BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(22)

# Synonymous codon usage in chloroplast genome of *Coffea arabica*

**Rahul R Nair[2], Manivasagam B Nandhini[1], Elango Monalisha[1], Kavitha Murugan[2], Thilaga Sethuraman[2], Sangeetha Nagarajan[2], Nayani Surya Prakash Rao[3] & Doss Ganesh[1]***

[1]Department of Plant Biotechnology, School of Biotechnology, Madurai Kamaraj University, Palkalai Nagar 625 021, Madurai, Tamil Nadu, India; [2]Plant Genetic Improvement Laboratory, Department of Biotechnology, SPK Centre for Environmental Sciences, Manonmaniam Sundaranar University, Alwarkurichi 627 412, Tirunelveli District, Tamil Nadu. India; [3]Division of Plant Breeding, Central Coffee Research Institute, Coffee Research Station Post 577 117, Chikmagalur District, Karnataka, India; Doss Ganesh – Email: ganeshdsneha@yahoo.co.in; *Corresponding author

Received October 22, 2012; Accepted October 26, 2012; Published November 13, 2012

**Abstract:**
Synonymous codon usage of 53 protein coding genes in chloroplast genome of *Coffea arabica* was analyzed for the first time to find out the possible factors contributing codon bias. All preferred synonymous codons were found to use A/T ending codons as chloroplast genomes are rich in AT. No difference in preference for preferred codons was observed in any of the two strands, viz., leading and lagging strands. Complex correlations between total base compositions (A, T, G, C, GC) and silent base contents ($A_3$, $T_3$, $G_3$, $C_3$, $GC_3$) revealed that compositional constraints played crucial role in shaping the codon usage pattern of *C. arabica* chloroplast genome. ENC Vs $GC_3$ plot grouped majority of the analyzed genes on or just below the left side of the expected $GC_3$ curve indicating the influence of base compositional constraints in regulating codon usage. But some of the genes lie distantly below the continuous curve confirmed the influence of some other factors on the codon usage across those genes. Influence of compositional constraints was further confirmed by correspondence analysis as axis 1 and 3 had significant correlations with silent base contents. Correlation of ENC with axis 1, 4 and CAI with 1, 2 prognosticated the minor influence of selection in nature but exact separation of highly and lowly expressed genes could not be seen. From the present study, we concluded that mutational pressure combined with weak selection influenced the pattern of synonymous codon usage across the genes in the chloroplast genomes of *C. arabica.*

**Keywords:** *Coffea arabica*, Synonymous codon usage, ENC Vs $GC_3$ plot, Codon adaptation index, Correspondence analysis.

**Background:**
In the universal genetic code, multiple codons differ only at the third position or occasionally in the second position in some of the aminoacids **[1].** Though a number of synonymous codons needed to regulate the translation process, but only particular codons are preferred, leaving the others as less preferred codons. This phenomenon, otherwise known as codon usage bias has been observed in all organism, including prokaryotes, animals and plants **[2-7]**. A number of investigations demonstrated that the synonymous codons usage (SCU) is not at same frequencies either within or between organisms **[8-10]**. The patterns of synonymous codon usage vary significantly among species and also among genes of the same species **[11]**. Though synonymous codon usage biases do not have direct impact on protein sequence, it may influence protein product and cellular processes since codon usage bias has been proved as an important evolutionary force **[12]**. Functional integrity of the genetic code is maintained by synonymous codons by providing a linkage between gene expression and evolution of proteins **[13-15].** The frequency of usage of preferred codons (common codons) may deviates due to mutational biases, caused by chemical decay of nucleotide bases **[16],** non uniform DNA repair and non random replication errors **[12]** or due to natural selection for optimal translation at the stage of synthesis of proteins. Codon bias has significant correlation with mRNA levels as there is a global optimization to reduce the time for the

 

ribosomes to participate in translation of mRNA **[12].** The other biological factors that affect patterns of synonymous codon usage bias are genome size **[17]**, length of the gene **[18],** codon context **[19]**, rate of recombination **[20]** and amino acid composition **[21].**

In plant molecular evolution, chloroplast genomes generate interest among biologist owing to its smaller size, larger copy number and known functions of many genes at molecular level **[22]**. Translational process in the chloroplast genome has been reported to be similar to that of unicellular organisms, indicating that synonymous codon usage may be identical to that of *Escherichia coli* **[23]**. The significance of mutational pressure in shaping the SCU variations in chloroplast genome was already established **[24]**. However, natural selection is also a driving force that frames SCU variations in plants and algal lineages **[25]**. The neutral theory of molecular evolution demonstrated the inverse relation between the rate of molecular evolution and the amount selective forces. Synonymous substitution in protein coding genes is in a slower rate than pseudo genes **[26, 27]**, indicating the influence of selective forces on the rate of synonymous codon evolution. Coffee is a most attractive beverage crop in the world. More than 100 species of *Coffea* are diploids in nature (2n=2x=22) except *Coffee arabica* (2n=4x=44) which is autogamous (self –fertile) and considered as allotetraploid. *C. arabica* is a species for centre of attraction owing to its inherent quality. Unfortunately, this species is highly susceptible to major pests and diseases. The complete nucleotide sequence of *C. arabica* chloroplast genome has been determined **[28]**. However, studies on synonymous codon usage bias in chloroplast genome of *C. arabica* (155,189 bp) have not been reported. Chloroplast genome of *C. arabica* comprises a total of 130 genes (79 protein coding genes, 29 tRNA genes, four ribosomal genes and 18 intron sequences). In this study, synonymous codon usage was analyzed using 53 protein coding genes (PCGs), having more than 100 codons by measuring codon usage indices such as relative synonymous codon usage, effective number of codons (ENC) and codon adaptation index (CAI) and the findings on synonymous codon usage are discussed.

## Methodology:

### *Sequence data of C. arabica chloroplast genes*

A total of 53 protein coding genes in the chloroplast genome of *Coffea arabica* (EF044213) were retrieved from the National Centre for Biotechnology Information (NCBI) and identity of those genes were presented **Table 1 (see supplementary material)**. To avoid sampling errors, PCGs contain more than 100 codons were chosen for analysis.

### *Relative Synonymous Codon Usage (RSCU)*

To analyze the characteristics of variations in synonymous codon usage by neglecting the influence of amino acid composition, the relative synonymous codon usage values of all sequences were determined according to equation **(described in supplementary material) [29].**

### *Effective number of codons (ENC)*

This index is used to measure the extent of synonymous codon usage bias. The ENC values would be 20 when only one codon is used for each amino acid, In contrast, when codons are used randomly, the ENC values is 61. If the calculated ENC is greater than 61 due to more even distribution of codon usage than expected, it is adjusted to 61. Thus, the expected ENCs are calculated by following equation **[30],**

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

Where s = GC3 (GC content at the third codon position)

### *Identification of optimal codons*

Optimal codons occur most frequently only in highly expressed genes. Difference in RSCU of a given codon between putative high and low expression data set was calculated and tested the significance ($p \leq 0.05$) using one-way analysis of variance (ANOVA). Codons that occur significantly at high frequency in highly expressed genes were regarded as putative optimal codons.

### *Codon adaptation index*

Codon adaptation index (CAI) is used to measure the extent of bias towards preferred codons in a gene by defining the translationally optimal codons that are mostly represented in a reference set of highly expressed genes **[31]**. It takes value from zero to one and a higher value reveals a stronger codon bias with high expression level. In this study, the ribosomal protein coding genes have been used as reference for estimating CAI values on the basis of equation **(described in supplementary material).**

### *Sequence analysis*

Nucleotide contents were calculated using a software viz., Dambe version 5.2.65 **[32]** and CAI values were estimated by CAI calculator 2 **[33]**.

## Statistical methods

### *Correspondence analysis*

Correspondence analysis (COA) is extensively used for analyzing multidimensional data **[34].** COA was performed to analyze RSCU values to explore the different features of synonymous codon usage (SCU) patterns across the 53 protein coding genes in the chloroplast genome of *C. arabica.*

### *Correlation analysis*

Correlation analysis was used to demonstrate the relationship between base compositions and synonymous codon usage patterns. This analysis was implemented based on the Pearson correlation analysis way. All statistical processes were carried out with software Past version 2.12 **[35]**.

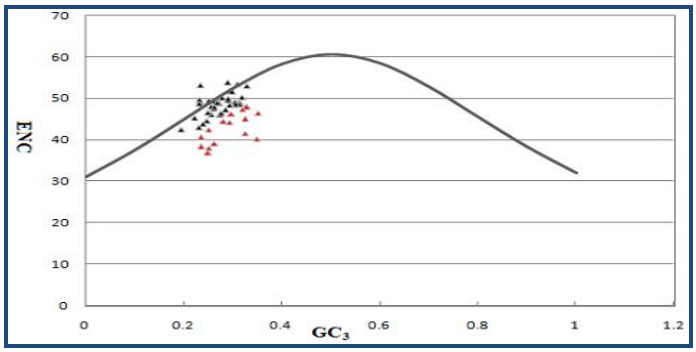


**Figure 1:** ENC Vs GC3 plot showing the grouping of majority of genes on or just below the expected GC3 curve. Red spots indicate genes that are independent of GC compositional constraints as they lie considerably below the continuous curve.

**Discussion:**

***Base compositional analysis of ORFs of 53 protein coding genes***

Total and silent base compositions were identified **Table 2 (see supplementary material)**. Intricate correlations were found between A, T, G, C, GC and $A_3$, $T_3$, $G_3$, $C_3$, $GC_3$ contents in the 53 ORFs of chloroplast genome of *C. arabica* **Table 3 (see supplementary material)**. Interestingly $GC_3$ has significant correlations with A, G, C, and GC contents, prognosticating that GC contents may have high influence, balancing between mutational pressure and natural selection. We observed that A was significantly correlated with $A_3$, $T_3$, $C_3$ and $GC_3$ contents and T had higher correlations only with $A_3$ and $T_3$. Base composition analysis of ORFs revealed that G, C content has significant correlations with $T_3$, $G_3$, $GC_3$, and $G_3$, $C_3$, $GC_3$ respectively. This implies that compositional constraints may have direct influence in the evolution of synonymous codon usage in the chloroplast genome of *C. arabica.*

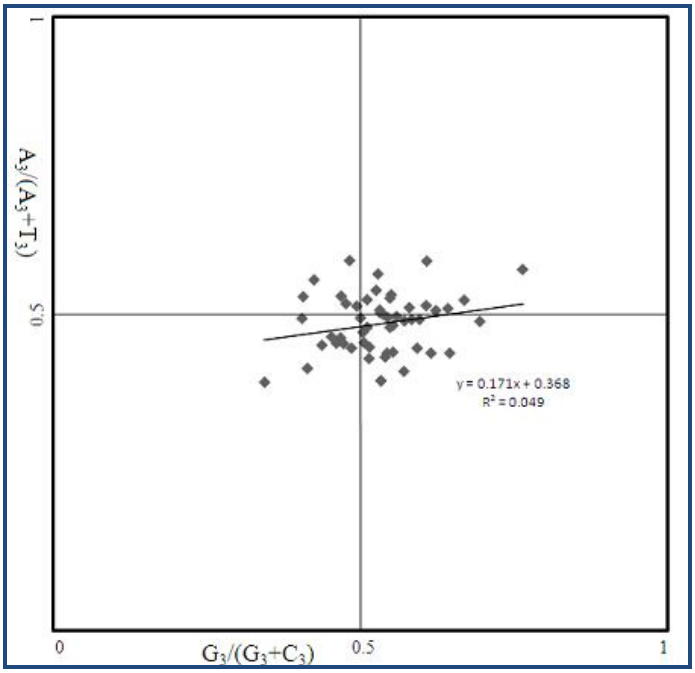


**Figure 2:** PR2 bias plot (A3/ (A3+T3) Vs G3/ (G3+C3)). Average position is y = 0.171x+0.368.

***Characteristics of relative synonymous codon usage***

Overall and strand specific codon usage patterns of 53 PCGs were analyzed **Table 4 (see supplementary material).** Strand asymmetry was reported in the chloroplast genome of *Euglina gracillis* with regard to base composition **[36]**. Strand specific analysis was carried out to examine the differences in preference for codons in leading and lagging strand and found that there was no difference in the selection of preferred codons for coding all the 18 degenerate amino acid in both the strand. Arg, Leu and Ser have six fold degeneracy. A ending codons TTA and AGA were preferred to code Leu and Arg respectively whereas TCT was most frequently used to code Ser. The amino acids Ala, Gly, Pro, Thr and Val possess four-fold degeneracy. We found that T ending codons were often used for coding Ala (TCT), Pro (CCT), and Thr (ACT) but A ending codons were preferred for coding Gly (GGA) and Val (GTA). In two fold degenerate family, A ending codons were used most frequently to code Glu (GAA), Gln (CAA) and Lys (AAA) whereas T ending codons were preferred to code His (CAT), Phe (TTT), Tyr (TAT), Asn (AAT), Asp (GAT) and Cys (TGT). Three fold degenerate Ile was preferably coded by ATT. All the rare of non preferred codons were observed to be ending in C or G. Putative optimal codons were identified for amino acids Asp (GAT), His (CAT), Ile (ATA), and Arg (AGA). No optimal codons were identified for any of the four fold degenerate amino acid family. The values of ENCs among 53 PCGs were found to be varying from 36.79 to 53.86 with mean of 46.87 and S.D of 4.02 in the chloroplast genome of *C. arabica* and the overall $GC_3$ values varied from 19.60% to 35.30% with a mean and S.D of 27.82% and 3.52 respectively, ensuring the heterogeneity of codon usage **(Table 2)**.

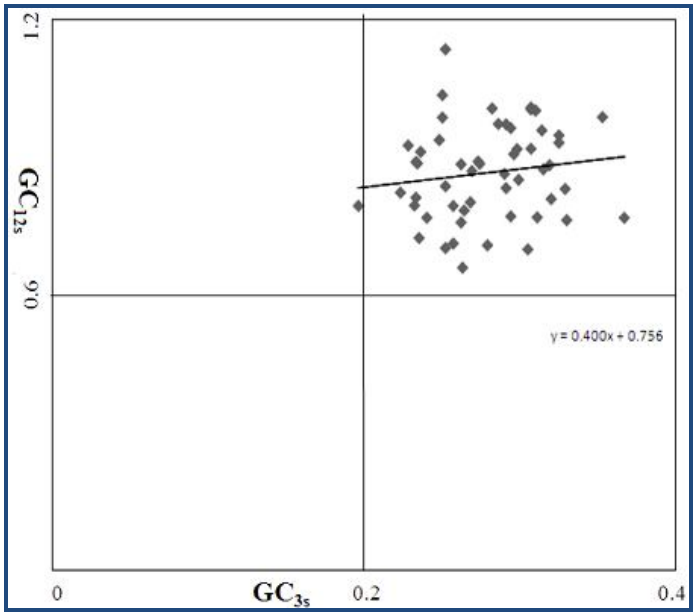


**Figure 3:** Neutrality plot (GC12 Vs GC3). Average position is y= 0.400+0.756. r = 0.142 (p = 0.305).

***ENC Vs $GC_3$ plot***

The ENC plot is developed for determining the variations in SCU across a number of genes by exhibiting intraspecific and interspecific variations. It has been considered as an alternative to complex multivariate statistical methods to analyze SCU patterns in a genome. This plot effectively demonstrates SCU variation when the genome under study has significantly different GC content from 0.50 **[30]**. ENCs and $GC_3$ were estimated and plotted for 53 PCGs, following the null hypothesis that $GC_3$ compositional constraints alone influences expected SCU patterns. If a gene is influenced by GC compositional constraints, it would lie on or below the expected curve. If translational selection acts on a gene, it lies distantly below the expected curve. Accumulation of majority of genes on or below the left hand side of the expected $GC_3$ based on null hypothesis suggested that SCU variations across majority of PCGs in *C. arabica* chloroplast genome were influenced by $GC_3$ compositional constraints **(Figure 1).** To confirm this further, Parity rule 2 (PR2) plot **[37]** was analyzed **(Figure 2).** If $GC_3$ mutational pressure acts on the genes, G and C (A and T) nucleotides should be used proportionally. To analyze whether codon preferences were restricted in strongly biased ORFs, relationship between synonymous G, C and A, T contents were analyzed. It was observed that G, C, and A, T contents were used proportionally **(Figure 2)** and this confirmed the influence

of $GC_3$ biased mutational pressure in framing codon usage biases. Neutrality plot **[38] (Figure 3)** revealed that some amount of selection may act on PCGs in the chloroplast genome of *Coffea arabica* as there was no correlation between $GC_{12}$ and $GC_3$ **[39]**. However, some of the genes, viz., *rps*12, *ndh*F, *ndh*E, *ndh*B, *atp*F, *pet*B, *pet*D, *psa*B, *psb*A, *psb*C, *psb*D, *rpl*16, *rps*8 and *rps*14 were grouped distantly below the expected $GC_3$ curve indicates the interaction of some other factors that are independent of base compositional constraints.

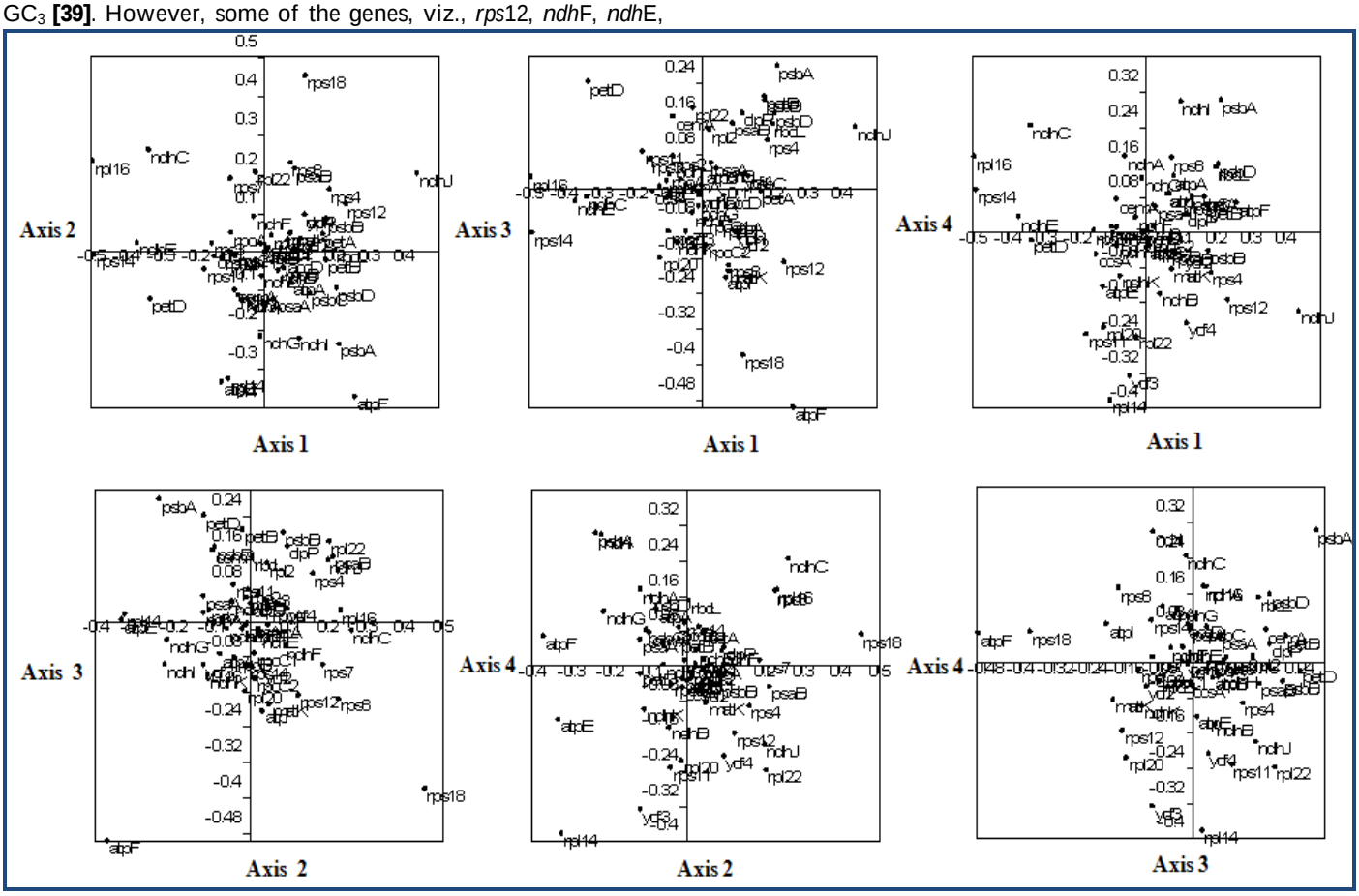


**Figure 4:** Correspondence analysis plots of RSCU values of 53 ORF's in the *C. arabica* chloroplast genome. Distribution of these ORFs was found to be mainly due to base compositions.

***Correspondence analysis (COA)***

Frequently used multivariate statistical analysis, viz., correspondence analysis was used for determining the SCU variation across 53 PCGs in the *C. arabica* chloroplast genome. To avoid effect of aminoacid composition, RSCU values were used instead of codon counts. In this study, ORF of each gene was represented as 59 dimensional vectors (excluding three stop codons, Met and Trp) and each dimension denotes RSCU value of one codon. The first, second, third and fourth axes account for 10.94%, 8.09%, 7.35% and 6.89% respectively **(Figure 4).** To unveil the factors that are responsible for the distribution of genes in the COA plot, ordination of 53 PCGs of axis 1 to 4 were analyzed for possible correlation with silent base composition and various indices of codon usage **Table 5 (see supplementary material).** Axis 1 was significantly correlated with $A_3$ ($r = -0.481$, $p < 0.01$) and with $C_3$ ($r = 0.542$, $p < 0.01$). Axis 3 was in correlation with $G_3$ ($r = -0.276$, $p < 0.01$) and $C_3$ ($r = 0.395$, $p < 0.01$). These correlations indicated the possible influence of nucleotide compositional constraints in framing codon usage. ENC, an index measuring the level of gene expression was found to be in correlation with axis 1 ($r = 0.364$, $p < 0.01$) and with axis 2 ($r = -0.377$, $p < 0.01$). Another index for measuring gene expression, viz., CAI was significantly correlated with axis 1 ($r = 0.498$, $p < 0.01$) and with axis 2 ($r = 0.381$, $p < 0.01$). Hence, it could be predicted that gene expression levels also slightly influences (weak selection) codon usage pattern in the chloroplast genome of *C. arabica*.

An equilibrium between positive and negative directional mutational pressure on GC base pairs results in the base composition of a genome **[40].** The non random usage of synonymous codons (codon bias) was detected in all coding sequences of prokaryotes and eukaryotes as a result of the interaction between two significant evolutionary forces such as directional mutational pressure and selection in nature for translational optimization **[2, 31, 6, 7]**. Codon bias of genes in the chloroplast genomes has been reported to be towards A and T ending codons due to the compositional bias towards AT rich content **[41, 42, 24].** Codon bias is reported to be relatively weaker in angiosperms **[25]**. A similar finding was also noticed from the ENC values of PCGs in the *C. arabica* chloroplast genome as ENCs of all the genes are greater than 35, indicating a weaker bias. Correlation analysis between total nucleotide contents and silent base contents revealed that base compositional constraints greatly influence in framing the SCU pattern in the PCGs of *C. arabica* chloroplast genome. Nucleotide composition is considered to be the most important factor that influences the SCU variations in chloroplast genomes

**[43].** Since all chloroplast genomes have high AT content, AT biased mutational pressure is believed to be the factor responsible for codon usage bias. But in the *psb*A gene of higher angiosperms, the codon usage is directly linked to tRNA population for translational optimization, indicating selection acts on this gene **[44].** Since selection influences codon usage of *psb*A, lower rate of synonymous substitutions was reported in *psb*A gene **[45].**

Identified putative optimal codons were found to be ending in A/T and found consistent with previous findings **[46].** In our study, ENC plot of selected chloroplast genes in *C. arabica* unravels the possibility that some amount of selection may act on *pet*B, *pet*D, *psa*B, *psb*A, *psb*C, *psb*D, *rpl*16, *rps*8 and *rps*14 as these genes lie considerably below the expected GC curve. However almost all the other genes lie on or just below the continuous curve, revealing the influence of compositional constraints in regulating SCU variation in *C. arabica* chloroplast genome, further confirmed using PR2 bias plot. COA analysis to find out the factors responsible for codon usage discrepancies shown that compositional constraints rather than selection plays crucial role in shaping the SCU variation across genes in *C. arabica* chloroplast genome as there was a grouping of A/T ending and C/G ending codons along the axis 1. Nevertheless, influence of selection on codon usage of some of the genes cannot be nullified because factors responsible for selective constraints are regarded as dynamic processes **[44].** It was reported that selection against mutational pressure may narrow the $GC_3$ distribution **[39].** Similar to previous finding **[47],** we observed a narrow distribution of $GC_3$ contents (19.36%-35.30%) in the *Coffea arabica* chloroplast genome and no correlation was observed between $GC_{12}$ and $GC_3$. This result demonstrates that some amount of selection may act in the chloroplast genome of *Coffea arabica*. Correlations of axis 1 & 2 with CAI, and axis 1 & 4 with ENC point out the influence of gene expression levels in framing the codon usage patterns. However, no exact separation is observed in the COA plot between highly and lowly expressed genes based on either ENC or CAI. In contrast to the previous findings that revealed significant correlation between length of CDS, hydropathy, aromaticity and codon bias **[48, 49, 22],** but our study could not find any significant correlation between these parameters and codon bias. Thus it can be concluded that the factor responsible for SCU variation across genes in the *C. arabica* chloroplast genome may possibly due to mutational pressure combined with weak selection. Moreover, information about the rare and preferred codons can be effectively used for enhancing expression of genes by optimizing synonymous codons. To the best of our knowledge, this is the very first report describing the codon usage bias in the genus *Coffea*.

**Acknowledgement:**
The authors acknowledge Central Coffee Research Institute, Chikmagalur, Karnataka, India for providing technical support in writing the manuscript. This study has been financially supported by the University Grants Commission, New Delhi, India under the major research project 'Establishment of genetic identity for Indian coffee germplasm using chloroplast genome sequences' F. No. 41-583/2012 (SR) during the year 2012 – 2015. Part of this work is supported by Department of Science and Technology under Promotion of University Research and Scientific Excellence (DST-PURSE).

**References:**

**[1]** Ermolaeva MD, *Curr Issues Mol Biol.* 2001 **3**: 91 [PMID: 11719972]
**[2]** Bonitz SG *et al. J Biol Chem.* 1980 **25**: 255 [PMID: 6254986]
**[3]** Ikemura T, *J Mol Biol*. 1982 **158**: 573 [PMID: 6750137]
**[4]** Sharp PM & Cowe E, *Yeast.* 1991 **7:** 657 [PMID: 1776357]
**[5]** Moriyama EN & Powell JR, *J Mol Evol.* 1997 **45**: 378 [PMID: 9321417]
**[6]** Akashi H, *Curr Opin Genet Dev.* 2001 **11**: 660 [PMID: 11682310]
**[7]** Duret L, *Curr Opin Genet Dev.* 2002 **12**: 640 [PMID: 12433576]
**[8]** Grantham R *et al. Nucleic Acids Res.* 1980 **8**: 49 [PMID: 6986610]
**[9]** Martin CE & Scheinbach S, *Biotechnol Adv.* 1989 **7**: 155 [PMID: 14545930]
**[10]** Lloyd AT & Sharp P, *Nucleic Acids Res.* 1992 **20:** 5289 [PMID: 1437548]
**[11]** Grocock RJ & Sharp PM, *Int J Parasitol.* 2001 **31**: 402 [PMID: 11306119]
**[12]** Carlini DB *et al. Genetics.* 2001 **159:** 623 [PMID: 11606539]
**[13]** Biro JC, *Theor Bio Med Mod.* 2008 **5**: 1
**[14]** Starmer WT & Sullivan DT, *Mol Biol Evol.* 1989 **6**: 546 [PMID: 2796728]
**[15]** Wall DP & Herback JT, *J Mol Evol*. 2003 **56**: 673 [PMID: 12911031]
**[16]** Kaufmann WK & Paules RS, *FASEB J.* 1996 **10:** 238 [PMID: 8641557]
**[17]** Reis DM *et al. Nucleic Acids Res.* 2004 **32**: 5036 [PMID: 15448185]
**[18]** Duret L & Mouchiroud D, *Proc Natl Acad Sci USA.* 1999 **96**: 4482 [PMID: 10200288]
**[19]** Irwin B *et al. J Biol Chem.* 1995 **270**: 22801 [PMID: 7559409]
**[20]** Zhou T *et al. Conf Proc IEEE Eng Med Bio Soc*. 2005 **5**: 4787
**[21]** D'Onofrio G *et al. Gene.* 1999 **238**: 3 [PMID: 10570978]
**[22]** Xu C *et al. Evol Bioinform.* 2011 **7**: 271 [PMID: 22253533]
**[23]** Sugiura M, *Plant Mol Biol.* 1992 **19**: 149 [PMID: 1600166]
**[24]** Morton BR, *J Mol Evol.* 2003 **56**: 616 [PMID: 12698298]
**[25]** Morton BR, *J Mol Evol.* 1998 **46**: 449 [PMID: 9541540]
**[26]** Li WH, *Nature.* 1981 **292**: 237 [PMID: 7254315]
**[27]** Miyata T & Hayashita H, *Proc Natl Acad Sci USA*. 1981 **78**: 5739 [PMID: 6795634]
**[28]** Samson N *et al. Plant Biotechnol J.* 2007 **5**: 339 [PMID: 17309688]
**[29]** Sharp PM *et al. Nucleic Acids Res.* 1986 **14**: 5125 [PMID: 3526280]
**[30]** Wright F, *Gene.* 1990 **87**: 23 [PMID: 2110097]
**[31]** Sharp PM & Li WH, *Nucleic Acids Res.* 1987 **15**: 1281 [PMID: 3547335]
**[32]** Xia X & Xie Z, *J Hered.* 2001 **92**: 371 [PMID: 11535656]
**[33]** Wu G *et al. Microbiol.* 2005 **151**: 2175 [PMID: 16000708]
**[34]** Perriere G & Thioulouse J, *Nucleic Acids Res.* 2002 **30**: 4548 [PMID: 12384602]
**[35]** Hammer Q *et al. Paleontologia Electronica.* 2001 **4**: 1
**[36]** Morton BR, *Proc Natl Acad Sci USA*. 1999 **96**: 5123 [PMID: 10220429]
**[37]** Sueoka N, *Gene*. 1999 **238**: 53 [PMID: 10570983]
**[38]** Sueoka N, *Pro Nat Acad Sci USA.* 1988 **85**: 2653 [PMID: 3357886]
**[39]** Kawabe A & Miyashita NT, *Genes Genet Syst.* 2003 **78**: 343 [PMID: 14676425]

 

**[40]** Sueoka N, *Proc Natl Acad Sci USA.* 1962 **48**: 582 [PMID: 13918161]
**[41]** Wolfe KH & Sharp PM, *Gene.* 1988 **66**: 215 [PMID: 3169573]
**[42]** Wolfe KH *et al. Proc Natl Acad Sci USA.* 1992 **89**: 10648 [PMID: 1332054]
**[43]** Meng Z *et al. J Forestry Res*. 2008 **9**: 293
**[44]** Morton BR, *J Mol Evol.* 1993 **37**: 273 [PMID: 8230251]
**[45]** Morton BR, *Mol Biol Evol.* 1997 **14**: 412 [PMID: 9100371]
**[46]** Sablok G *et al. Mol Biotechnol.* 2011 **49**: 116 [PMID: 21308422]
**[47]** Liu Q & Xue Q, *J Genet.* 2005 **84**: 55 [PMID: 15876584]
**[48]** Duret L & Mouchiroud D, *Mol Biol Evol.* 2000 **17**: 68 [PMID: 10666707]
**[49]** Ingvarsson PK, *Mol Bio Evol*. 2007 **24**: 836 [PMID: 17204548]

**Edited by P Kangueane**
**Citation: Nair *et al*.** Bioinformation 8(22): 1096-1104 (2012)

BIOINFORMATION

open access

## Supplementary material:

**Methodology:**

***Relative Synonymous Codon Usage (RSCU)***

$$RSCU = \frac{\text{Observed frequency of a codon}}{\text{Expected frequency provided all synonymous codons for those amino acids used Equally}}$$

***Codon adaptation index***

$$CAI = \exp\ 1/L \sum_{n=1}^{L} \ln w_n$$

Where w = relative adaptness of $n^{th}$ codon, L = Number of codons

**Table 1:** Details of genes analyzed in the present study

| Genes | ID | Length | Genes | ID | Length |
|---|---|---|---|---|---|
| *acc*D | 4421772 | 1536 | *psb*A | 4421839 | 1062 |
| *atp*A | 4421812 | 1524 | *psb*B | 4421754 | 1527 |
| *atp*B | 4421770 | 1497 | *psb*C | 4421869 | 1422 |
| *atp*E | 4421769 | 399 | *psb*D | 4421868 | 1062 |
| *atp*F | 4421813 | 573 | *rbc*L | 4421771 | 1446 |
| *atp*I | 4421787 | 735 | *rpl*2 | 4421821 | 831 |
| *ccs*A | 4421850 | 969 | *rpl*14 | 4421816 | 369 |
| *cem*A | 4421775 | 690 | *rpl*16 | 4421817 | 408 |
| *clp*P | 4421753 | 591 | *rpl*20 | 4421752 | 399 |
| *mat*K | 4421803 | 1518 | *rpl*22 | 4421819 | 468 |
| *ndh*A | 4421856 | 1092 | *rpo*A | 4421760 | 1008 |
| *ndh*B | 4421826 | 1623 | *rpo*B | 4421747 | 3201 |
| *ndh*C | 4421834 | 363 | *rpoC*1 | 4421790 | 2052 |
| *ndh*D | 4421851 | 1503 | *rpoC*2 | 4421789 | 4176 |
| *ndh*E | 4421853 | 306 | *rps*2 | 4421788 | 711 |
| *ndh*F | 4421847 | 2220 | *rps*3 | 4421818 | 657 |
| *ndh*G | 4421854 | 531 | *rps*4 | 4421827 | 606 |
| *ndh*H | 4421857 | 1188 | *rps*7 | 4421879 | 468 |
| *ndh*I | 4421855 | 504 | *rps*8 | 4421815 | 405 |
| *ndh*J | 4421832 | 477 | *rps*11 | 4421761 | 417 |
| *ndh*K | 4421833 | 690 | *rps*12 | 4421880 | 372 |
| *pet*A | 4421776 | 960 | *rps*14 | 4421874 | 303 |
| *pet*B | 4421758 | 648 | *rps*18 | 4421751 | 306 |
| *pet*D | 4421759 | 483 | *ycf*1 | 4421859 | 5625 |
| *psa*A | 4421876 | 2253 | *ycf*2 | 4421824 | 6846 |
| *psa*B | 4421875 | 2205 | *ycf*3 | 4421877 | 507 |
| *ycf*4 | 4421774 | 555 | --- | --- | --- |

**Table 2:** Identified nucleotide contents in the coding sequences of 53 protein coding genes of chloroplast genome of *Coffea arabica*

| Genes | A | T | G | C | $A_3$ | $T_3$ | $G_3$ | $C_3$ | $GC_3$ | ENC | CAI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *acc*D | 0.326 | 0.314 | 0.216 | 0.144 | 0.252 | 0.438 | 0.180 | 0.131 | 0.311 | 48.85 | 0.711 |
| *Atp*A | 0.296 | 0.283 | 0.227 | 0.194 | 0.325 | 0.384 | 0.163 | 0.128 | 0.291 | 49.50 | 0.704 |
| *atp*B | 0.303 | 0.277 | 0.229 | 0.192 | 0.359 | 0.355 | 0.138 | 0.148 | 0.286 | 47.25 | 0.710 |
| *atp*E | 0.333 | 0.271 | 0.231 | 0.165 | 0.353 | 0.398 | 0.150 | 0.098 | 0.248 | 44.56 | 0.704 |
| *atp*F | 0.335 | 0.300 | 0.223 | 0.141 | 0.330 | 0.340 | 0.215 | 0.115 | 0.330 | 47.91 | 0.663 |
| *atp*I | 0.259 | 0.355 | 0.200 | 0.186 | 0.306 | 0.420 | 0.131 | 0.143 | 0.274 | 46.06 | 0.714 |
| *ccs*A | 0.301 | 0.370 | 0.173 | 0.156 | 0.344 | 0.378 | 0.158 | 0.121 | 0.279 | 50.12 | 0.705 |
| *cem*A | 0.301 | 0.364 | 0.164 | 0.171 | 0.291 | 0.404 | 0.135 | 0.170 | 0.305 | 49.13 | 0.667 |
| *clp*P | 0.289 | 0.291 | 0.235 | 0.184 | 0.360 | 0.345 | 0.152 | 0.142 | 0.294 | 51.81 | 0.699 |
| *mat*K | 0.310 | 0.366 | 0.154 | 0.169 | 0.308 | 0.435 | 0.134 | 0.123 | 0.257 | 48.06 | 0.727 |
| *ndh*A | 0.279 | 0.372 | 0.179 | 0.170 | 0.368 | 0.409 | 0.113 | 0.110 | 0.223 | 45.29 | 0.730 |
| *ndh*B | 0.278 | 0.346 | 0.177 | 0.199 | 0.322 | 0.359 | 0.131 | 0.189 | 0.320 | 47.31 | 0.651 |
| *ndh*C | 0.242 | 0.422 | 0.204 | 0.132 | 0.306 | 0.455 | 0.157 | 0.083 | 0.240 | 43.84 | 0.750 |
| *ndh*D | 0.270 | 0.376 | 0.173 | 0.182 | 0.321 | 0.385 | 0.152 | 0.142 | 0.294 | 48.37 | 0.694 |
| *ndh*E | 0.284 | 0.395 | 0.173 | 0.147 | 0.265 | 0.500 | 0.127 | 0.108 | 0.235 | 40.57 | 0.745 |
| *ndh*F | 0.284 | 0.398 | 0.168 | 0.151 | 0.301 | 0.447 | 0.157 | 0.095 | 0.252 | 42.49 | 0.708 |
| *ndh*G | 0.254 | 0.403 | 0.179 | 0.164 | 0.311 | 0.458 | 0.130 | 0.102 | 0.232 | 42.97 | 0.708 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *ndh*H | 0.310 | 0.307 | 0.222 | 0.161 | 0.361 | 0.376 | 0.141 | 0.121 | 0.262 | 47.57 | 0.745 |
| *ndh*I | 0.314 | 0.329 | 0.179 | 0.179 | 0.310 | 0.423 | 0.125 | 0.143 | 0.268 | 49.24 | 0.705 |
| *ndh*J | 0.279 | 0.325 | 0.233 | 0.164 | 0.283 | 0.403 | 0.189 | 0.126 | 0.315 | 48.92 | 0.709 |
| *ndh*K | 0.312 | 0.315 | 0.183 | 0.191 | 0.348 | 0.417 | 0.130 | 0.104 | 0.234 | 49.17 | 0.738 |
| *pet*A | 0.313 | 0.287 | 0.213 | 0.189 | 0.303 | 0.378 | 0.163 | 0.156 | 0.319 | 50.26 | 0.686 |
| *pet*B | 0.244 | 0.355 | 0.216 | 0.185 | 0.278 | 0.426 | 0.162 | 0.134 | 0.294 | 44.18 | 0.713 |
| *pet*D | 0.273 | 0.344 | 0.203 | 0.180 | 0.373 | 0.391 | 0.149 | 0.087 | 0.236 | 38.30 | 0.791 |
| *psa*A | 0.256 | 0.320 | 0.224 | 0.201 | 0.290 | 0.395 | 0.161 | 0.153 | 0.314 | 48.59 | 0.685 |
| *psa*B | 0.254 | 0.341 | 0.214 | 0.192 | 0.275 | 0.427 | 0.155 | 0.143 | 0.298 | 46.23 | 0.716 |
| *psb*A | 0.239 | 0.342 | 0.214 | 0.205 | 0.232 | 0.444 | 0.113 | 0.212 | 0.325 | 41.49 | 0.668 |
| *psb*B | 0.240 | 0.323 | 0.263 | 0.175 | 0.279 | 0.415 | 0.169 | 0.138 | 0.307 | 48.40 | 0.708 |
| *psb*C | 0.235 | 0.318 | 0.250 | 0.198 | 0.281 | 0.367 | 0.167 | 0.186 | 0.353 | 46.48 | 0.683 |
| *psb*D | 0.220 | 0.355 | 0.219 | 0.205 | 0.251 | 0.424 | 0.136 | 0.189 | 0.325 | 44.98 | 0.688 |
| *rbc*L | 0.279 | 0.282 | 0.249 | 0.190 | 0.286 | 0.407 | 0.151 | 0.156 | 0.307 | 48.68 | 0.682 |
| *rpl*2 | 0.329 | 0.234 | 0.253 | 0.185 | 0.343 | 0.347 | 0.155 | 0.155 | 0.310 | 53.45 | 0.694 |
| *rpl*14 | 0.331 | 0.268 | 0.225 | 0.176 | 0.374 | 0.350 | 0.114 | 0.163 | 0.277 | 46.43 | 0.735 |
| *rpl*16 | 0.333 | 0.238 | 0.240 | 0.189 | 0.441 | 0.309 | 0.154 | 0.096 | 0.250 | 36.79 | 0.775 |
| *rpl*20 | 0.363 | 0.288 | 0.201 | 0.148 | 0.391 | 0.376 | 0.158 | 0.075 | 0.233 | 49.68 | 0.721 |
| *rpl*22 | 0.385 | 0.265 | 0.184 | 0.167 | 0.385 | 0.359 | 0.122 | 0.135 | 0.257 | 46.09 | 0.751 |
| *rpo*A | 0.364 | 0.296 | 0.182 | 0.159 | 0.384 | 0.354 | 0.146 | 0.116 | 0.262 | 49.33 | 0.736 |
| *rpo*B | 0.319 | 0.292 | 0.222 | 0.167 | 0.363 | 0.365 | 0.168 | 0.105 | 0.273 | 48.79 | 0.713 |
| *rpo*C1 | 0.317 | 0.303 | 0.212 | 0.169 | 0.342 | 0.389 | 0.156 | 0.113 | 0.269 | 48.94 | 0.709 |
| *rpo*C2 | 0.327 | 0.298 | 0.199 | 0.176 | 0.343 | 0.367 | 0.157 | 0.134 | 0.291 | 49.91 | 0.708 |
| *rps*2 | 0.331 | 0.286 | 0.219 | 0.165 | 0.329 | 0.371 | 0.177 | 0.122 | 0.299 | 51.65 | 0.719 |
| *rps*3 | 0.403 | 0.266 | 0.181 | 0.149 | 0.452 | 0.352 | 0.105 | 0.091 | 0.196 | 42.48 | 0.791 |
| *rps*4 | 0.342 | 0.284 | 0.187 | 0.188 | 0.406 | 0.361 | 0.124 | 0.109 | 0.233 | 48.74 | 0.726 |
| *rps*7 | 0.353 | 0.235 | 0.214 | 0.199 | 0.442 | 0.308 | 0.122 | 0.128 | 0.250 | 46.54 | 0.742 |
| *rps*8 | 0.353 | 0.284 | 0.217 | 0.146 | 0.370 | 0.378 | 0.148 | 0.104 | 0.252 | 37.97 | 0.768 |
| *rps*11 | 0.290 | 0.247 | 0.264 | 0.199 | 0.360 | 0.388 | 0.137 | 0.115 | 0.252 | 49.46 | 0.747 |
| *rps*12 | 0.331 | 0.239 | 0.218 | 0.212 | 0.395 | 0.323 | 0.121 | 0.161 | 0.282 | 44.38 | 0.734 |
| *rps*14 | 0.347 | 0.244 | 0.238 | 0.172 | 0.396 | 0.297 | 0.238 | 0.069 | 0.350 | 40.24 | 0.702 |
| *rps*18 | 0.346 | 0.304 | 0.186 | 0.163 | 0.343 | 0.392 | 0.186 | 0.078 | 0.264 | 39.07 | 0.769 |
| *ycf*1 | 0.393 | 0.299 | 0.161 | 0.147 | 0.378 | 0.358 | 0.146 | 0.117 | 0.263 | 47.93 | 0.724 |
| *ycf*2 | 0.311 | 0.311 | 0.193 | 0.186 | 0.265 | 0.500 | 0.127 | 0.108 | 0.235 | 53.18 | 0.646 |
| *ycf*3 | 0.333 | 0.282 | 0.207 | 0.178 | 0.337 | 0.373 | 0.160 | 0.130 | 0.290 | 53.86 | 0.691 |
| *ycf*4 | 0.263 | 0.350 | 0.218 | 0.169 | 0.292 | 0.378 | 0.151 | 0.178 | 0.329 | 53.00 | 0.667 |

**Table 3:** Pearson correlation analysis between A, T, G, C, GC contents and $A_3$, $T_3$, $G_3$, $C_3$, $GC_3$ contents in ORFs of 53 chloroplast protein coding genes of *Coffea arabica*

| | $A_3$ | $T_3$ | $G_3$ | $C_3$ | $GC_3$ |
|---|---|---|---|---|---|
| **A** | 0.727** | -0.528** | 0.000 | -0.447** | -0.371* |
| **T** | -0.608** | 0.728** | -0.106 | 0.062 | -0.054 |
| **G** | -0.034 | -0.270* | -0.307* | 0.182 | 0.451** |
| **C** | -0.095 | -0.160 | -0.270* | 0.590** | 0.329* |
| **GC** | -0.070 | -0.299* | 0.132** | 0.425** | 0.486** |

** Figures are significant at $p < 0.01$

* Figures are significant at $p < 0.05$

**Table 4**: Overall and strand specific relative synonymous codon usage data of protein coding genes in the chloroplast genome of *Coffea arabica*

| AA | Codon | N (RSCU) | | | AA | Codon | N (RSCU) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **All** | **(+) strand** | **(-) strand** | | | **All** | **(+) strand** | **(-) strand** |
| Phe | UUU | **815 (1.35)** | **274 (1.22)** | **635 (1.33)** | Tyr | UAU | **649 (1.62)** | **196 (1.54)** | **542 (1.64)** |
| | UUC | 395 (0.65) | 177 (0.78) | 317 (0.67) | | UAC | 153 (0.38) | 58 (0.46) | 120 (0.36) |
| Leu | UUA | **723 (1.92)** | **198 (1.57)** | **594 (1.92)** | TER | UAA | **27 (1.53)** | **6 (1.12)** | **23 (1.68)** |
| | UUG | 461 (1.23) | 178 (1.41) | 366 (1.18) | | UAG | 13 (0.74) | 5 (0.94) | 10 (0.73) |
| | CUU | 491 (1.31) | 168 (1.33) | 406 (1.31) | His | CAU | **375 (1.47)** | **121 (1.48)** | **313 (1.48)** |
| | CUC | 133 (0.35) | 53 (0.42) | 116 (0.37) | | CAC | 134 (0.53) | 43 (0.52) | 110 (0.52) |
| | CUA | 302 (0.80) | 111 (0.88) | 251 (0.81) | Gln | CAA | **609 (1.54)** | **172 (1.49)** | **521 (1.54)** |
| | CUG | 147 (0.39) | 51 (0.40) | 126 (0.41) | | CAG | 182 (0.46) | 59 (0.51) | 155 (0.46) |
| Ile | AUU | **929 (1.51)** | **256 (1.40)** | **793 (1.52)** | Asn | AAU | **815 (1.58)** | **262 (1.53)** | **689 (1.57)** |
| | AUC | 327 (0.53) | 123 (0.67) | 266 (0.51) | | AAC | 219 (0.42) | 80 (0.47) | 187 (0.43) |
| | AUA | 591 (0.96) | 169 (0.93) | 510 (0.98) | Lys | AAA | **860 (1.52)** | **246 (1.39)** | **742 (1.53)** |
| Met | AUG | 487 (1.00) | 158 (1.00) | 403 (1.00) | | AAG | 269 (0.48) | 107 0.61) | 226 0.47) |
| Val | GUU | 445 (1.51) | 142 (1.46) | 345 (1.51) | Asp | GAU | **709 (1.63)** | **255 (1.63)** | **588 (1.64)** |
| | GUC | 130 (0.44) | 52 (0.53) | 97 (0.42) | | GAC | 159 (0.37) | 58 (0.37) | 127 0.36) |
| | GUA | **460 (1.57)** | **146 (1.50)** | **359 (1.57)** | Glu | GAA | **877 (1.54)** | **267 (1.42)** | **717 (1.54)** |
| | GUG | 140 (0.48) | 49 (0.50) | 115 (0.50) | | GAG | 261 (0.46) | 109 (0.58) | 213 (0.46) |
| Ser | UCU | **473 (1.80)** | **161 (1.66)** | **401 (1.83)** | Cys | UGU | **169 (1.48)** | **55 (1.28)** | **143 (1.53)** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | UCC | 236 (0.90) | 104 (1.07) | 194 (0.89) | | UGC | 60 (0.52) | 31 (0.72) | 44 (0.47) |
| | UCA | 310 (1.18) | 104 (1.07) | 269 (1.23) | TER | UGA | 13 (0.74) | 5 (0.94) | 8 (0.59) |
| | UCG | 152 (0.58) | 57 (0.59) | 133 (0.61) | Trp | UGG | 391 (1.00) | 149 (1.00) | 295 (1.00) |
| Pro | CCU | **337 (1.50)** | **120 (1.57)** | **266 (1.47)** | Arg | CGU | **280 (1.34)** | **97 (1.51)** | **218 (1.22)** |
| | CCC | 183 (0.82) | 63 (0.82) | 147 (0.81) | | CGC | 88 (0.42) | 35 (0.55) | 66 (0.37) |
| | CCA | 265 (1.18) | 85 (1.11) | 219 (1.21) | | CGA | **294 (1.41)** | **78 (1.22)** | **265 (1.48)** |
| | CCG | 112 (0.50) | 38 (0.50) | 93 (0.51) | | CGG | 101 (0.48) | 28 (0.44) | 90 (0.50) |
| Thr | ACU | **416 (1.54)** | **150 (1.65)** | **318 (1.47)** | Ser | AGU | **315 (1.20)** | **114 (1.18)** | **250 (1.14)** |
| | ACC | 216 (0.80) | 77 (0.85) | 166 (0.77 | | AGC | 89 (0.34) | 42 (0.43) | 67 (0.31) |
| | ACA | 336 (1.25) | 100 (1.10) | 286 (1.32) | Arg | AGA | **369 (1.77)** | **105 (1.64)** | **329 (1.84)** |
| | ACG | 111 (0.41) | 37 (0.41) | 97 (0.45) | | AGG | 121 (0.58) | 42 (0.65) | 104 (0.58) |
| Ala | GCU | **540 (1.81)** | **166 (1.78)** | **423 (1.81)** | Gly | GGU | 482 (1.28) | 171(1.31) | 354 (1.21) |
| | GCC | 195 (0.65) | 62 (0.66) | 157 (0.67) | | GGC | 178 (0.47) | 57 (0.44) | 135 (0.46) |
| | GCA | 330 (1.11) | 104 (1.11) | 260 (1.11) | | GGA | **589 (1.56)** | **188 1.44)** | **487 (1.66)** |
| | GCG | 127 (0.43) | 42 (0.45) | 97 (0.41) | | GGG | 263 (0.70) | 108 (0.82) | 197 (0.67) |

*Preferred codon for each amino acid is shown in bold

**Table 5:** Correlation analysis between four different axes of COA and various codon usage indices

| Axes | $A_3$ | $T_3$ | $G_3$ | $C_3$ | $GC_3$ | ENC | CAI | Gravy score | Aromaticity | Length (CDS) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Axis 1** | -0.481** | 0.176 | -0.007 | 0.542** | 0.411** | 0.364** | 0.498** | -0.036 | 0.144 | 0.094 |
| **Axis 2** | 0.154 | -0.057 | 0.071 | -0.227 | -0.152 | -0.193 | 0.381** | -0.250 | 0.061 | -0.089 |
| **Axis 3** | -0.191 | 0.098 | -0.276* | 0.395** | 0.138 | -0.377** | 0.034 | 0.263 | 0.234 | -0.061 |
| **Axis 4** | -0.258 | 0.239 | 0.091 | 0.014 | 0.092 | -0.038 | 0.035 | 0.261 | 0.210 | -0.223 |

** Figures are significant at $p < 0.01$

* Figures are significant at $p < 0.05$