BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(22)

Trends in the codon usage patterns of *Chromohalobacter salexigens* genes

Rajkumari Sanjukta[1]*, Mohammad Samir Farooqi[1]*, Naveen Sharma[1], Anil Rai[1], Dwijesh Chandra Mishra[1] & Dhananjaya P Singh[2]

[1]Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Pusa, New Delhi – 110 012; [2]National Bureau of Agriculturally Important Microorganisms, Mau Nath Bhanjan, UP – 275 101; Mohammad Samir Farooqi – Email: samir@iasri.res.in; Rajkumari Sanjukta – Email: sanjuktark@iasri.res.in; Phone: 011-25847122-25 Ext 4217; Fax: 011-25841564; *Corresponding authors

Received September 22, 2012; Accepted October 06, 2012; Published November 13, 2012

Abstract:
Chromohalobacter salexigens, a Gammaproteobacterium belonging to the family Halomonadaceae, shows a broad salinity range for growth. In order to reveal the factors influencing architecture of protein coding genes in *C. salexigens,* pattern of synonymous codon usage bias has been investigated. Overall codon usage analysis of the microorganism revealed that C and G ending codons are predominantly used in all the genes which are indicative of mutational bias. Multivariate statistical analysis showed that the genes are separated along the first major explanatory axis according to their expression levels and their genomic GC content at the synonymous third positions of the codons. Both N_C plot and correspondence analysis on Relative Synonymous Codon Usage (RSCU) indicates that the variation in codon usage among the genes may be due to mutational bias at the DNA level and natural selection acting at the level of mRNA translation. Gene length and the hydrophobicity of the encoded protein also influence the codon usage variation of genes to some extent. A comparison of the relative synonymous codon usage between 10% each of highly and lowly expressed genes determines 23 *optimal codons*, which are statistically over represented in the former group of genes and may provide useful information for salt-stressed gene prediction and gene-transformation. Furthermore, genes for regulatory functions; mobile and extrachromosomal element functions; and cell envelope are observed to be highly expressed. The study could provide insight into the gene expression response of halophilic bacteria and facilitate establishment of effective strategies to develop salt-tolerant crops of agronomic value.

Keywords: Codon usage pattern, Correspondence analysis, Relative synonymous codon usage, Mutational bias, Halophilic bacteria.

Background:
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, thus for a gene using the universal code, 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 **[1]**. The unequal or preferred usage of a particular codon by an amino acid among the SF family is termed as synonymous codon usages (SCU). Specific SCU patterns may be due to mutational bias, bias in G+C content, natural selection etc. However, SCU pattern is non-random and species-specific **[2]**. It has also been reported that there is significant variation of codon usage bias among different genes within the same organism **[3-4]**.

Microorganisms belonging to diverse genetic lineages of bacteria and archeans are adapted to unusual limits of one or more abiotic factors in environment such as temperature, pressure, pH, radiation, salinity, etc. Salinity is an important deterrent to agriculture in many parts of the world, but investigations on its molecular effects are very few. Moderately halophilic bacteria, which are distributed through wide range of saline environments, constitute a heterogeneous group of microorganisms of different genera. These organisms are characterized by optimum growth at concentrations between 0.5M and 2.5M-NaCl **[5]**. Most of these have been isolated from either salted food or the Dead Sea, which are specialized hypersaline environments. The molecular basis of microbial resistance to salt stress is not fully understood in relation to regulation of gene expression during salinity, or anaerobic stress and this has little been examined in microorganisms. Understanding of molecular mechanisms involved in the halophilic adaptation will not only provide insight into factors responsible for genomic and proteomic stability under high salt conditions, but also, has importance for potential applications in the field of agriculture.

Chromohalobacter salexigens is a gram-negative aerobic bacterium, which is moderately halophilic in nature. It grows on a wide range of simple carbon compounds at NaCl concentrations between 0.5M and 4M, with an optimum growth at 2.0–2.5M and at an optimum temperature of 37°C **[6-9]**. So far codon usage bias in *C. salexigens* has not been investigated in detail, and it is not clear how different genes are expressed under saline environment in this organism. Therefore, it is of interest to understand factors that shape codon usage bias in this organism. Thus in this study, codon usage bias of *C. salexigens* genes was investigated using codon usage statistic, multivariate statistical technique and correlation analysis. The pattern of codon usage of this organism was studied based on values of codon usage indices and their correlation. The factors responsible for codon usage variation among genes were determined. Moreover, the expressivity level of genes, according to various functions was also determined with a view to understand the highly expressed genes and their optimal codons.

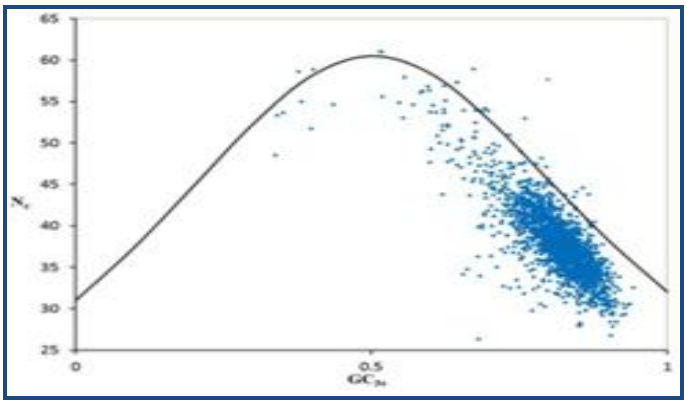


Figure 1: N_c plot of 2147 genes of *C. salexigens*. The continuous curve between GC_{3s} and N_c under random condon usage.

Methodology:

The gene sequences (2230) in FASTA format related to various functions in C*hromohalobacter salexigens* were retrieved from comprehensive microbial resource (http://www.tigr.org/CMR) and given in **Table 1 (see supplementary material)**. In order to minimise sampling errors, sequences with length < 300bp, redundant data and sequences with intermediate termination codons were excluded for this study. Thus the remaining dataset consisting of 2147 gene sequences were used for the analysis. PERL script was developed to merge all the individual gene sequences together for further data processing and analysis.

Codon Usage indices

In order to investigate the base composition of codons used by these genes, different statistic has been calculated. For each individual gene, the percentage of codons for 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. 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}) and GC skewness [(G-C)/(G+C)] were also calculated for each gene.

Measures of codon usage

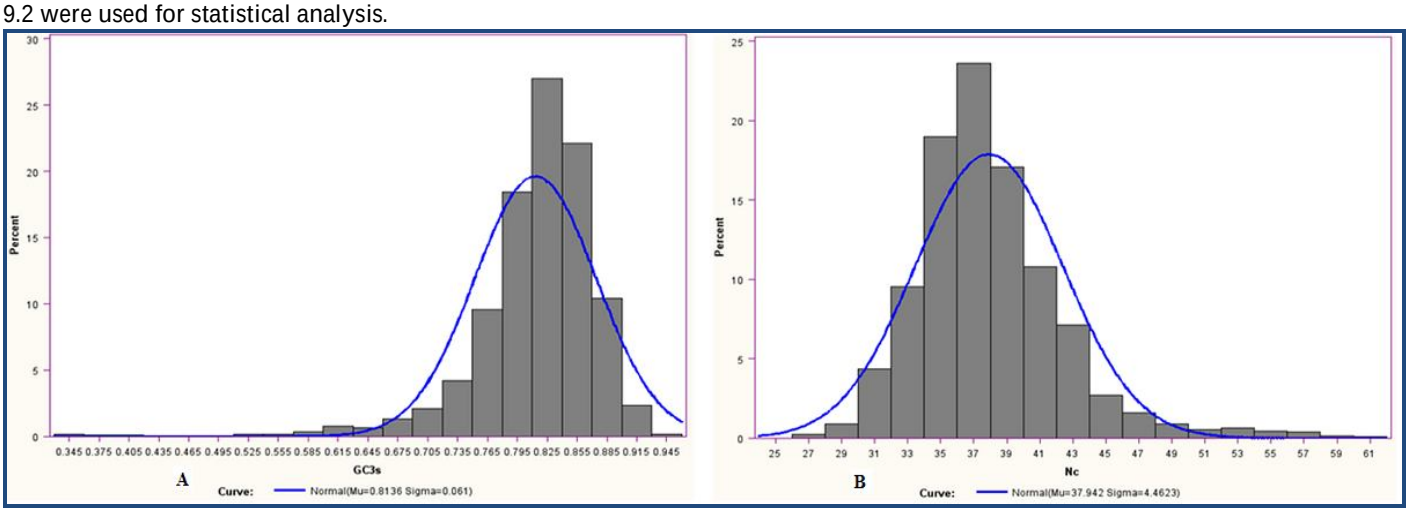
In order to investigate the characteristics of synonymous codon usage without the confounding influence of amino acid composition, the Relative Synonymous Codon Usage (RSCU) values among different codons in each gene was calculated. RSCU is defined as the ratio of the observed frequency of a codon to the expected frequency, if all the synonymous codons for those amino acids are used equally **[10].** RSCU values greater than 1.0 indicate that the corresponding codons are used more frequently than the expected frequency, whereas, the reverse is true for RSCU value less than 1.0. The effective number of codons of a gene (N_C) **[1]** was also used to quantify the codon usage bias of a gene (Wright 1990). It is calculated by the equation: $N_c = 2 + s + [29 / \{s^2 + (1- s^2)\}]$, where, s is the value of GC_{3s}. N_c can take values from 20 to 61, when only one codon, or all synonymous in equal frequencies were used per amino acid respectively. The sequences in which N_c values are <30 are considered as highly expressed, while those with >55 are considered as poorly expressed genes **[11-12]**. Another measure used for identification of gene expression is Codon Adaptation Index (CAI) is described in **supplementary material**. CAI value ranges from 0 to 1.0, and a higher value indicates a stronger codon bias and higher expression level **[13]**. In order to understand the properties of protein coding sequences, hydrophobicity 'GRAVY' score **[14]** and frequency of aromatic amino-acids 'Aromo' **[15]** in the translated gene product were also estimated.

Multivariate Statistical Analysis

Correspondence analysis (CA) is a data dimension reduction technique which takes multivariate data and combines them into a small number of variables (axes) that explains most of the variation among the original variables **[16]**. In this study, CA was applied to RSCU values of 59 codons (excluding Met, Trp, and stop codons). Further, correlation analysis ($P<0.01$) was used for explanation of variation and association of gene feature values with axes scores. Correlation analysis was also applied for the statistics obtained from base composition with N_c value. Chi squared contingency test ($P<0.01$) was performed to estimate the optimal codons i.e. synonymous codons frequently used in highly expressed genes.

Software implementation

CodonW **[17]** was employed for calculating the indices and measures of codon usage and also for CA. SPSS 17.0 and SAS 9.2 were used for statistical analysis.



Fiugure 2: **(A)** Frequency of $GC_{3s;}$ **(B)** Frequency of N_c

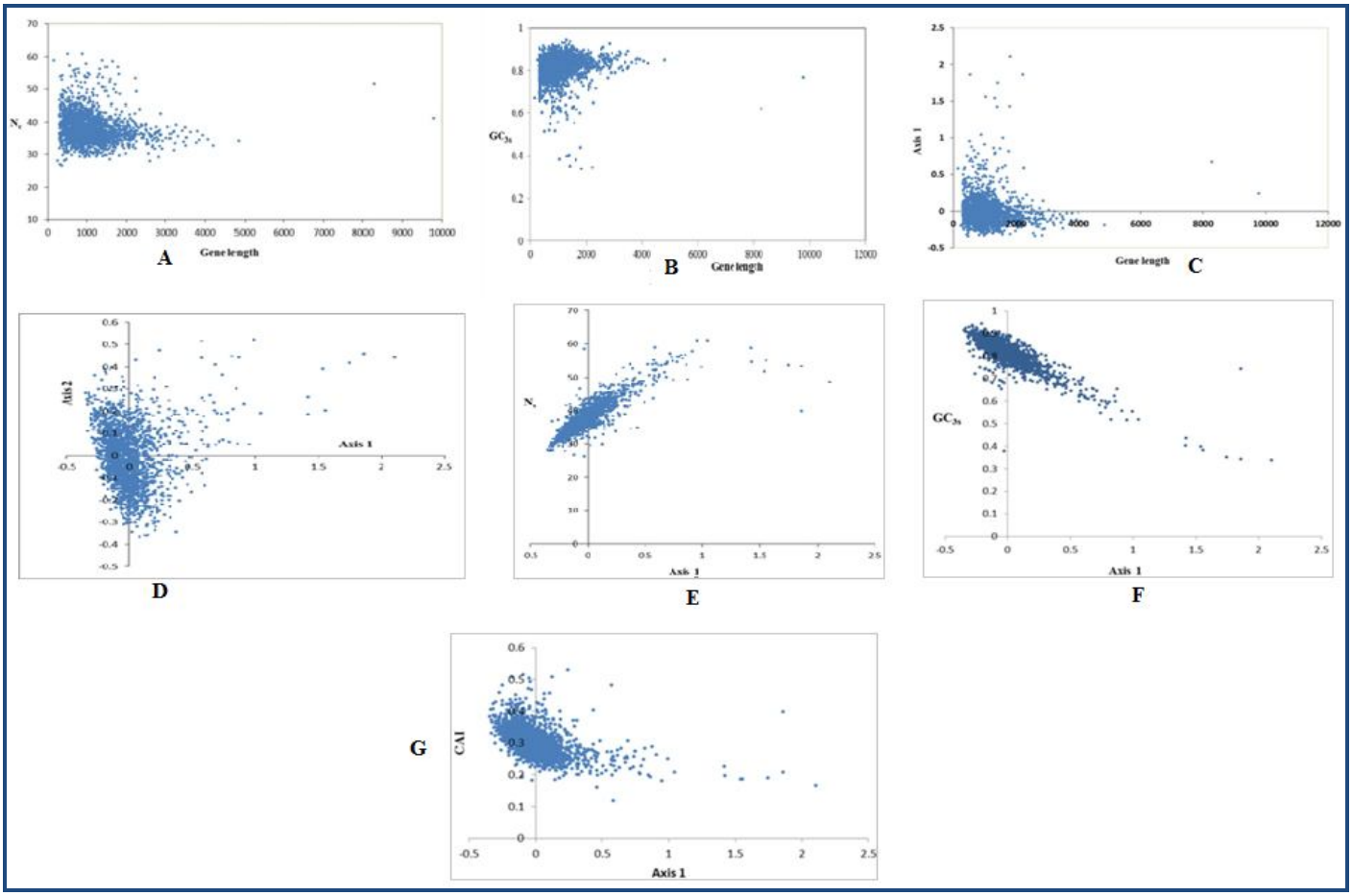


Figure 3: (A) Plot of Nc versus gene length; **(B)** Plot of GC3s versus gene length; **(C)** Plot of gene position on axis1 versus gene length; **(D)** Correspondence analysis of Relative Synonymous Codon Usage values of *C. salexigens genes*; **(E)** Scatter plot of gene position on axis 1 and Nc values; **(F)** Scatter plot of gene position on axis 1 and GC3s values; **(G)** Scatter plot of gene position on axis 1 and CAI values.

Discussion:

Codon usage analysis

Table 2 (see supplementary material) depicts the codon usage data and RSCU values for each codon. These values were calculated by summing all the 2147 genes of *C. salexigen.* It was derived that C_{3s} and G_{3s} accounted for 54.2% and 45.3% respectively, whereas, T_{3s} and A_{3s} accounted for 13.2% and 8.3% respectively. This suggests that, there was much greater

preferential stability in the usages of codons with C and G nucleotides at the third position as compared to A and T in the genes of this organism. Moreover, C ending codons were preferred over G ending codons. The average percentage of GC content among genes of this group was 64.9%, which is quite high and average value of GC skewness of the genes was low (0.0079), this indicated the influence of mutational bias.

Further, analysis and the usage of amino acid depict maximum usage of acidic amino acids (Asp, Glu), low proportion of hydrophobic amino acids and a high frequency of amino acids such as Gly and Ser which corroborated the work reported by Lanyi **[18]** for halophilic protein.

Heterogeneity of codon usage

In order to study the codon usage variation among genes of this organism, two different indices, namely, N_c and GC_{3s} were used to detect the codon usage variation among the genes **[19]**. Wright **[1]** suggested that 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 (a plot of N_c versus GC_{3s}). It is evident from N_c plot **(Figure 1)**, that a 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 considerably below the expected curve. This suggests that apart from compositional bias, majority of genes have an additional codon usage bias **[2]**. Furthermore, most of the genes in the N_c plot fall within a restricted cloud, at a relatively narrow range of GC_{3s} between 0.765 to 0.890 **(Figure 2 A)** with large variation of N_c values ranging between 30 and 45 **(Figure 2 B)**. This suggests that translational selection is also responsible for codon bias among the genes. However, the presence of significant negative correlation between GC_{3s} and N_c (r= -0.81, P<0.01) suggests strong influence of compositional constraints on codon usage bias in the genes of this organism.

Relationship between codon bias and gene length

In order to study the relationship between codon bias and gene length, a plot was drawn between gene length and N_c. From **Figure 3A,** it is understood that shorter genes have a much wider variance in N_c values, and vice versa for longer genes. The 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. A low negative correlation was observed with gene length against Nc (r= -0.15). This reveals that gene length has little influence on codon usage of these genes. However Eyre-Walker **[20]** 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. Correlation between gene length and values of GC_{3s} as well as gene position has also been worked out and similar type of result has been obtained i.e. low correlation **(Figure 3B & Figure 3C).** These findings suggest that gene length is not playing major role in the case of codon usage bias in this organism.

Multivariate Statistical Analysis

In order to investigate the major possible trends in shaping codon usage variation among the genes, Correspondence Analysis (CA)**,** a method of multivariate statistical analysis, was performed on the data of RSCU values of genes. Axis 1 shows the largest fraction (14.52%) of the variation in the data. Axis 2 describes the second largest trend (6.79%), and so on with each subsequent axis describing a progressively smaller amount of variation. Although, the first principal axis explains a substantial amount of variation of codon usage among the genes in this microorganism, its value is still lower than found in other organisms studied earlier **[22]**. The low value might be due to the extreme genomic composition of this genome. It is also obvious from **Figure 3D** that the majority of the points are clustered in an elliptical shape around the origin of axes. This indicates that, these genes have more or less similar codon usage biases. There are very few points that are widely scattered along the positive side of the first major axis, indicating that codon usage biases of these genes are not homogeneous. The first major axis is negatively correlated with G_{3s} and C_{3s} while correlated positively with A_{3s} and T_{3s} **Table 3 (see supplementary material)**. Also, strong positive correlation exists between position of genes along the first axis with N_c **(Figure 3E)** and high degree of negative correlation with GC_{3s} **(Figure 3F)**. These findings suggest, that highly biased genes, those ending with G and C, are clustered on the negative side, whereas, the codons ending in A and T predominate on the positive side of the first major axis. Further, 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), which suggests that highly expressed genes tend to use “C” or “G” at the synonymous positions as compared to lowly expressed genes in this organism. Further study reveals 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 **[23]**. Thus, compositional mutation bias possibly plays an important role in shaping the genome of this organism.

Effect of gene expression level on synonymous codon usage bias

CAI has been widely used to estimate the expressivities of genes by many workers and has been considered to be the best measure of gene expressivities **[24-26]**. In order to assess the effect of expressivities of genes on codon usage biases, codon adaptation index (CAI) values of all the genes of *C. salexigens* were calculated. In order to investigate, whether, there is a correlation between the codon usage bias and the gene expression level, the correlation coefficients were estimated between CAI values against the positions of genes along the first major axis, nucleotide composition and N_c values. From **Table 4 (see supplementary material)**, it is found that significant negative correlations exist between the gene expression level assessed by CAI value and the positions of genes along axis 1 and N_c values. A significant positive correlation between CAI value 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 bias among genes of *C. salexigens* is also affected by gene expression level. From the analysis, it can be suggested that genes with higher expression level, exhibiting a greater degree of codon usage bias and distributed at the left side of the first axis, are GC-rich and prefer to the codons with C or G at the synonymous position. A scatter diagram of the position of genes along the first major axis produced by CA on RSCU and their corresponding CAI values is shown **(Figure 3G)** and 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 (r= -0.49 (P< 0.01), confirming that axis 1 is significantly correlated with the expression level of each gene of *C. salexigens*. This is a clear indication that gene expression also affects the codon usage variation among the genes in this organism. Correlation analysis of synonymous codon usage bias against hydrophobicity of each protein was also investigated (r=0.22, P< 0.01). The findings indicated that genes, encoding more hydrophobic protein and bias to G/C bases at synonymous third codon positions, showed a stronger codon bias. This was also reported by Liu *et al.* **[27]** on synonymous codon usage in maize. Although, the absolute value of this correlation coefficient is low, but it is statistically significant. Subsequently, it has been inferred that the hydrophobicity of the encoded protein played a minor role in affecting codon usage in *C. salexigens* too. However, no significant correlation has been observed between synonymous codon bias and aromaticity scores.

Translational optimal codons

The χ^2 test (P<0.01) has been applied between top 10% genes (highly expressed) having higher value of the major axis and 10% genes (lowly expressed) having lower value of the axis. Through this analysis, twenty-three codons were determined as the 'optimal codons', which were significantly more frequent among the highly expressed genes. From **Table 5 (see supplementary material)**, it was derived that among 23 codons, there are 16 C-ending (69.6%) and 7-G ending codons (30.4%). These optimal codons might be significant to introducing point mutation, and modifying heterologous genes in order to increase the product of specific protein. Ikemura showed that there is a match between these codons and the most abundant tRNAs. In *Escherichia coli* **[28]**, *Drosophila melanogaster* **[29]** and *Caenorhabditis elegans* **[30]** highly expressed genes have a strong selective preference for codons with a high concentration for the corresponding acceptor tRNA molecule; the preferred codons are those which are best recognized by the most abundant tRNAs. This trend has been interpreted as the co-adaptation 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 codons in each gene and respective CAI value. This suggests that translational selection influenced the codon usage of *C. salexigens* and the optional codons are more frequent in highly expressed genes.

Categorization of genes based on expressivity

In order to differentiate the genes on the basis of expression level, the scores of axis 1 calculated for CA on RSCU values, were classified into three quantiles representing three different gene expression levels with corresponding cut-off percentages *viz.*, high (75% and above), moderate (between 25% and 75%), and low (below 25%) expressed genes as shown in **Table 6 (see supplementary material)**. The analysis depicted maximum percentage of genes belonging to category of moderately expressed genes. Further analysis revealed that genes related to regulatory functions, mobile and extrachromosomal element functions, cell envelope, protein function and cellular processes are highly expressed. Moderately expressed genes include functions related to DNA metabolism, transcription, signal transduction and protein synthesis, whereas, low expressed genes regulate central intermediary metabolism, purines, pyrimidines, nucleosides and nucleotides; amino acid biosynthesis; energy metabolism; and transport and binding protein **(Figure 4)**. This pattern of expression is expected for this halophilic bacterium.

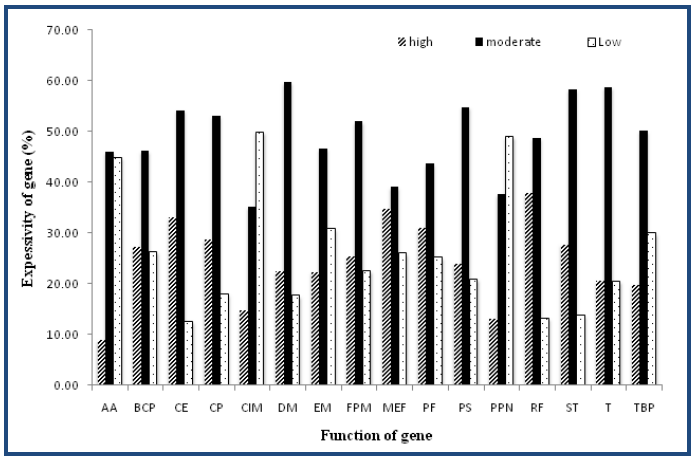


Figure 4: Expression level of functional genes of *C. salixigens* on the basis of Axis 1 based on RSCU values. **AA**= Amino acid biosynthesis; **BCP**=Biosynthesis of cofactor, Prosthetic groups and carriers; **CE**=Cell envelope; **CP**=Cellular processes; **CIM**=Central intermediary metabolism; **DM**=DNA metabolism; **EM**=Energy metabolism; **FPM**=Fatty acid & phospholipid metabolism; **MEF**=Mobile & extrachromosomal element functions; **PF**=Protein fate; **PS**=Protein synthesis; **PPN**=Purines, pyrimidines, nucleosides & nucleotides; **RF**=Regulatory functions; **ST**=Signal transduction; **T**=Transcription; **TBP**=Transport & binding protein

Conclusion:

It can be inferred that high level of heterogeneity is seen within the genes of various functions in *Chromohalobacter salexigens*. Though codon usage of *C. salexigens* is largely determined by compositional constraints, translational selection is also operating in shaping the codon usage variation among the genes. The study revealed that G/C-ending codons are preferred over A/T-ending codons in highly expressed genes. Total number of codons in highly expressed genes is much higher than those in lowly expressed genes. Length of the genes also affects the codon usage bias, while aromaticity and hydrophobicity of the encoded proteins play minor role in shaping codon usage bias. A set of twenty-three codons are identified as the optimal codons. Using χ^2 test at P<0.01, it was found that these codons are significantly more frequent among the highly expressed genes. As more genomes of halophilic bacteria with known gene sequences become available at public databases, it will be interesting to see if these effects are universal and whether these bacteria follow a similar trend of codon usage pattern for haloadaptation. The study could be explored to derive unique salt tolerant traits and for prediction of genes responsible for salt stress which could potentially be used in agricultural crops that are almost exclusively glycophytes.

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] Wright F, *Gene.* 1990 **87:** 23 [PMID: 2110097]
[2] Gupta SK *et al. J Biomol Struc Dyn.* 2004 **21:** 527 [PMID: 14692797]
[3] Gouy M & Gautier C, *Nucleic Acid Res.* 1982 **10:** 7055 [PMID: 6760125]
[4] Ikemura T, *Mol Biol Evol.* 1985 **2:** 13 [PMID: 3916708]
[5] Ventosa A *et al. Can J Microbiol.* 1984 **30:** 1279
[6] Canovas D *et al. J Bacteriol.* 1996 **178:** 7221 [PMID: 8955405]
[7] Arahal DR *et al. Int J Syst Evol Microbiol.* 2001 **51:** 1443 [PMID: 11491344]
[8] Connor KO & Csonka LN, *Appl Environ Microbiol.* 2003 **69:** 6334 [PMID: 14532102]
[9] Vargas C *et al. Saline Systems.* 2008 **4:** 14 [PMID: 18793408]
[10] Sharp PM & Li WH, *J Mol Evol.* 1986 **24:** 28 [PMID: 3104616]
[11] Sharp PM *et al. Nucleic Acid Res.* 1986 **14:** 5125 [PMID: 3526280]
[12] Sau K *et al. Virus Res.* 2005 **113:** 123 [PMID: 15970346]
[13] Sharp PM & Li WH, *Nucleic Acid Res.* 1987 **15:** 1281 [PMID: 3547335]
[14] Kyte J & Doolittle RF, *J Mol Biol.* 1982 **157:** 105 [PMID: 7108955]
[15] Lobry JR & Gautier C, *Nucleic Acid Res.* 1994 **22:** 3174 [PMID: 8065933]
[16] Suzuki H *et al. DNA Res.* 2008 **15:** 357 [PMID: 18940873]
[17] http://codonw.sourceforge.net/
[18] Lanyi JK, *Bacteriol Rev.* 1974 **38:** 272 [PMID: 4607500]
[19] Sahu K *et al. J Biochem Mol Biol.* 2004 **37:** 487 [PMID: 15469738]
[20] Eyre-Walker A, *Mol Biol Evol.* 1996 **13:** 864 [PMID: 8754221]
[21] Pan A *et al. Gene.* 1998 **215:** 405 [PMID: 9714839]
[22] Alvarez F *et al. Mol Biol Evol.* 1994 **11:** 790 [PMID: 7968492]
[23] Shepherd JC, *Proc Natl Acad Sci U S A.* 1981 **78:** 1596 [PMID: 6940175]
[24] Gutierrez G *et al. Nucleic Acids Res.* 1996 **24:** 2525 [PMID: 8692691]
[25] Nakamura Y & Tabata S, *Microbiol Comp Genomics.* 1997 **2:** 299 [PMID: 9689228]
[26] Tiller ER & Collins RA, *J Mol Evol.* 2000 **50:** 249 [PMID: 10754068]
[27] Liu H *et al. Mol Biol Rep.* 2010 **37:** 677 [PMID: 19330534]
[28] Ikemura T, *J Mol Biol.* 1981 **151:** 389 [PMID: 6175758]
[29] Moriyama EN & Powell JR, *J Mol Evol.* 1997 **45:** 514 [PMID: 9342399]
[30] Duret L, *Trends Genet.* 2000 **16:** 287 [PMID: 10858656]

Edited by P Kangueane
Citation: Sanjukta *et al*. Bioinformation 8(22): 1087-1095 (2012)

BIOINFORMATION

open access

Supplementary material:

Methodology:

Measures of codon usage

Formula used for calculating CAI

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L}\ln \omega_k\right)$$

Where ω_k = relative adaptedness of the k^{th} codon and L = number of synonymous codons in the gene.

Table 1: List of the functions and their sub-functions along with number of identified genes in genome of *C. salexigens (prior to sample minimization)*

Sl. No.	Function of genes	Total
1	Amino Acid Biosynthesis	101
2	Biosynthesis Of Cofactor, Prosthetic Groups & Carriers	109
3	Cell Envelope	146
4	Cellular Processes	188
5	Central Intermediary Metabolism	36
6	DNA Metabolism	113
7	Energy Metabolism	312
8	Fatty Acid & Phospholipid Metabolism	77
9	Mobile & Extrachromosomal Element Functions	34
10	Protein Fate	176
11	Protein Synthesis	160
12	Purines, Pyrimidines, Nucleosides & Nucleotides	53
13	Regulatory Functions	225
14	Signal Transduction	36
15	Transcription	52
16	Transport & binding Protein	412
	Grand total	2230

Table 2: Overall codon usage data of *C. salexigens* genes

AA	Codons	N	RSCU	AA	Codons	N
Phe	UUU	3989	0.29	Ser	UCU	1167
	UUC	23346	1.71		UCC	9092
Leu	UUA	545	0.04		UCA	1406
	UUG	9973	0.67		UCG	13627
	CUU	3265	0.22	Pro	CCU	2377
	CUC	21829	1.47		CCC	15410
	CUA	1280	0.09		CCA	1823
	CUG	52246	3.52		CCG	19154
Ile	AUU	5722	0.45	Thr	ACU	1681
	AUC	31012	2.46		ACC	23742
	AUA	1032	0.08		ACA	1921
Met	AUG	20118	1		ACG	14669
Val	GUU	2311	0.16	Ala	GCU	3899
	GUC	25639	1.73		GCC	49504
	GUA	2881	0.19		GCA	7171
	GUG	28379	1.92		GCG	31220
Tyr	UAU	7301	0.77	Cys	UGU	1372
	UAC	11775	1.23		UGC	5936
TER	UAA	321	0.45	TER	UGA	1546
	UAG	280	0.39	Trp	UGG	11103
His	CAU	8634	0.87	Arg	CGU	11852
	CAC	11311	1.13		CGC	33421
Gln	CAA	5091	0.34		CGA	2934
	CAG	25158	1.66		CGG	9394
Asn	AAU	5684	0.57	Ser	AGU	3163

	AAC	14280	1.43		AGC	14019
Lys	AAA	2679	0.26	Arg	AGA	471
	AAG	17678	1.74		AGG	953
Asp	GAU	17438	0.71	Gly	GGU	8201
	GAC	31573	1.29		GGC	40390
Glu	GAA	18269	0.73		GGA	4506
	GAG	31987	1.27		GGG	12063

Table 3: Correlation coefficient values of Axis 1 with four nucleotides at third codon position, N_c and GC_{3s}

	T_{3s}	C_{3s}	A_{3s}	G_{3s}	N_c	GC_{3s}
Axis 1	0.75*	-0.68*	0.70*	-0.15*	0.83*	-0.85*

* represents significantly correlated with probability, P<0.01, NS represents non-significant correlation

Table 4: Correlation coefficient values of CAI values with axis 1, nucleotide composition and N_c

	Axis 1	N_c	GC_{3s}	GC
CAI	-0.49*	-0.58*	0.30*	-0.13*

* represents significantly correlated with probability, P<0.01

Table 5: 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	0.1	74	0.76	492	Ser	UCU	0.04	14	0.7	286
	UUC*	1.9	1456	1.24	805		UCC*	1.82	643	0.95	389
Leu	UUA	0	2	0.3	205		UCA	0.07	26	0.65	266
	UUG	0.23	152	1.11	753		UCG*	1.69	595	1.29	525
	CUU	0.13	85	0.72	486	Pro	CCU	0.07	32	0.83	317
	CUC*	1.85	1232	1.19	804		CCC*	1.84	875	1.19	454
	CUA	0.03	18	0.37	250		CCA	0.05	22	0.65	247
	CUG*	3.76	2499	2.31	1561		CCG*	2.04	969	1.33	508
Ile	AUU	0.28	201	0.83	537	Thr	ACU	0.07	37	0.65	295
	AUC*	2.72	1982	1.77	1143		ACC*	2.95	1659	1.48	669
	AUA	0.01	7	0.4	261		ACA	0.06	32	0.58	260
Met	AUG	1	1157	1	890		ACG	0.93	522	1.29	581
Val	GUU	0.06	51	0.69	429	Ala	GCU	0.11	123	0.65	575
	GUC*	2.08	1652	1.41	875		GCC*	2.7	2982	1.61	1418
	GUA	0.11	85	0.5	313		GCA	0.18	202	0.66	580
	GUG*	1.74	1382	1.4	872		GCG	1.01	1115	1.08	955
Tyr	UAU	0.47	264	1.16	557	Cys	UGU	0.11	19	0.82	156
	UAC*	1.53	859	0.84	401		UGC*	1.89	338	1.18	225
TER	UAA	0.76	27	0.67	24	TER	UGA	2.02	72	1.88	67
	UAG	0.22	8	0.45	16	Trp	UGG	1	507	1	429
His	CAU	0.57	276	1.09	431	Arg	CGU	0.89	385	1.45	567
	CAC*	1.43	684	0.91	363		CGC*	4.34	1872	2.04	798
Gln	CAA	0.17	137	0.72	490		CGA	0.1	45	0.7	272
	CAG*	1.83	1446	1.28	880		CGG	0.63	270	0.98	384
Asn	AAU	0.29	194	0.93	482	Ser	AGU	0.17	61	0.85	347
	AAC*	1.71	1137	1.07	558		AGC*	2.2	775	1.56	635
Lys	AAA	0.13	103	0.79	423	Arg	AGA	0	2	0.42	164
	AAG*	1.87	1494	1.21	644		AGG	0.03	12	0.42	163
Asp	GAU	0.42	603	1.06	1023	Gly	GGU	0.37	328	0.92	654
	GAC*	1.58	2294	0.94	908		GGC*	3.08	2740	1.65	1178
Glu	GAA	0.73	1080	1	1023		GGA	0.11	102	0.62	442
	GAG*	1.27	1884	1	1032		GGG	0.44	389	0.81	574

∗Codons whose occurrences are significantly higher (P < .01) in the extreme left side of axis 1 than the genes present on the extreme right of the first major axis. Each group contains 10% of genes at either extreme of the major axis generated by correspondence analysis. AA: amino acid; N: number of codon; [1]: genes on extreme left of axis 1; 2 genes on extreme right of axis 1.

Table 6: List of number and percentage of genes on basis of their expression level

Sl. No	Function	High		Moderate		Low		Total
		No. of genes	% of genes	No. of genes	% of genes	No. of genes	% of genes	
1	Amino acid biosynthesis	26	26	55	55	19	19	**100**
2	Biosynthesis of cofactor, prosthetic groups and carriers	25	23.6	60	56.6	21	19.8	**106**
3	Cell envelope	34	23.9	61	43.0	47	33.1	**142**
4	Cellular processes	39	22.0	93	52.5	45	25.4	**177**
5	Central intermediary metabolism	12	35.3	15	44.1	7	20.6	**34**
6	DNA metabolism	18	16.8	60	56.1	29	27.1	**107**
7	Energy metabolism	92	29.0	155	48.9	70	22.1	**317**
8	Fatty acid and phospholipid metabolism	20	26.7	30	40.0	25	33.3	**75**
9	Mobile and extrachromosomal element functions	6	26.1	12	52.2	5	21.7	**23**
10	Protein fate	51	29.3	85	48.9	38	21.8	**174**
11	Protein synthesis	33	24.8	60	45.1	40	30.1	**133**
12	Purines, pyrimidines, nucleosides and nucleotides	18	34.0	27	50.9	8	15.1	**53**
13	Regulatory functions	47	22.3	114	54.0	50	23.7	**211**
14	Signal transduction	7	19.4	20	55.6	9	25.0	**36**
15	Transcription	5	10.0	28	56.0	17	34.0	**50**
16	Transport and binding protein	103	25.2	199	48.7	107	26.2	**409**
	Total	**536**		**1074**		**537**		**2147**