Statistical Analysis of codon usage in extremely halophilic bacterium, 
*Salinibacter ruber* DSM 13855
Sanjukta RK¹, Farooqi MS¹*, Sharma N¹, Rai N², Mishra DC¹, Rai Â¹, Singh DP³ and Chaturvedi KK¹

¹Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Pusa, New Delhi – 110 012, India
²Central Drug Research Institute, Jhankipuram, Lucknow-226 021
³National Bureau of Agriculturally Important Microorganisms, Mau Nath Bhanjan, UP – 275 101, India

*E-mail: samir@iasri.res.in, Tel: 011-25847121-2 Ext 4305, Fax: 011-25841564

ABSTRACT

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

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

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

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

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

**MATERIALS AND METHODS**

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

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

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Functions</th>
<th>No. of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Amino Acid Biosynthesis</td>
<td>96</td>
</tr>
<tr>
<td>2</td>
<td>Biosynthesis Of Cofactor, Prosthetic Groups &amp; Carriers</td>
<td>76</td>
</tr>
<tr>
<td>3</td>
<td>Cell Envelope</td>
<td>91</td>
</tr>
<tr>
<td>4</td>
<td>Cellular Processes</td>
<td>131</td>
</tr>
<tr>
<td>5</td>
<td>Central Intermediary Metabolism</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>Disrupted Reading Frame</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>DNA Metabolism</td>
<td>113</td>
</tr>
<tr>
<td>8</td>
<td>Energy Metabolism</td>
<td>232</td>
</tr>
<tr>
<td>9</td>
<td>Fatty Acid &amp; Phospholipid Metabolism</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>Mobile and Extrachromosomal element functions</td>
<td>51</td>
</tr>
<tr>
<td>11</td>
<td>Protein Fate</td>
<td>135</td>
</tr>
<tr>
<td>12</td>
<td>Protein Synthesis</td>
<td>113</td>
</tr>
<tr>
<td>13</td>
<td>Purines, Pyrimidines, Nucleosides and Nucleotides</td>
<td>62</td>
</tr>
<tr>
<td>14</td>
<td>Regulatory Functions</td>
<td>67</td>
</tr>
<tr>
<td>15</td>
<td>Signal Transduction</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td>Transcription</td>
<td>45</td>
</tr>
<tr>
<td>17</td>
<td>Transport &amp; binding Protein</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>1450</strong></td>
</tr>
</tbody>
</table>
The codon usage analysis of these 1450 gene sequences has been performed to study (i) base composition analysis of the codons (ii) codon usage bias.

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

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

\[
\text{RSCU} = \frac{g_{ij}}{\sum_{j} n_{ij}}
\]

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

\[
N_{C} = 2 + s + \left[\frac{29}{(s^{2} + (1 - s^{2})}\right],
\]

where, \(s\) is the value of GC\(_{3s}\).

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

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

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

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

RESULTS AND DISCUSSION

Codon usage analysis:
The frequency of different codons in their respective synonymous family of 1450
genes along with its RSCU values within each family are shown in Table 1. From this
table, the value of C\textsubscript{s} and G\textsubscript{s} can be derived as 44.6% and 37.2% respectively,
whereas, it can be seen that value of T\textsubscript{s} and A\textsubscript{s} consist of 11.1% and 7.1%
respectively. This suggests that the codons with G and/or C at the third position in all
synonymous codon family are predominately used. Further, C ending codons are
preferred over G ending codons. The result depicts maximum usage of acidic amino
acids i.e. Asp, Glu, low proportion of hydrophobic amino acids and a high frequency
of amino acids such as Gly and Ser. These results are in the line of work reported by
Lanyi 1974; Oren and Mana, 1987 for halophilic protein. Furthermore, it is observed
that the average percentage of GC content among genes is 65%, which is quite high
and average value of GC skewness of the genes is low (0.0079) which is an indicative
of mutational bias. This clearly shows that, there is much greater preferential stability in the usages of codons with C and G nucleotides at the third position. RSCU value of all codons ending with C and G are >1.0 indicating the biased codon usage behaviour towards these codons in their respective synonymous codon family.

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

Table 2 Overall RSCU value for the genes of all functions

<table>
<thead>
<tr>
<th>AA</th>
<th>Codon</th>
<th>N</th>
<th>RSCU</th>
<th>AA</th>
<th>Codon</th>
<th>N</th>
<th>RSCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>UUU</td>
<td>3642</td>
<td>0.58</td>
<td>Glu</td>
<td>GAG</td>
<td>17774</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>UUC</td>
<td>9008</td>
<td>1.42</td>
<td>Ser</td>
<td>UCU</td>
<td>6012</td>
<td>0.63</td>
</tr>
<tr>
<td>Leu</td>
<td>UUA</td>
<td>1043</td>
<td>0.17</td>
<td>UCC</td>
<td>11095</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>4289</td>
<td>0.71</td>
<td>UCA</td>
<td>7289</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUU</td>
<td>4410</td>
<td>0.73</td>
<td>UCG</td>
<td>18538</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td>13142</td>
<td>2.17</td>
<td>AGU</td>
<td>3799</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUA</td>
<td>2275</td>
<td>0.37</td>
<td>AGC</td>
<td>10404</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td>11254</td>
<td>1.85</td>
<td>Pro</td>
<td>CCU</td>
<td>8681</td>
<td>0.65</td>
</tr>
<tr>
<td>Ile</td>
<td>AUU</td>
<td>3295</td>
<td>0.75</td>
<td>CCC</td>
<td>15010</td>
<td>1.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>9162</td>
<td>2.08</td>
<td>CCA</td>
<td>8183</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td>736</td>
<td>0.17</td>
<td>CCG</td>
<td>21749</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>AUG</td>
<td>6241</td>
<td>1</td>
<td>Thr</td>
<td>ACU</td>
<td>4170</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>GUU</td>
<td>3179</td>
<td>0.46</td>
<td>ACC</td>
<td>14511</td>
<td>1.29</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>10211</td>
<td>1.47</td>
<td>ACA</td>
<td>6075</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2257</td>
<td>0.32</td>
<td>ACG</td>
<td>20205</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12217</td>
<td>1.75</td>
<td>Ala</td>
<td>7523</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1047</td>
<td>0.29</td>
<td>GCC</td>
<td>22967</td>
<td>1.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>UAC</td>
<td>6278</td>
<td>1.71</td>
<td>GCA</td>
<td>9063</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>742</td>
<td>0.32</td>
<td>Cys</td>
<td>UGU</td>
<td>3658</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5660</td>
<td>2.42</td>
<td>UGC</td>
<td>9831</td>
<td>1.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>CAU</td>
<td>4401</td>
<td>0.64</td>
<td>Trp</td>
<td>UGG</td>
<td>12693</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>CAC</td>
<td>9333</td>
<td>1.36</td>
<td>Arg</td>
<td>CGU</td>
<td>9200</td>
<td>0.66</td>
</tr>
<tr>
<td>Gln</td>
<td>CAA</td>
<td>4209</td>
<td>0.61</td>
<td>CGC</td>
<td>19369</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAG</td>
<td>9563</td>
<td>1.39</td>
<td>CGA</td>
<td>14869</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>AAU</td>
<td>2052</td>
<td>0.48</td>
<td>CGG</td>
<td>22871</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>6450</td>
<td>1.52</td>
<td>AGA</td>
<td>6479</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>AAA</td>
<td>3090</td>
<td>0.61</td>
<td>AGG</td>
<td>11392</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG</td>
<td>7039</td>
<td>1.39</td>
<td>Gly</td>
<td>GGU</td>
<td>5912</td>
<td>0.45</td>
</tr>
<tr>
<td>Asp</td>
<td>GAU</td>
<td>5528</td>
<td>0.44</td>
<td>GGC</td>
<td>20183</td>
<td>1.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAC</td>
<td>19546</td>
<td>1.56</td>
<td>GGA</td>
<td>10112</td>
<td>0.76</td>
<td></td>
</tr>
</tbody>
</table>
Heterogeneity of codon usage:
Two indices, viz. \( N_c \) and \( GC_{3s} \) are generally used to study the codon usage variation among the genes in any organism (Sahu et al. 2004). Figure 1 shows the \( N_c \) distribution of different genes in *S. ruber*. The \( N_c \) values range from 29.17 to 61 (with a mean of 37.85 and standard deviation of 5.73), indicating that there is a wide variation of codon usage bias among the genes. The heterogeneity of codon usage biases among the genes is further confirmed from the distributions of \( GC_{3s} \), shown in Figure 2. It is obvious that \( GC_{3s} \) varies from 23 to 97% with a mean of 85% and standard deviation of 7.8%. These results indicate that apart from compositional constraints, other factors might influence the overall codon usage variation among the genes in *S. ruber*.

![Figure 1: Distribution of effective number of codons \( (N_c) \) in *S. ruber* genes.](image)

Glu GAA 6763 0.55 GGG 16853 1.27
AA represents amino acid, \( N \) is the number of codons and RSCU represents relative synonymous codon usage.
Total number of genes=1450, total number of codons= 585618.
Exploration of different factors in determining the codon usage variation

i. The $N_c$ plot:

It was demonstrated by Wright 1990 that the comparison of the actual distribution of genes with the expected distribution under no selection could be indicative, if the codon usage bias of the genes had some influence other than the genomic GC composition. In other words, genes whose codon choice is constrained only by a G+C mutational bias, will lie on or just below the curve of the predicted value in the $N_c$ plot ($N_c$ versus GC$_{3s}$). From $N_c$ plot (Figure 3), it is evident that few points lie on the expected curve towards GC rich regions, which certainly originates from the extreme compositional constraints. However, considerable number of points with low $N_c$ values lie below the expected curve. This suggests that majority of genes have an additional codon usage bias apart from compositional bias. Moreover, it can be seen that almost all genes lie below the curve with low $N_c$ value (30.60 to 42.74) and falls in the narrow range of higher GC$_{3s}$ value (0.8 to 0.95). This suggests that translational selection is also responsible for codon bias among the genes. However, strong influence of compositional constraints on codon usage bias in the genes could be understood from the presence of significant negative correlation between GC$_{3s}$ and $N_c$ ($r = -0.94$, $P< 0.01$).
Figure 3: $N_c$ plot of 1450 genes of *S. ruber*. The continuous curve represents the expected curve between $GC_{3s}$ and $N_c$ under random codon usage.

**ii. Multivariate Statistical Approach:**

The dataset of RSCU values of 1,450 genes is subjected to correspondence analysis (CA), a method of multivariate statistical analysis (MVA). For large multidimensional datasets, CA allows a reduction in the dimensionality of the data so that an efficient visualization that captures most of the variation can occur (Greenacre 1984). In this study, CA has been performed on RSCU values to minimize the effects of amino acid composition. The most prominent axes contributing to the codon usage variation among the genes are determined. It is seen that axis 1 has the largest fraction of the variation (12.61%) in the data. Axis 2 (6.79%) describes the second largest trend, and so on with each subsequent axis describing a progressively smaller amount of variation. It must be remembered that although the first axis explains a substantial amount of variation, its value is still lower than found in other organisms studied earlier (Eyre-Walker 1996). The low value might be due to the extreme genomic composition of this organism. It is also obvious from Figure 4 that the majority of the points are clustered around the origin of axes indicating that these genes have more or less similar codon usage biases. However, few points are widely scattered along the negative side of axis 1, which suggest that codon usage biases of these genes are not homogeneous. Genes with low $N_c$ value indicate highly bias and vice-versa for the unbiased genes. It is noticeable from the figure that highly biased genes ($N_c \leq 35$) are scattered towards extreme positive side of axis 1 (highlighted with blue). Genes of $N_c$ values (35 to 50) are dispersed in the middle while high $N_c$ value (>50) genes are widely scattered towards the extreme negative side of the axis. Axis 1 is significantly positively correlated with $G_{3s}$ ($r=0.55$, $P<0.01$) and $C_{3s}$ ($r=0.75$, $P<0.01$) while significantly negatively correlated with $A_{3s}$ ($r=-0.91$, $P<0.01$) and $T_{3s}$ ($r=-0.92$, $P<0.01$).
Also, strong significant negative correlation exists between position of genes along the first axis with \( N_c \) (\( r=-0.95, \ P<0.01 \)) and high degree of significant positive correlation with \( GC_{3s} \) (\( r=0.97, \ P<0.01 \)). These findings suggest that highly biased genes, those ending with G and C, are clustered on the positive side, whereas those of A and T predominate on the negative side of the first major axis. Additionally, significant negative correlation is observed with \( N_c \) against \( GC_{3s} \) (\( r=-0.81, \ P<0.01 \)) and GC (\( r=-0.43, \ P<0.01 \)). Highly expressed genes tend to use C or G at the synonymous positions as compared to lowly expressed genes. It is also studied that C-ending codons are preferred over G-ending codons in highly expressed genes. Preference of C-ending codons in the highly expressed genes might be related to the translational efficiency of the genes as it has been reported that RNY (R-Purine, N-any nucleotide base, and Y-pyrimidine) codons are more advantageous for translation (Alvarez et al. 1994). Thus, compositional mutation bias possibly plays an important role in shaping the codon usage pattern of this organism.

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

It has been noted that in organisms with a highly skewed base composition, mutational bias is the main factor in shaping the codon usage variation among the genes whereas translational selection plays a minor role (Gupta and Ghosh 2001). Overall RSCU values (Table 1) and \( N_c \) plot (Figure 3) markedly indicate that mutational bias is key determinant of codon usage variation among the genes. However, correspondence analysis indicates that there is a single major trend in the codon usage among the genes in this bacterium. To assess the effect of expressivities of genes on codon usage biases, codon adaptation index (CAI) of *S. ruber* genes has been calculated. CAI has been considered as an
effective measure of gene expressivities (Gutierrez et al. 1996; Nakamura and Tabata 1997; Tiller and Collins 2000). The correlation coefficients are estimated for CAI values against the positions of genes along the first major axis, nucleotide compositions and \( N_c \) values.

Table 3 Correlation analysis data

<table>
<thead>
<tr>
<th>CAI</th>
<th>Axis1</th>
<th>(N_c)</th>
<th>GC(_{3s})</th>
<th>G(_{3s})</th>
<th>C(_{3s})</th>
<th>A(_{3s})</th>
<th>T(_{3s})</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.47*</td>
<td>-0.48*</td>
<td>0.42*</td>
<td>0.19*</td>
<td>0.68*</td>
<td>-0.36*</td>
<td>-0.36*</td>
<td>-0.11</td>
<td></td>
</tr>
</tbody>
</table>

* represents significantly correlated with probability, \( P<0.01 \)

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

![Figure 5: The Scatter diagram of gene position on axis 1 and CAI values](image)

Correlation analysis of synonymous codon usage bias against hydrophobicity of each protein has also been investigated \((r=0.22, \ P<0.01)\). The findings indicate that genes, encoding more hydrophobic protein and bias to G/C bases at synonymous third codon positions, show a stronger codon bias which was also reported by Liu et al. 2010. Although the absolute value of this correlation
coefficient is low, it is statistically significant. Subsequently, it can be inferred that the hydrophobicity of the encoded protein play a minor role in affecting codon usage. However, no significant correlation has been observed between synonymous codon bias and aromaticity scores.

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

![Figure 6: Plot of gene position on axis1 versus gene length](image)
Translational optimal codons:

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

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

<table>
<thead>
<tr>
<th>AA</th>
<th>Codon</th>
<th>RSCU₁</th>
<th>N₁</th>
<th>RSCU²</th>
<th>N²</th>
<th>AA</th>
<th>Codon</th>
<th>RSCU₁</th>
<th>N₁</th>
<th>RSCU²</th>
<th>N²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>UUU</td>
<td>0.36</td>
<td>214</td>
<td>0.91</td>
<td>333</td>
<td>Glu</td>
<td>GAG*</td>
<td>1.77</td>
<td>3017</td>
<td>1.17</td>
<td>1020</td>
</tr>
<tr>
<td></td>
<td>UUC*</td>
<td>1.64</td>
<td>975</td>
<td>1.09</td>
<td>395</td>
<td>Ser</td>
<td>UCU</td>
<td>0.07</td>
<td>20</td>
<td>0.86</td>
<td>250</td>
</tr>
<tr>
<td>Leu</td>
<td>UUA</td>
<td>0</td>
<td>0</td>
<td>0.21</td>
<td>68</td>
<td>UCC*</td>
<td>1.75</td>
<td>518</td>
<td>1.13</td>
<td>326</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>0.07</td>
<td>40</td>
<td>0.77</td>
<td>245</td>
<td>UCA</td>
<td>0.03</td>
<td>9</td>
<td>0.71</td>
<td>206</td>
<td></td>
</tr>
</tbody>
</table>
CUU  0.23  131  1.2  382  UCG*  2.11  624  1.23  356
CUC*  3.02  1719  1.79  572  Ser  AGU  0.16  46  0.76  219
CUA  0.05  29  0.49  158  AGC*  1.89  558  1.31  380
CUG*  2.63  1496  1.54  491  Pro  CCU  0.03  14  0.98  283
Ile  AUU  0.47  269  0.98  315  CCC*  1.71  714  0.78  224
AUC*  2.53  1439  1.66  531  CCA  0.04  16  0.88  253
AUA  0  1  0.36  116  CGG*  2.22  926  1.36  393
Met  AUG  1  838  1  350  Thr  ACU  0.02  10  0.67  209
Val  GUU  0.09  63  0.72  274  ACC*  1.95  1101  1.28  400
GUC*  1.58  1127  1.34  509  ACA  0.05  27  0.7  220
GUA  0.06  41  0.64  245  ACG*  1.98  1117  1.35  423
GUG*  2.27  1621  1.29  492  Ala  GCU  0.03  30  0.93  417
Tyr  UAU  0.07  35  0.74  243  GCC*  2.48  2366  1.23  551
UAC*  1.93  953  1.26  414  GCA  0.07  70  0.79  351
TER  UAA  0.58  14  0.67  39  GCG*  1.41  1346  1.05  467
UGA  1.5  36  0.79  46  Cys  UGU  0.13  17  0.79  125
UGA  0.92  22  1.53  89  UGC*  1.87  251  1.21  190
His  CAU  0.05  23  0.85  270  Trp  UGG  1  354  1  404
CAC*  1.95  830  1.15  364  Arg  CGU  0.17  73  0.8  327
Gln  CAA  0.08  56  0.72  326  CGC*  3.74  1611  1.2  487
CAG*  1.92  1313  1.28  581  CGA  0.14  60  1.19  486
Asn  AAU  0.1  51  0.78  247  CGG*  1.92  828  1.55  631
AAC*  1.9  937  1.22  383  AGA  0.01  3  0.64  262
Lys  AAA  0.18  97  0.75  402  AGG  0.02  9  0.61  250
AAG*  1.82  983  1.25  676  Gly  GGU  0.05  32  0.66  326
Asp  GAU  0.12  190  0.8  489  GGC*  2.71  1910  1.37  677
GAC*  1.88  2959  1.2  740  GGA  0.09  65  1.09  540
Glu  GAA  0.23  391  0.83  717  GGG*  1.15  808  0.88  433

*Codons whose occurrences are significantly higher (P < 0.01) in the extreme left side of axis 1 than the genes present on the extreme right of the first major axis. AA: amino acid; N: number of codon; 1: genes on extreme left of axis 1; 2: genes on extreme right of axis 1.

In conclusion, high level of heterogeneity is seen within the genes of *S. ruber*. The findings reveal that there are large number of genes with high G+C content, and that the G+C content at the third codon position is higher than that of A+T. Accordingly, it is supposed that the usage frequency of codons ending with G or C bases is higher than that ending with A or T bases. In this study, the general association between codon usage bias and base composition suggests that mutational pressure, rather than natural selection, is mainly supported by the highly significant correlation between GC3s and Nc. The (G+C) content is another factor which is found to play an important role in codon usage bias. The overall degree of synonymous codon usage bias is high as suggested by Nc value (mean Nc=37.85 and sd=5.73; less than 40). Apart from the two main factors, gene length also influences the codon usage while aromaticity and hydrophobicity of the encoded proteins play minor role in codon usage bias. The observation that genes expressed at high levels have increased frequencies of those codons that are expected to be translationally optimal is strongly suggestive that these codons are selectively favoured. Identification of the
codon usage patterns of halophilic bacterium may prove useful in the design of oligonucleotide probes, in deducing whether open reading frames are likely to be protein coding, determining the probable level of expression of genes, and indicating the codons to be used in synthetic genes are likely to be expressed in non-salt tolerant bacteria.

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

REFERENCES


