has almost zero and other with negative correlation between EN_c and GC_{3s}. Thus, strong influence of compositional constraints on codon usages bias could be stated from the presence of significant negative correlation between GC_{3s} and EN_c in case of *S. ruber* and *C. salexigens* as compared to *R. etli*.

Mutational bias effect on codon usage variation

In order to determine differences between nucleotide composition and codon selection in each species, Pearson correlation between EN_c and CAI were obtained. Significant negative correlation was observed in *S. ruber* ($r = -0.43711$, $P < 0.0001$), *C. salexigens* ($r = -0.57703$, $P < 0.0001$) and *R. etli* ($r = -0.72062$, $P < 0.0001$). The correlation values indicates that codon usage bias of genes of these species have very distinct relationships with nucleotide composition of coding sequences.

Correlation analysis between EN_c and GC_{3s} showed lower correlation coefficient for *R. etli* ($r = -0.67$, $P < 0.01$) than *C. salexigens* ($r = -0.81$, $P < 0.01$) and *S. ruber* ($r = -0.94$, $P < 0.01$). This shows expression of genes in *R. etli* is more dependent on composition biased mutational pressure than *S. ruber* and *C. salexigens*.

In order to further test, the compositional constraints, Poisson regression was performed between EN_c as dependent variable and nucleotide composition at the third codon position as predictor variables. The estimated coefficients are shown in Table 4. All the coefficients are found to be significant in *S. ruber*. The coefficient of T3s is non-significant in *C. salexigens* and T3s and G3s are non-significant

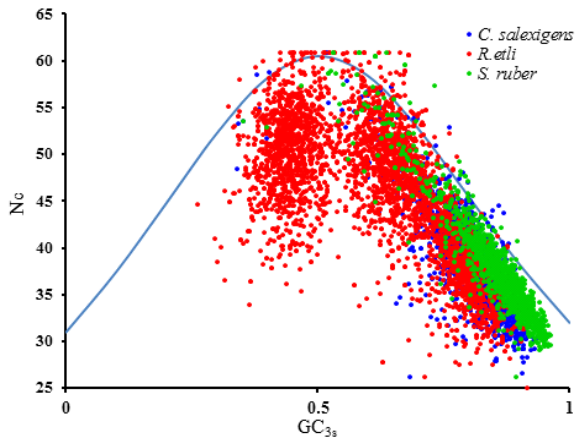

Figure 3: Nc plot (EN_c vs. GC_{3s}). The continuous curve represents the expected curve between GC_{3s} and EN_c under random codon usage.

Species	T3 (coefficient)	A3 (coefficient)	C3 (coefficient)	G3 (coefficient)
S. ruber	0.7058*	1.0241*	-0.5335*	-0.2828*
C. salexigens	0.2854	0.6669*	-0.8904*	-0.4468*
R. etli	-0.0477	0.3448*	-0.6824*	-0.1007
* significant value with P < 0.01				

Table 4: Poisson regression coefficients of effective number of codons of genes within each genome as a function of base compositions at the 3rd position of codons.

in *R. etli*. It is concluded that nucleotide composition is playing a major role in determining the EN_c variation.

Multivariate statistical approach

The dataset of RSCU values of genes of these bacteria was subjected to correspondence analysis (CA), a method of multivariate statistical analysis (MVA). 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; axis 2 describes the second largest trend, and so on with each subsequent axis describing a progressively smaller amount of variation as shown in Table 5. 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 [35]. The low value might be due to the extreme genomic composition [36] 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 bias of these genes are not homogeneous. It is interesting to note that the scatter plot (Figure 4) drawn between axis 1 and axis 2, scores for *R. etli* genes are clearly differentiated into three clusters, whereas in case of *S. ruber* and *C. salexigens* single cluster is observed. Genes falling in same cluster indicate that these genes have more or less similar codon usage bias.

Translational optimal codons

In order to identify the optimal codons, 10% of genes each from both extremes of axis 1 were analysed for these species under study (Table 5). Ikemura 1981 [37] 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 [38,39]. This trend has been interpreted as the co-adaptation between amino acid composition of protein and tRNA-pools to enhance the translational efficiency. The possible reasons for the varying GC bias in bacteria under saline habitats [40], although not very strict, could be linked with tRNA affinity as deciphered in this study, selection on genomic base composition [41] and presence of highly acidic proteome in halophiles mostly lacking basic proteins and over representation of acidic residues (e.g. Asp and Glu) in amino acids [42].

Codon context analysis

Data clustering helps to know the identification patterns of preferred and rejected codon pairs which can give a better understanding of genetic diversity. The codon context maps along with cluster trees were generated. The 5' codons are in rows and the 3' codons are in columns in 64 x 64 contingency (Table 6). The green color represents highest number of the context i.e., positive values and red color represents the lowest number of context i.e., negative values. It has been observed that the highest and lowest number of codon context is comparatively lower in *S. ruber* as compared to that of *C. salexigens* and *R. etli*. (Figures 5a, 5b and 5c). Hierarchical clustering of codon context data based on single linkage highlights discrete groups of good and bad codon context. It can be seen from the Figures 5a, 5b and 5c that major numbers of codons do not fall into any cluster which indicates preferences or rejections of codons are defined on one to one basis. Further, species specific codon context maps indicate that each species has specific set of codon context rules and there is no clear distinguishable common features present among these species.

Distribution of the adjusted residuals from the codon context map of *S. ruber*, *C. salexigens* and *R. etli*. (Figures 6a, 6b and 6c) show that 57.75, 49.78 and 52.28 percent of the residuals respectively fall within the non-significant -5 to +5 interval, indicating that a very large number of codon combinations are not significant to the rejection of

	S. ruber	***C. salexigens***	***R. etli***
Axis 1	12.61%	14.52%	22.71%
Axis 2	6.79%	6.79%	13.46%

Table 5: Percentages of prominent axes contributing to the codon usage variation among the genes.

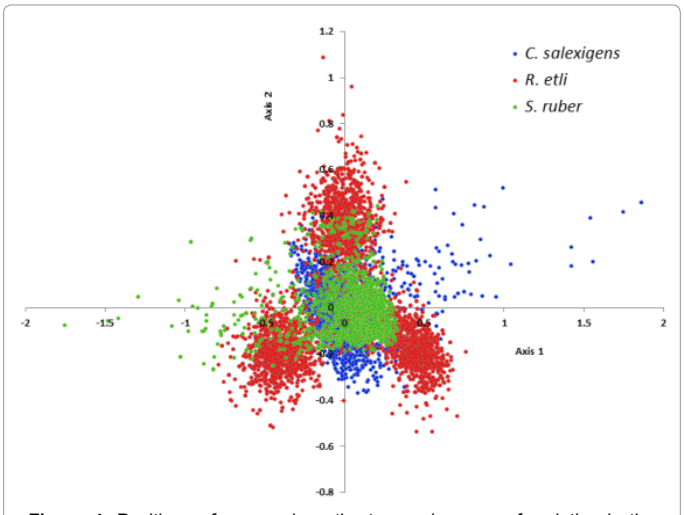

Figure 4: Positions of genes along the two major axes of variation in the correspondence analysis on RSCU values.

AA	Codon	RSCU[1]			N[1]			RSCU[2]			N[2]		
		S. ruber	*C. salexigens*	*R. etli*	*S. ruber*	*C. salexigens*	*R. etli*	*S. ruber*	*C. salexigens*	*R. etli*	*S. ruber*	*C. salexigens*	*R. etli*
Phe	UUU	0.36	0.1	0.2	214	74	253	0.91	0.76	1.15	333	492	530
	UUC*	1.64	1.9	1.8	975	1456	2268	1.09	1.24	0.85	395	805	393
Leu	UUA	0	0	0.01	0	2	12	0.21	0.3	0.45	68	205	279
	UUG	0.07	0.23	0.18	40	152	168	0.77	1.11	0.16	245	753	100
	CUU	0.23	0.13	0.46	131	85	433	1.2	0.72	1.82	382	486	1122
	CUC*	3.02	1.85	2.37	1719	1232	2245	1.79	1.19	1.09	572	804	675
	CUA	0.05	0.03	0.03	29	18	27	0.49	0.37	1.59	158	250	980
	CUG*	2.63	3.76	2.96	1496	2499	2809	1.54	2.31	0.89	491	1561	550
Ile	AUU	0.47	0.28	0.28	269	201	367	0.98	0.83	1.39	315	537	270
	AUC*	2.53	2.72	2.69	1439	1982	3529	1.66	1.77	0.69	531	1143	133
	AUA	0	0.01	0.03	1	7	37	0.36	0.4	0.92	116	261	178
Met	AUG	1	1	1	838	1157	1635	1	1	1	350	890	127
Val	GUU	0.09	0.06	0.35	63	51	412	0.72	0.69	1.94	274	429	597
	GUC*	1.58	2.08	2.47	1127	1652	2924	1.34	1.41	1.23	509	875	377
	GUA	0.06	0.11	0.05	41	85	64	0.64	0.5	0.43	245	313	133
	GUG*	2.27	1.74	1.13	1621	1382	1338	1.29	1.4	0.4	492	872	122
Tyr	UAU	0.07	0.47	1.12	35	264	949	0.74	1.16	1.35	243	557	414
	UAC*	1.93	1.53	0.88	953	859	752	1.26	0.84	0.65	414	401	198
TER	UAA	0.58	0.76	0.81	14	27	50	0.67	0.67	0.59	39	24	128
	UAG	1.5	0.22	0.34	36	8	21	0.79	0.45	0.17	46	16	37
	UGA	0.92	2.02	1	22	72	114	1.53	1.88	2.24	89	67	485
His	CAU	0.05	0.57	0.96	23	276	486	0.85	1.09	1.34	270	431	1835
	CAC*	1.95	1.43	1.04	830	684	531	1.15	0.91	0.66	364	363	908
Gln	CAA	0.08	0.17	0.13	56	137	126	0.72	0.72	1.61	326	490	1923
	CAG*	1.92	1.83	1.87	1313	1446	1856	1.28	1.28	0.39	581	880	467
Asn	AAU	0.1	0.29	0.63	51	194	678	0.78	0.93	1.2	247	482	235
	AAC*	1.9	1.71	1.37	937	1137	1487	1.22	1.07	0.8	383	558	158
Lys	AAA	0.18	0.13	0.22	97	103	340	0.75	0.79	1.6	402	423	264
	AAG*	1.82	1.87	1.78	983	1494	2724	1.25	1.21	0.4	676	644	65
Asp	GAU	0.12	0.42	0.67	190	603	1168	0.8	1.06	1.26	489	1023	171
	GAC*	1.88	1.58	1.33	2959	2294	2328	1.2	0.94	0.74	740	908	101
Glu	GAA	0.23	0.73	0.96	391	102	1714	0.83	1	1.63	717	1023	114
	GAG*	1.77	1.27	1.04	3017	389	1853	1.17	1	0.37	1020	1032	
Ser	UCU	0.07	0.04	0.1	20	14	53	0.86	0.7	2.3	250	286	865
	UCC*	1.75	1.82	1.57	518	643	859	1.13	0.95	1.15	326	389	433
	UCA	0.03	0.07	0.12	9	26	64	0.71	0.65	0.93	206	266	350
	UCG*	2.11	1.69	2.68	624	595	1468	1.23	1.29	0.51	356	525	191
	AGU	0.16	0.17	0.05	46	61	28	0.76	0.85	0.41	219	347	153
	AGC*	1.89	2.2	1.49	558	775	819	1.31	1.56	0.7	380	635	265
Pro	CCU	0.03	0.07	0.21	14	32	149	0.98	0.83	1.54	283	317	1356
	CCC*	1.71	1.84	1.02	714	875	739	0.78	1.19	0.47	224	454	410
	CCA	0.04	0.05	0.11	16	22	79	0.88	0.65	0.93	253	247	821
	CCG*	2.22	2.04	2.67	926	969	1932	1.36	1.33	1.06	393	508	928
Thr	ACU	0.02	0.07	0.08	10	37	70	0.67	0.65	2.06	209	295	331
	ACC*	1.95	2.95	2.37	1101	1659	2091	1.28	1.48	0.4	400	669	64
	ACA	0.05	0.06	0.12	27	32	110	0.7	0.58	0.8	220	260	129
	ACG*	1.98	0.93	1.42	1117	522	1253	1.35	1.29	0.74	423	581	119
Ala	GCU	0.03	0.11	0.24	30	123	410	0.93	0.65	1.71	417	575	2191
	GCC*	2.48	2.7	2.34	2366	2982	4015	1.23	1.61	0.76	551	1418	978
	GCA	0.07	0.18	0.24	70	202	410	0.79	0.66	0.75	351	580	954
	GCG*	1.41	1.01	1.18	1346	1115	2019	1.05	1.08	0.78	467	955	997
Cys	UGU	0.13	0.11	0.12	17	19	27	0.79	0.82	0.66	125	156	467
	UGC*	1.87	1.89	1.88	251	338	415	1.21	1.18	1.34	190	225	955
Trp	UGG	1	1	1	354	507	833	1	1	1	404	429	329
Arg	CGU	0.17	0.89	0.53	73	385	288	0.8	1.45	1.04	327	567	1661
	CGC*	3.74	4.34	4.38	1611	1872	2379	1.2	2.04	1.22	487	798	1943
	CGA	0.14	0.1	0.08	60	45	45	1.19	0.7	1.99	486	272	3173
	CGG*	1.92	0.63	0.72	828	270	392	1.55	0.98	1.46	631	384	2320
	AGA	0.01	0	0.05	3	2	29	0.64	0.42	0.14	262	164	216

	AGG	0.02	0.03	0.23	9	12	125	0.61	0.42	0.15	250	163	233
Gly	GGU	0.05	0.37	0.4	32	328	524	0.66	0.92	0.97	326	654	1138
	GGC*	2.71	3.08	3.3	1910	2740	359	1.37	1.65	1.35	677	1178	184
	GGA	0.09	0.11	0.09	65	102	118	1.09	0.62	1.1	540	442	1290
	GGG*	1.15	0.44	0.21	808	389	275	0.88	0.81	0.59	433	574	694

*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.

Table 6: RSCU for the highly and lowly expressed genes highlighting translational optimal codons in the three bacteria.

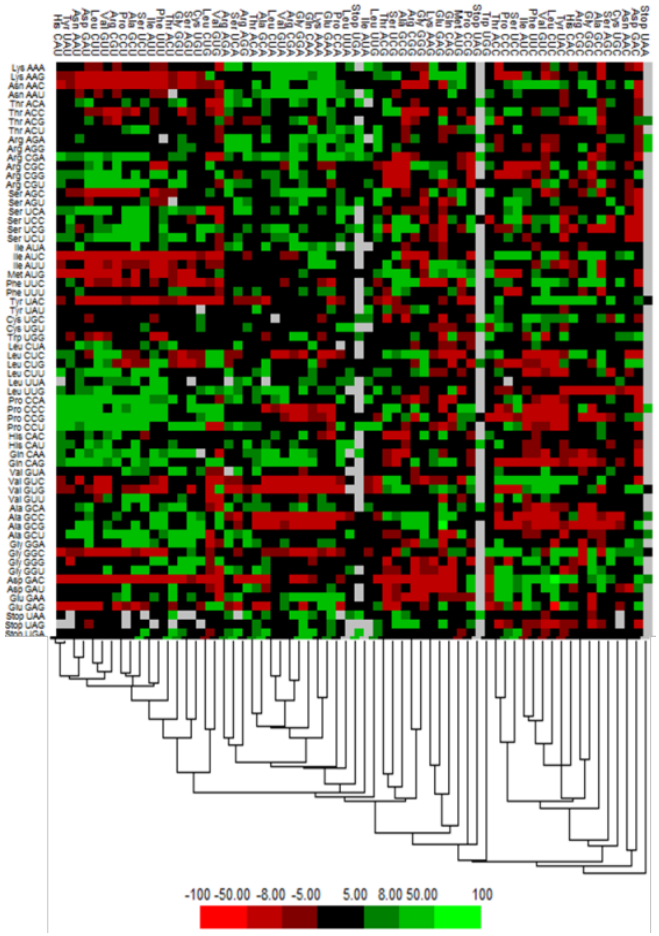

Figure 5a: The cell corresponding to each pair of codons was given by a colour scale in which red stands for rejected and green stands for preferred codon pairs. ORFeome codon context map of *S. ruber* is obtained using a colour coded map of 64 x 64 matrix.

independence. This is in accordance with above clustering result of the codon context.

Occurrence of codon context frequencies in each bacterial species were analysed and found to be variable. The frequent and rare codon contexts of each species are listed in Table 7. In *S. ruber,* GAC-GAG was most abundant (2848) and GUA-AUC was lowest (37) whereas in *C. salexigens,* codons CUG-GCC (3611) was most frequently presented and AUC-GUA with least frequency of 50. However, in case of *R. etli,* GCC-GGC was most frequently presented (7559) and UAC-CGU was the rarest with occurrence of (122).

Conclusion

Codon usage bias is the parameter that delineates the differences in the occurrence of synonymous codons in genomic coding sequences. This codon bias is calculated for all coding sequences of the three bacteria. On analysing codon usage bias of these bacteria, it has been observed that *S. ruber* and *C. salexigens* follow almost similar pattern in codon usage bias and *R. etli* varying in a noticeably different manner. The pattern of codon usage bias within *S. ruber* and *C. salexigens* is remarkably similar. Although, these bacteria show a similarity in the overall codon bias pattern, but some prominent differences are also seen. These differences in the codon bias pattern of all the bacteria are due to mutation and genetic drift as well as translation selection acting on coding sequences. Selection favours the preferred codons over the non-preferred ones. Nevertheless the existences of non-preferred or non-optimal codons are due to the action of mutational and genetic drift forces.

In this study, it was found that most frequent codons end with 'G or C' mostly at 3rd codon position with greater preference of 'C' in all the three species. This finding may be the result of compositional

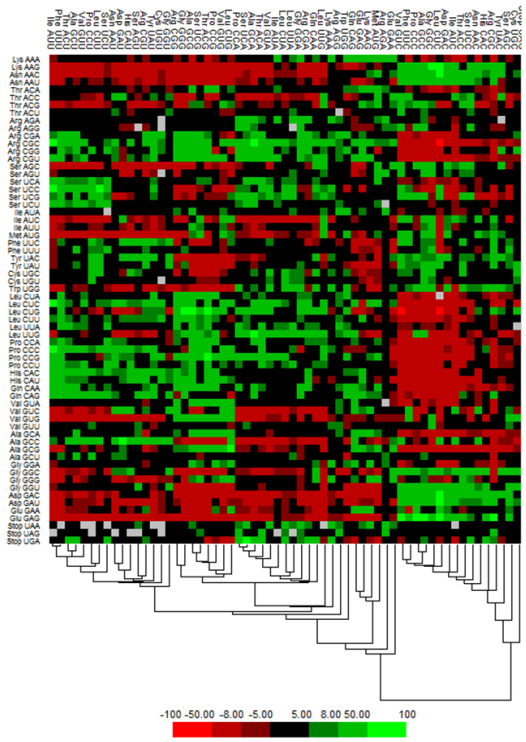

Figure 5b: The cell corresponding to each pair of codons was given by a colour scale in which red stands for rejected and green stands for preferred codon pairs. ORFeome codon context map of *C. salexigens* is obtained using a colour coded map of 64 x 64 matrix.

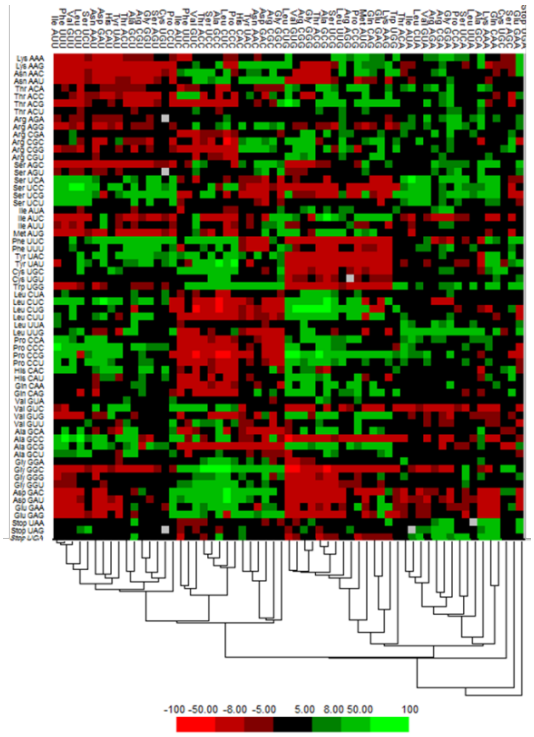

Figure 5c: The cell corresponding to each pair of codons was given by a colour scale in which red stands for rejected and green stands for preferred codon pairs. ORFeome codon context map of *R. etli* is obtained using a colour coded map of 64 x 64 matrix.

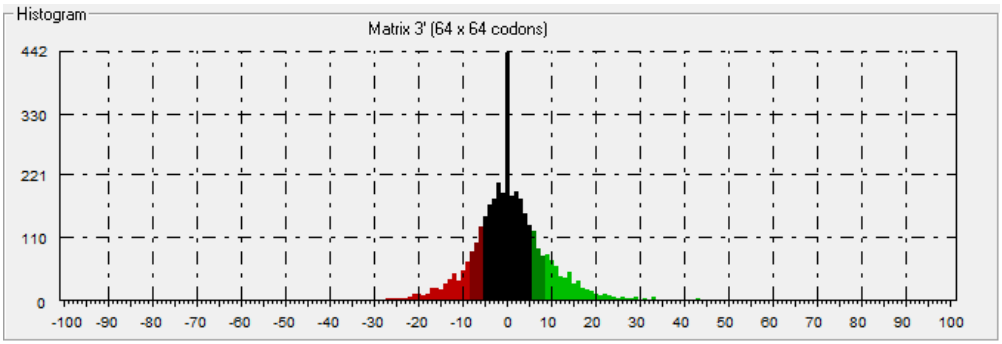

Figure 6a: Histogram graph showing distribution of negative and positive residues for codon context in the 64 x 64 codon matrix for *S. ruber*.

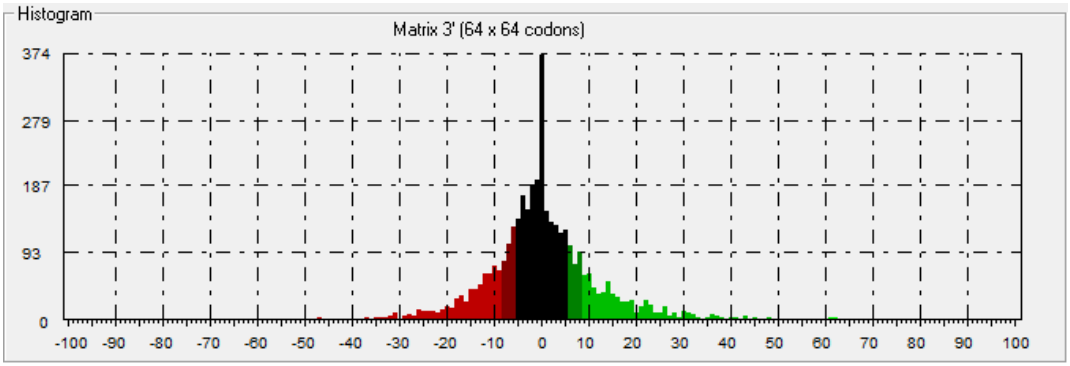

Figure 6b: Histogram graph showing distribution of negative and positive residues for codon context in the 64 x 64 codon matrix for *C. salexigens*.

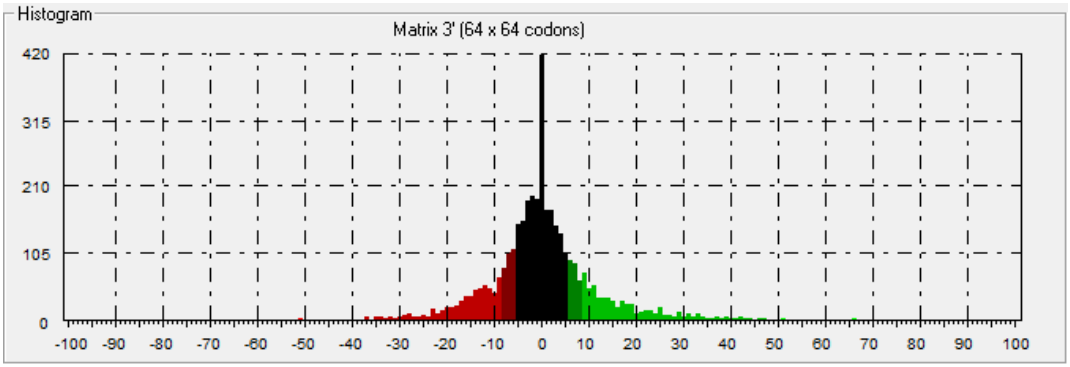

Figure 6c: Histogram graph showing distribution of negative and positive residues for codon context in the 64 x 64 codon matrix for *R. etli.*

Species	Frequent context	Rare context
S. ruber	GAC-GAG (2848)	GUA-AUC (37)
C. salexigens	CUG-GCC (3611)	AUC-GUA (50)
R. etli	GCC-GGC (7559)	UAC-CGU (122)

Table 7: Codon frequency.

constraint that occurred in codon usage pattern in these bacteria. This comparative study will be useful for understanding the pattern of codon usage in these bacterial species.

References

1. 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.
2. 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 2581T. J Bacteriology 184: 3078-3085.
3. Empadinhas N, Albuquerque L, Mendes V, Macedo-Ribeiro S, da Costa MS (2008) Identification of the mycobacterial glucosyl-3-phosphoglycerate synthase. FEMS Microbiol Lett 280: 195-202.
4. 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: 217-223.
5. Corcelli A, Veronica MT, Lattanzio, Mascolo G, Babudri F, et al. (2004) Novel sulfonolipid in the extremely halophilic bacterium, *Salinibacter ruber*. Appl Environ Microbiol 70: 6678-6685.
6. O'Connor K, Csonka LN (2003) The high salt requirement of the moderate halophile *Chromohalobacter salexigens* DSM3043 can be met not only by NaCl but by other ions. Appl Environ Microbiol 69: 6334-6336.
7. Vargas C, Argandona M, Bueno MR, Moya JR, Aunion CF, et al. (2008) Unravelling the adaptation responses to osmotic and temperature stress in *Chromohalobacter salexigens,* a bacterium with broad salinity tolerance. Saline Systems: 4-14.
8. Andrews M, James EK, Cummings SP, Zavalin AA, Vinogradova LV, et al. (2003) Use of nitrogen fixing bacteria inoculants as a substitute for nitrogen fertilizer for dry land graminaceous crops: Progress made mechanisms of action and future potential. Symbiosis 35: 209-229.
9. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87: 23-29.
10. Chen SL, Lee W, Hottes AK, McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA 101: 3480-3485.
11. Gupta SK, Bhattacharyya TK, Ghosh TC (2002) Compositional correlation and codon usage studies in *Buchnera aphidicola*. Indian J Biochem Biophys 39: 35-48.
12. Gupta SK, Bhattacharyya TK, Ghosh TC (2004) Synonymous codon usage in Lactococcus lactis: mutational bias versus translational selection. J Biomol Struct Dyn 21: 527-536.
13. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9: r43-74.
14. Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10: 7055-7074.
15. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2: 13-34.
16. Andersson GE, Sharp PM (1996) Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142: 915-925.
17. Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12: 640-649.
18. 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: 1141-1153.
19. Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 15: 8023-8040.
20. Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84: 166-169.
21. Lu J, Wu CI (2005) Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci USA 102: 4063-4067.
22. Ermolaeva MD (2001) Synonymous codon usage in bacteria. Curr Issues Mol Biol 3: 91-97.
23. Comeron JM, Aguadé M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47: 268-274.
24. Boycheva S, Chkodrov G, Ivanov I (2003) Codon pairs in the genome of *Escherichia coli.* Bioinformatics 19(8):987-98.
25. Berg OG, Silva PJ (1997) Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. Nucleic Acids Res 25: 1397-1404.
26. Sanjukta R, Farooqi MS, Sharma N, Rai A, Mishra DC, et al. (2012) Trends in the codon usage patterns of *Chromohalobacter salexigens* genes. Bioinformation 8: 1087-1095.
27. Sanjukta RK, Farooqi MS, Sharma N, Rai N, Mishra DC, et al. (2013) Statistical analysis of codon usage in extremely halophilic bacterium, *Salinibacter ruber* DSM13855. Online Journal of Bioinformatics 14: 14-31.
28. Sanjukta RK, Farooqi MS, Rai N, Rai A, Sharma N, et al. (2013) Expression analysis of genes responsible for amino acid biosynthesis in halophilic bacterium *Salinibacter ruber*. Indian J Biochem Biophys 50: 177-185.
29. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24: 28-38.
30. 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.

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

32. Farooqi MS, Sanjukta RK, Sharma N, Rai A, Mishra DC, et al. (2013) Statistical and computational methods for detection of synonymous codon usage patterns and gene expression. Int J Agricult Stat Sci 9: 303-310.

33. Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, et al (2007) Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. PLoS One 2: e847.

34. Pinheiro M, Afreixo V, Moura G, Freitas A, Santos MA, et al. (2006) Statistical, computational and visualization methodologies to unveil gene primary structure features. Methods Inf Med 45: 163-168.

35. 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.

36. Butt AM, Nasrullah I, Tong Y3 (2014) Genome wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS One 9: e90905.

37. 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.

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

39. 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.

40. Paul S, Bag SK, Das S, Harvill ET, Dutta C (2008) Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol 9: R70.

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

42. Kennedy SP, Ng WV, Salzberg SL, Hood L, Sarma SD (2001) Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res 11: 1641-1650.