Indian Journal of Biochemistry & Biophysics
Vol. 50, June 2013, pp. 177-185

Expression analysis of genes responsible for amino acid biosynthesis in halophilic bacterium *Salinibacter ruber*

R K Sanjukta[1], Md. Samir Farooqi[1]*, Niyati Rai[1], Anil Rai[1], Naveen Sharma[1], Dwijesh C Mishra[1] and Dhananjaya P Singh[2]

[1]Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute Pusa, New Delhi 110012, India
[2]National Bureau of Agriculturally Important Microorganisms, Mau Nath Bhanjan 275101, UP, India

Received 06 September 2012; revised 24 March 2013

The degeneracy of the genetic code allows for multiple codons to encode the same amino acid. However, alternative codons and amino acids are used unevenly among genes, a phenomenon termed codon-usage bias. Genes regulating amino acid biosynthesis of *Salinibacter ruber*, an extremely halophilic bacterium were studied in order to determine the synonymous codon usage patterns. Factors responsible for codon usage variation among the genes were investigated using codon usage indices and multi-variate statistical approach. Overall codon usage data analysis indicated that codons ending in G and/or C were predominant among the genes. Multi-variate statistical analysis showed that there was a single major trend in the codon usage variation among the genes, which had a strong positive correlation ($r = 0.93$, $P<0.01$) with (G + C) content of the genes. Further, correlation analysis indicated that genes with higher expression level and showing a greater degree of codon usage bias were GC-rich and preferred codons with C or G nucleotides at the third position. A set of thirteen codons were identified through Chi-square test as optimal codons, which were preferred in highly expressed genes. It could be concluded that mutational bias had a profound effect on codon usage pattern. In addition, translational selections also operated with a proper balance, making the genes translationally more efficient. The frequency of these codons appeared to be correlated with the level of gene expression and might be a useful indicator in the case of genes (or open-reading-frames) whose expression levels are unknown.

Keywords: Synonymous codon usage, Multi-variate statistical analysis, Mutational bias, *Salinibacter rubber*, Halophilic bacterium

Extreme salt habitats are dynamic and unique environments to offer excellent opportunity to increase understanding of hypersaline cellular physiology and genes responsible for salt tolerance. Halophilic organisms living in saline environments are usually challenged by two stress factors *viz*., high inorganic ion concentration and low water potential. Halophiles mainly fall in three categories depending upon their salinity optima for survival, growth and development[1]: (i) halotolerant (1-6% salt), (ii) moderate (6-15% salt), and (iii) extreme (15-30% salt concentration). *Salinibacter ruber*, an extremely halophilic rod-shaped gram-negative bacterium isolated from saltern crystallizers in Spain[2] that make-up about 5-25% of the total prokaryotic community in salt-saturated ponds[3] can grow optimally at 200-300 g/l salt concentration and high pH conditions[4]. In contrast to halophilic and halotolerant aerobic bacteria, *S. ruber* contains high concentrations of KCl[5] with acidic proteome having an isoelectric point of 5.2[6], low content of hydrophobic amino acids and high abundance of serine[7]. At genomic level, the organism shares similarity with extremely halophilic Archaea, suggesting an extensive gene exchange between *S. ruber* and *Halobacteriaceae*[4].

Most organisms synthesize amino acids, which play a central role in cellular metabolism for making of proteins to carry out many biologically important functions. These proteins provide the energy and essential nutrients for organisms. Many amino acids play important role as chemical messengers in communication between cells. The relative uses of

*Author for correspondence
E-mail: samir@iasri.res.in
Fax: 011-25841564
Ph: 91-11-25847121-2/Ext: 4305

Abbreviations: CA, correspondence analysis; CAI, codon adaptation index; CDS, coding sequences; CV, coefficient of variance; MVA, multi-variate statistical analysis; RSCU: relative synonymous codon usage; SE, standard error.

amino acid biosynthetic pathways vary widely among species because different synthesis pathways have involved fulfilling unique metabolic needs in different organisms. Although some pathways are present in certain organisms, they are absent in others. Every protein has a biosynthetic cost to the cell based on the synthesis of its constituent amino acids[8].

Energy availability reasonably limits prokaryotic growth and natural selections favour substitutions, resulting in the utilization of energetically low-cost amino acids because reduced biosynthetic costs may confer a fitness advantage to the organisms. Manifestation of such substitution biases is preferred in highly expressed genes in the same manner, as adherence to codon-usage biases tend to be greatest in genes that are expressed at high levels[9-15]. Akashi and Gojobori[16] demonstrated that genes that adhere to codon-usage biases most strongly (and are most highly expressed) tend to incorporate low-cost amino acids in *E. coli* and *Bacillus subtilis*[17].

An important linkage between the metabolism of a cell and the evolution of its genome sequence has been established[8]. Codon-usage biases have inversely been correlated with a coding region's average amino acid biosynthetic cost in a fashion that is independent of chemoheterotrophic, photoautotrophic, or thermophilic lifestyle of different prokaryotes[17]. Other reasons that make amino acid biosynthetic pathway important in organisms under salinity stress are osmolyte synthesis and the pathways such as glutamic acid (proline), aspartate (ectoine), choline metabolism (glycine betaine) and myoinositol synthesis (pinitol) from which these osmolytes originate are situated in amino acid biosynthesis[18].

In the present study, codon usage pattern of the genes of *S. ruber* responsible for amino acid biosynthesis have been analyzed in order to determine codon composition and various factors affecting the synonymous codon usage bias. The pattern has been 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 multi-variate statistical analysis. The results have been presented graphically for better understanding of whole phenomena. Further, the optimal codons have been identified in highly expressed genes, 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 amino acid biosynthesis pathway in *S. ruber*. It may also help in understanding the effect of codon biasness on cellular systems and explaining the behaviour of the organism which may be replicated in similar prokaryotes[19].

Methodology

Ninety-six gene sequences involved in amino acid biosynthesis of *Salinibacter ruber* were retrieved from the Comprehensive Microbial Resource (http://cmr.jcvi.org/cgi-bin/CMR/CmrHome Page.cgi) in FASTA format. The list of genes according to their sub-functions is given in Table 1.

CDs with less than 300 base pairs (bps) were excluded from this analysis to avoid sampling error. Summary statistics such as mean, variance, percentage coefficient of variance etc. were calculated for the statistical properties of the codon usage. The percentage of codons with different values of each of the four nucleotides *i.e.* A, G, T and C at third position was calculated separately for all genes and represented as A_{3s}, G_{3s}, T_{3s} and C_{3s} respectively. The values of total number of G and C nucleotides in gene (GC), frequency of codons with G or C at third position (GC_{3s}), GC skewness [(G - C)/(G + C)], AT skewness [(A - T)/(A + T)], GC_{3s} skewness $[(G_{3s} - C_{3s})/(G_{3s} + C_{3s})]$, AT_{3s} skewness $[(A_{3s} - T_{3s})/(A_{3s} + T_{3s})]$ were calculated for each gene. RSCU which is defined as a ratio of observed frequency and expected frequency of a synonymous codon as well as value of effective number of codons under random codon usage condition by a gene. The effective number of codons used by gene (N_c) was calculated using following equation:

$N_c = 2 + s + [29/s^2 + (1- s^2)\}]$, where s is the value of GC_{3s}

For a single codon per amino acid, the N_c value is minimum (20), whereas for all codons with equal

Table 1—Genes of *S. ruber* involved in amino acid biosynthesis

Function	Sub Function	No. of genes
Amino acid biosynthesis	Aromatic amino acid family	16
	Asparate family	29
	Glutamate family	19
	Pyruvate family	12
	Serine family	11
	Histidine family	9
Total		96

probability, it has maximum value (61)[20]. The sequences having N_c values <30 are highly expressed, while those with >55 are poorly expressed genes[21,22].

In order to derive valid biological conclusions, multi-variate statistics using CA was applied. Major sources of variations in the genes lying along the axes were studied through correlation analysis for the explanation of variation and association of gene feature values with axes scores. The CA was also used for determining highly expressed genes and optimal codons *i.e.* synonymous codons frequently used in highly expressed genes. Further, stochastic distributional properties of RSCU for each codon within genes was studied using statistics related to central tendency (mean), dispersion (variance, standard errors) and stability (co-efficient of variance).

Results and Discussion

Analysis of codon usage pattern has both practical and theoretical importance in providing further insight into genetic organization of halophilic bacteria for its survival under high saline conditions. Gene sequence features, such as codon bias, base composition, GC_{3s}, GC_{3s} skew, gene expressivity level etc. can be better understood at the genomic scale level by combining statistical methodologies with advanced computer algorithms and data visualization through sophisticated graphical interfaces.

Codon usage analysis

A comprehensive analysis of codon usage is an essential prerequisite for understanding the trend of biased usage of synonymous codons in salt stress bacteria. In addition, codon usage profiles may be used to facilitate the design of primers and improve the accuracy of gene prediction from genomic sequences, as well as protein functional classification[23].

Summary statistics i.e. mean, standard error (SE), variance and percentage CV for RSCU value of each codon of the genes belonging to amino acid biosynthesis function in *S. ruber* is given in Table 2. Most preferred codons (codons preferred by the organism for translation to a particular amino acid) had less CV percentage as compared to their synonymous counter parts. For example, CUC had mean RSCU value of 2.92 and %CV 17.77% as compared to UUA (mean RSCU value 0.02 with %CV as 346.87%) for translation in leucine. This trend was observed for all synonymous codon usage families (Table 2). It was also found that as the preference of a particular codon decreased in translation process within the family, its stability of usage also decreased across genes. These results indicated that *S. ruber* consistently preferred optimal codons across genes belonging to the same function.

Average number of codons at A_{3s} (6%) and T_{3s} (8%) with % CV's of 45% and 37% respectively varied from that with C_{3s} (59%) and G_{3s} (48%) with % CV's of 10% and 11%. This indicated greater preferential stability in the usage of codons with C and G nucleotides at third position, as compared to A and T. Synonymous codon usage bias among the genes regulating amino acid biosynthesis in *S. ruber* reflected that the codons ending with C and G nucleotides were most preferred (Table 3). Earlier study reported that most bacterial genome use more G and C-ending codons, as compared to A and T-ending codons[24]. The results further depicted 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 were in agreement with the earlier reports in halophilic bacteria[25-27].

Codon usage in genomes with extreme composition (very high or very low GC content) is mostly shaped by mutational bias[26]. GC content (67%) in amino acid biosynthetic genes of *S. ruber* was quite high, indicating mutational bias. The average value of codon percentage with C or G nucleotides at third position was calculated to be 88% with %CV of 5%, while average number of codon of a gene in this function was 1172 with %CV of 44% indicated that the gene length was moderately varying in this function. The average number of effective codons in a gene that are responsible for translation into amino acids was 36 with % CV of 9%, suggesting high stability of the effective codons under random selection among genes in this function.

Nc plot analysis

Two different indices, namely Nc and GC_{3s} have been widely used to detect the codon usage variation among the genes[20]. A plot of Nc *versus* GC_{3s} was used to explore the codon usage variation among the amino acid biosynthesis genes of *S. ruber*.

The Nc plot (Fig. 1) showed that all the genes of amino acid biosynthetic pathway had high GC_{3s} content with low Nc value. Only few genes lied on the

Table 2—Summary statistics of codons of 96 genes from amino acid biosynthesis function

Amino acid	Codons	Mean	Std error	Variance	%CV	Amino acid	Codons	Mean	Std error	Variance	%CV
Phe	UUU	0.54	0.03	0.11	60.7	Try	UAU	0.2	0.03	0.07	131.26
	UUC	1.46	0.03	0.11	22.5		UAC	1.78	0.03	0.1	17.76
Leu	UUA	0.02	0.01	0	346.87	His	CAU	0.23	0.03	0.06	109.37
	UUG	0.23	0.02	0.05	99.28		CAC	1.77	0.03	0.06	14.03
	CUU	0.46	0.04	0.13	78.41	Gln	CAA	0.21	0.02	0.06	112.25
	CUC	2.92	0.05	0.27	17.77		CAG	1.79	0.02	0.06	13.42
	CUA	0.09	0.01	0.02	142.19	Asn	AAU	0.28	0.03	0.07	95.53
	CUG	2.29	0.06	0.35	25.73		AAC	1.7	0.03	0.1	18.96
Ile	AUU	0.72	0.05	0.23	66.6	Lys	AAA	0.35	0.04	0.14	105.56
	AUC	2.26	0.05	0.23	21.44		AAG	1.65	0.04	0.14	22.34
	AUA	0.02	0.01	0.01	370.62	Asp	GAU	0.25	0.02	0.02	61.75
Val	GUU	0.23	0.02	0.04	86.5		GAC	1.75	0.02	0.02	8.9
	GUC	1.59	0.05	0.21	28.55	Glu	GAA	0.38	0.02	0.03	48.49
	GUA	0.13	0.02	0.02	114.95		GAG	1.62	0.02	0.03	11.39
	GUG	2.05	0.05	0.23	23.45	Cys	UGU	0.43	0.05	0.24	114.24
Ser	UCU	0.28	0.04	0.13	126.7		UGC	1.49	0.06	0.33	38.51
	UCC	1.97	0.08	0.54	37.4	Arg	CGU	0.47	0.04	0.14	77.94
	UCA	0.1	0.02	0.03	177.04		CGC	2.8	0.07	0.49	25.16
	UCG	1.94	0.08	0.57	38.93		CGA	0.33	0.03	0.11	98.36
Pro	CCU	0.17	0.02	0.04	115.34		CGG	2.3	0.07	0.43	28.63
	CCC	1.81	0.05	0.28	29.28	Ser	AGU	0.24	0.03	0.09	121.73
	CCA	0.14	0.02	0.03	133.27		AGC	1.47	0.08	0.56	50.82
	CCG	1.88	0.05	0.28	28	Arg	AGA	0.04	0.01	0.01	229.28
Thr	ACU	0.07	0.01	0.01	158		AGG	0.05	0.01	0.01	208.1
	ACC	1.86	0.05	0.26	27.2	Gly	GGU	0.12	0.02	0.02	130.02
	ACA	0.16	0.02	0.04	120.22		GGC	2.27	0.05	0.27	22.92
	ACG	1.91	0.05	0.22	24.78		GGA	0.27	0.03	0.07	97.84
Ala	GCU	0.11	0.02	0.02	135.59		GGG	1.35	0.05	0.21	33.67
	GCC	2.31	0.04	0.12	15.21						
	GCA	0.17	0.01	0.02	84.19						
	GCG	1.41	0.03	0.12	24.25						

Table 3—Overall RSCU value for the amino acid biosynthetic genes in *S. ruber*

Amino acid	Codon	RSCU	Amino acid	Codon	RSCU	Amino acid	Codon	RSCU	Amino acid	Codon	RSCU
Phe	UUU	0.55	Ser	UCU	0.29	Tyr	UAU	0.17	Cys	UGU	0.49
	UUC	1.45		UCC	1.93		UAC	1.83		UGC	1.51
Leu	UUA	0.02		UCA	0.13	TER	UAA	0.53	TER	UGA	1.31
	UUG	0.22		UCG	1.99		UAG	1.16	Trp	UGG	1
	CUU	0.44	Pro	CCU	0.17	His	CAU	0.21	Arg	CGU	0.47
	CUC	2.91		CCC	1.79		CAC	1.79		CGC	2.81
	CUA	0.09		CCA	0.14	Gln	CAA	0.21		CGA	0.33
	CUG	2.31		CCG	1.9		CAG	1.79		CGG	2.29
Ile	AUU	0.72	Thr	ACU	0.07	Asn	AAU	0.3	Ser	AGU	0.25
	AUC	2.26		ACC	1.85		AAC	1.7		AGC	1.42
	AUA	0.02		ACA	0.16	Lys	AAA	0.34	Arg	AGA	0.04
Met	AUG	1		ACG	1.92		AAG	1.66		AGG	0.06
	GUU	0.23	Ala	GCU	0.1	Asp	GAU	0.24	Gly	GGU	0.1
Val	GUC	1.58		GCC	2.31		GAC	1.76		GGC	2.3
	GUA	0.13		GCA	0.16	Glu	GAA	0.37		GGA	0.25
	GUG	2.06		GCG	1.43		GAG	1.63		GGG	1.35

expected number of effective codon curve, indicating that there were few genes in which bias was dictated by GC_{3s}. Most of the genes lied below the curve with low Nc value and were in the narrow range of higher GC_{3s} values. These results indicated that apart from compositional bias, other factors, such as mutational as well as translational selection also play important role in shaping codon selection pattern in *S. ruber.*

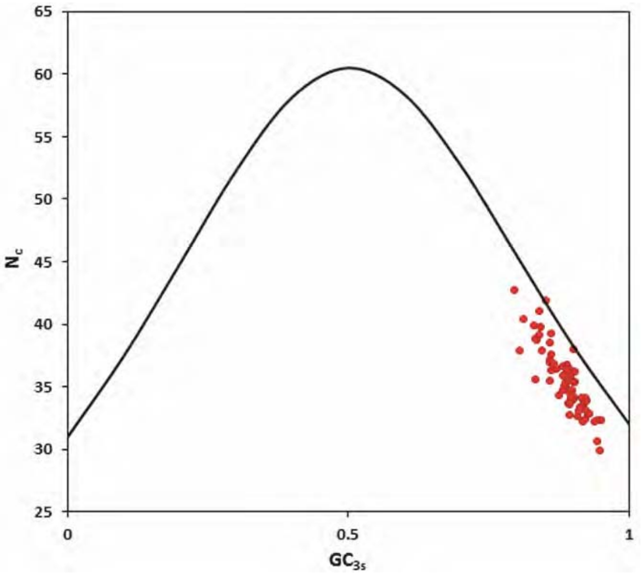

Fig. 1—N_c plot of 96 genes involved in amino acid biosynthesis pathway [The continuous curve represents the expected curve between GC_{3s} and N_c under random codon usage]

Multi-variate statistical analysis

Multi-variate statistical analysis has been widely used to study the codon usage variation among the genes in different organisms. Correspondence analysis (CA) is one of the multi-variate statistical technique in which the data are plotted in a multi-dimensional space of 59 axes (excluding Met, Trp and stop codons) and then it determines the most prominent axes contributing the codon usage variation among the genes[28].

In order to examine, if amino acid compositions exerted any constraint on synonymous codon usage, CA was performed on the data of RSCU values of genes. Major sources of variation among genes were identified through CA carried out with respect to four of its axes. It was observed that CA on codon count accounted for 12.30%, 7.65%, 6.06% and 4.46% of the total variation on the first, second, third and fourth major axes respectively, indicating first axis could be used for the analysis of codon usage variation in amino acid biosynthesis (Table 4). It was also observed that first axis of CA had statistically high positive correlation with GC_{3s} (0.93) and C_{3s} (0.71), whereas it had statistically high negative correlation coefficient with Nc (-0.92), T_{3s} (-0.86), A_{3s} (-0.79) and GC_{3s} skewness (-0.33). Variation of the first axis of CA could, therefore, be explained by these codon usage statistics.

Furthermore, correlation was also observed within different variables (Table 5). Nucleotides A and T at

Table 4—Correlation between axes generated by CA on codon usage with T_{3s}, C_{3s}, A_{3s}, GC_{3s}, GC skew, GC_{3s} skew, Gravy and N_c

	T_{3s}	C_{3s}	A_{3s}	G_{3s}	GC_{3s}	GC skew	GC_{3s} skew	Gravy	N_c
Axis 1	-0.86*	0.71*	-0.79*	0.16	0.93*	-0.13	-0.33	0.10	-0.92*
Axis 2	-0.13	0.43*	0.14	-0.50*	0.02	-0.49*	-0.50*	-0.10	0.04

*indicates significantly correlated with probability, P<0.01

Table 5—Correlation between T_{3s}, C_{3s}, A_{3s}, GC_{3s}, Gravy and N_c

	T_{3s}	C_{3s}	A_{3s}	G_{3s}	GC_{3s}	N_c	Gravy
T_{3s}	1.00*	-0.63*	0.64*	-0.11*	-0.92*	0.84*	-0.13
C_{3s}	-0.63*	1.00*	-0.42	-0.44*	0.61*	-0.63*	-0.10
A_{3s}	0.64*	-0.42*	1.00*	-0.39*	-0.88*	0.77*	-0.19
G_{3s}	-0.11	-0.44*	-0.39*	1.00*	0.27	-0.21	-0.10
GC_{3s}	-0.92*	0.61*	-0.88*	0.27	1.00*	-0.91*	0.13
N_c	0.84*	-0.63*	0.77*	-0.21*	-0.91*	1.00*	-0.09*
Gravy	-0.13	-0.10	-0.19	-0.10	0.13	-0.09	1.00

*indicates significantly correlated with probability, P<0.01

third position i.e. A_{3s} and T_{3s} showed a significant positive correlation with Nc, whereas *vice versa* was observed for G_{3s} and C_{3s}. Nc showed a negative correlation with GC_{3s}, which suggested a bias for A and/or T at third position. Also, GC_{3s} showed significant correlation with C_{3s} in comparison to G_{3s}, indicating that most biased codons used C at third position, which was also evident from Table 5. 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 (Rpurine, N- any nucleotide base and Y- pyrimidine) codons are more advantageous for translation[29]. Thus, compositional mutation bias possibly plays an important role in shaping the genes of *S. ruber*.

The pattern of correlation coefficient with respect to axis 1 of CA and with Nc is shown in Fig. 2a, whereas that with T_{3s}, C_{3s}, A_{3s}, G_{3s}, GC_{3s} is shown in Fig. 2b. The pattern indicated that axes of CA could provide information about the amino acid composition of the genes. Figure 3 shows the positions of the genes along first and second major axes produced by CA on the basis of RSCU value. The scatter plot of axis 1 and 2 of the CA with respect to RSCU values could differentiate genes through these axes of CA, but pattern of distribution varied according to the statistics. RSCU value indicated more differentiation of genes along axis 2 and concentration of genes was more around origin falling in I and IV quadrants.

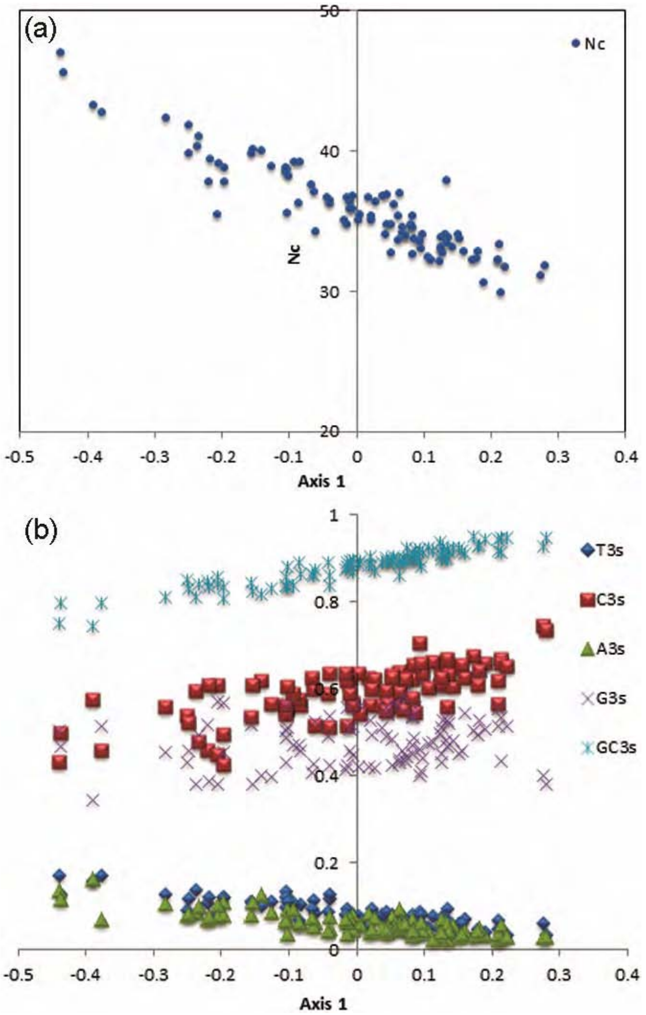

Fig. 2—(a): Scatter plot of axis 1 with Nc; and (b): Scatter plot of axis 1 with T_{3s}, C_{3s}, A_{3s}, G_{3s} and GC_{3s}

CA calculated on RSCU values showed 12.61% and 8.12% of the total variation on the first and second major axes, respectively. Thus, it could be speculated that in the genes responsible for amino acid biosynthesis, there was a single major explanatory axis which accounted the codon usage variation in *S. ruber*. These results indicated that gene expression levels were sufficient to discriminate genes according to their codon usage along the first major explanatory axis and amino acid compositions could not exert any constraints on this axis.

Effect of gene expression level on synonymous codon usage

Codon adaptation index (CAI) is a measurement of the relative adaptedness of the codon usage of a gene towards the codon usage of highly expressed genes[30]. It has been used widely to estimate the expressivities

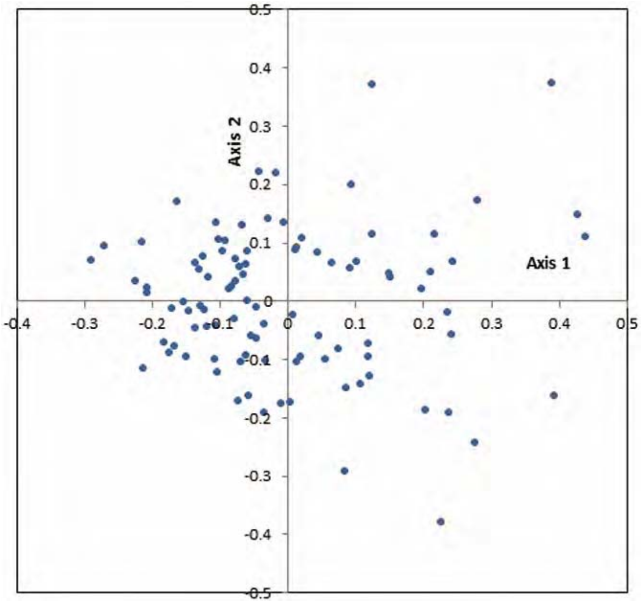

Fig. 3—Scatter plots of axis-1 and 2 of the correspondence analysis with RSCU value

of genes by many workers and is considered a well-accepted measure of gene expressivities[31-35].

In order to explore whether there was a correlation between the codon usage bias and the gene expression level, we calculated the correlation coefficient between CAI *versus* the positions of genes along the first major axis, nucleotide composition and N$_C$. Significant negative correlation was observed between the gene expression level assessed by CAI value and position of genes along axis 1 and Nc values (r = -0.67 and -0.50, respectively, P<0.01), while significant positive correlation was found between CAI value and GC3s content (r = 0.47, P<0.01). All the above correlations, except GC content were statistically significant and suggested that codon usage in particular gene was affected by gene expression level. Complete analysis suggested that genes with higher expression level, exhibiting a greater degree of codon usage bias were GC-rich and preferred the codons with C or G at the third position.

Aromaticity and hydrophobicity

GRAVY is an index which represents global hydrophobicity of proteins and measures variation in

Table 6—Codon usage of highly and lowly expressed amino acid biosynthetic genes of *S. ruber*

AA	Codon	RSCU^{a}	N^{a}	RSCU^{b}	N^{b}	AA	Codon	RSCU^{a}	N^{a}	RSCU^{b}	N^{b}
Phe	UUU	0.29	(7)	0.81	(25)	Ser	UCU	0.00	(0)	0.84	(14)
	UUC*	1.71	(42)	1.19	(37)		UCC	2.09	(31)	1.62	(27)
Leu	UUA	0.00	(0)	0.05	(1)		UCA	0.13	(2)	0.18	(3)
	UUG	0.04	(1)	0.59	(12)		UCG	2.02	(30)	1.98	(33)
CUU	0.16	(4)	0.98	(20)	Pro	CCU	0.05	(1)	0.24	(6)	
	CUC	3.00	(74)	2.61	(53)		CCC	1.95	(38)	1.70	(42)
	CUA	0.00	(0)	0.20	(4)		CCA	0.00	(0)	0.32	(8)
	CUG*	2.80	(69)	1.57	(32)		CCG	2.00	(39)	1.74	(43)
Ile	AUU	0.63	(18)	1.10	(15)	Thr	ACU	0.07	(2)	0.18	(4)
	AUC	2.37	(68)	1.90	(26)		ACC*	2.16	(66)	1.38	(31)
	AUA	0.00	(0)	0.00	(0)		ACA	0.03	(1)	0.40	(9)
Met	AUG	1.00	(51)	1.00	(30)		ACG	1.74	(53)	2.04	(46)
Val	GUU	0.03	(1)	0.51	(13)	Ala	GCU	0.02	(1)	0.31	(12)
	GUC	1.54	(50)	1.61	(41)		GCC*	2.55	(113)	1.97	(76)
	GUA	0.12	(4)	0.20	(5)		GCA	0.07	(3)	0.34	(13)
	GUG$^{@}$	2.31	(75)	1.69	(43)		GCG	1.36	(60)	1.38	(53)
Tyr	UAU	0.05	(1)	0.47	(10)	Cys	UGU	0.32	(3)	0.64	(7)
	UAC*	1.95	(36)	1.53	(33)		UGC	1.68	(16)	1.36	(15)
TER	UAA	0.00	(0)	0.75	(1)	TER	UGA	1.50	(2)	1.50	(2)
	UAG	1.50	(2)	0.75	(1)	Trp	UGG	1.00	(22)	1.00	(14)
His	CAU	0.17	(3)	0.46	(11)	Arg	CGU	0.12	(2)	0.71	(12)
	CAC	1.83	(32)	1.54	(37)		CGC*	3.76	(62)	1.65	(28)
Gln	CAA	0.03	(1)	0.59	(19)		CGA	0.24	(4)	1.00	(17)
	CAG*	1.97	(67)	1.41	(45)		CGG	1.82	(30)	2.41	(41)
Asn	AAU	0.05	(1)	0.93	(14)	Ser	AGU	0.13	(2)	0.36	(6)
	AAC*	1.95	(41)	1.07	(16)		AGC	1.62	(24)	1.02	(17)
Lys	AAA	0.18	(3)	0.72	(13)	Arg	AGA	0.06	(1)	0.06	(1)
	AAG*	1.82	(30)	1.28	(23)		AGG	0.00	(0)	0.18	(3)
Asp	GAU	0.13	(10)	0.32	(15)	Gly	GGU	0.00	(0)	0.21	(7)
	GAC$^{@}$	1.87	(149)	1.68	(80)		GGC*	2.99	(100)	1.56	(51)
Glu	GAA	0.24	(17)	0.49	(23)		GGA	0.18	(6)	0.40	(13)
	GAG$^{@}$	1.76	(123)	1.51	(70)		GGG	0.84	(28)	1.83	(60)

Superscript "a" represents for highly expressed genes and "b" for lowly expressed genes. Asterisk represents the codons occurring significantly more often in the highly expressed genes than that of lowly expressed genes.

amino acid composition among proteins[36]. Aromaticity is the frequency of aromatic amino acids in the translated gene product[37]. Both indices have been previously used to quantify the major CA trends in the amino acid usage of *E. coli* genes[35,36]. In our study, no significant correlation was observed for GRAVY with any of the variable. The average value of aromaticity was 0.06 with moderately high % CV i.e. 30%, whereas negative average Gravy value (-0.20) was observed with moderately low stability (% CV 82) among the genes of the function of amino acid biosynthesis of *S. ruber*. Thus, it can be inferred that both aromaticity and hydrophobicity of encoded proteins play minor role in shaping the codon usage bias among the genes.

Translational optimal codons

An optimal codon is the codon whose frequency of usage is significantly higher in putatively highly expressed genes[31]. Significance is estimated using $\chi 2$-based analysis with a cut-off at $p<0.01$ that compares observed codon frequencies to those expected, if codon usage reflects only local base composition at synonymous sites[38].

Based on $\chi 2$ contingency test, between highly expressed genes (10% of genes with high value of axis 1 of CA) and very low expressed genes (10% of genes with least value of axis 1 of CA), it was found that codons UUC, CUG, GUG, UAC, CAG, AAC, AAG, GAC, GAG, ACC, GCC, CGC and GGC were optimal codons (*i.e.* higher frequency of these codons in the highly expressed genes for this function in *S. ruber*) as shown in Table 6. Earlier, it has been shown that codons UUC, UAC, AAC and GAC are optimal in *E. coli, Bacillus subtilis, Saccharomyces cerevisiae, S. pombe* and *Drosophila melanogaste*[39] and are preferentially used in highly expressed genes. However, certain codons, such as AGG and AGA are commonly avoided in highly expressed prokaryotic genes[39]. Interestingly in our study, out of 13 codons that were found to be statistically over-represented in the genes located on the extreme left side of axis 1, there was 8 C-ending and 5 G-ending codons, which accounted for 62% C-ending and 38% G-ending codons. Moreover, axis 1 of CA based on RSCU value was highly correlated with GC_{3s} ($r = 0.93$, $P<0.01$).

In conclusion, the pattern of synonymous codon usage bias among the genes and their expressivity level can be derived with the help of statistical method including codon usage analysis. The study shows that three main factors *viz.* base composition, GC (AT) skew and translational selection play important role in shaping codon bias among the amino acid biosynthesis genes of *S. ruber*. Correspondence analysis has revealed differentiation of two groups of genes i.e., highly expressed and lowly expressed genes in *S. ruber*. It may be possible in the near future to determine the highly expressed genes in other prokaryotes that are salt stress in nature. The findings may also facilitate to determine the unique salt tolerant traits for their adaptation to high-salt environments. Genes responsible for salt tolerance from halophilic bacteria can further be employed for developing crop varieties with enhanced salt stress tolerance.

Acknowledgement

The authors acknowledge the NAIP for financial assistance of this study under the project entitled ‘Establishment of National Agricultural Bioinformatics Grid in ICAR’.

References

1 Garabito M J, Arahal D R, Mellado E, M. Marquez C & Ventosa A (1997) *Int J Syst Bacteriol* 47, 735-741

2 Anton J, Oren A, Benlloch S, Rodriguez-Valera F, Amann R & Rossello-Mora R (2002) *Int J Syst Evol Microbiol* 52, 485-491

3 Anton J, Rossello-Mora R, Rodriguez-Valera F & Amann R (2000) *Appl Environ Microbiol* 66, 3052-3057

4 Oren A (2008) *Saline Syst* 4:2 doi:10.1186/1746-1448-4-2

5 Oren A, Heldal M, Norland S & Galinski E A (2002) *Extremophiles* 6, 491-498

6 Mongodin M E F, Nelson K E, Duagherty S, DeBoy R T, Wister J, Khouri H, Weidman J, Balsh D A, Papke R T, Sanchez P G, Sharma A K, Nesbo C L, MacLeod D, Bapteste E, Doolittle W F, Charlebois R L, Legault B & Rodríguez-Valera F (2005) *Proc Natl Acad Sci* (USA) 102, 18147-18152

7 Oren A & Mana L (2002) *Extremophiles* 6, 217-223

8 Barton M D, Delneri D, Oliver S G, Rattray M & Bergman C M (2010) *PLoS ONE* 5(8): e11935. doi:10.1371/journal.pone.0011935

9 Grantham R, Gautier C, Gouy M & Pave A (1980) *Nucleic Acids Res* 8, r49-62

10 Grantham R, Gautier C & Gouy M (1980) *Nucleic Acids Res* 8, 1893-912

11 Ikemura T (1981) *J Mol Biol* 151, 389-410

12 Robinson M, Lilley R, Little S, Emtage J, Yarranton G, Stephens P, Millican A, Eaton M & Humphreys G (1984) *Nucleic Acids Res* 12, 6663-6671

13 Kanaya S, Yamada Y, Kudo Y & Ikemura T (1999) *Gene* 238, 143-155

14 Lafay B, Atherton J C & Sharp P M (2000) *Microbiology* 146, 851-860

15 Reis M, Wernisch L & Savva R (2003) *Nucleic Acids Res* 31, 6976-6985

16 Akashi H & Gojobori T (2002) *Proc Natl Acad Sci* (USA*)* 99, 3695-700

17 Heizer Jr E M, Raiford D W, Raymer M L, Doom T E, Miller R V & Krane D E (2006) *Mol Biol Evol* 23, 1670-1680

18 Hasegawa P M & Bressan R A (2000) *Annu Rev Plant Physiol Plant Mol Biol* 51, 463-499

19 Raiford D W, Heizer E M, Miller R V, Akashi H, Raymer M L & Krane D E (2008) *J Mol Evol* 67, 621-630

20 Wright F (1990) *Gene* 87, 23-29

21 Sau K, Gupta S K, Sau S & Ghosh T C (2005) *Virus Res* 113, 123-131

22 Sharp P M & Li W H (1986) *J Mol Evol* 24, 28-38

23 Lin K, Kuang Y, Joseph J S & Kolatkar P R (2002) *Nucleic Acids Res* 30, 2599-2607

24 Ermolaeva M D (2001) *Curr Issues Mol Biol* 3, 91-97

25 Lanyi J K (1974) *Bacteriol Rev* 38, 272-290

26 Oren A & Mana L (2002) *Extremophiles* 6, 217-223

27 Corcelli A, Lattanzio V M T, Mascolo G, Babudri F, Oren A & Kates M (2004) *Appl Environ Microbiol* 70, 6678-6685

28 Gupta S K, Bhattacharyya T K & Ghosh T C (2004) *J Biomol Struct Dynam* 21, 1-9

29 Shepherd J C (1981) *Proc Natl Acad Sci* (USA) 78, 1596-1600

30 Sharp P M & Li W H (1987) *Nucleic Acids Res* 15, 81-1295

31 Sharp P M & Cowe E (1991) *Yeast* 7, 657-678

32 Gutierrez G, Marquez L & Marin A (1996) *Nucleic Acids Res* 24, 2525-2527

33 Nakamura Y & Tabata S (1997) *Microbiol Comp Genomics* 2, 299-312

34 Tiller E R & Collins R A (2000) *J Mol Evol* 50, 249-257

35 Pan A, Dutta C & Das J (1998) *Gene* 215, 405-413

36 Lobry J R & Gautier C (1994) *Nucleic Acid Res* 22, 3174-80

37 Kyte J & Doolittle R F (1982) J *Mol Biol* 157, 105-132

38 Shields D C, Sharp P M, Higgins D G & Wright F (1988) *Mol Biol Evol* 5, 704-716

39 Sharp P M & Devine K M (1989) *Nucleic Acids Res* 17, 5029-5038