BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(10)

Codon usage pattern in human SPANX genes

Monisha Nath Choudhury & Supriyo Chakraborty*

Department of Biotechnology, Assam University, Silchar-788011, Assam, India; Supriyo Chakraborty – Email: supriyoch_2008@rediffmail.com; *Corresponding author

Received September 22, 2015; Accepted October 02, 2015; Published October 31, 2015

Abstract:
Background: SPANX (sperm protein coupled with the nucleus in the X chromosome) genes play a crucial role in human spermatogenesis. Codon usage bias (CUB) is a well-known phenomenon that exists in many genomes and mainly determined by mutation and selection. CUB is species specific and a unique characteristic of a genome. Analysis of compositional features and codon usage pattern of SPANX genes in human has contributed to explore the molecular biology of this gene. In our current study, we have retrieved the sequences of different variants of SPANX gene from NCBI using accession number and a perl script was used to analyze the nucleotide composition and the parameters for codon usage bias. **Results:** Our results showed that codon usage bias is low as measured by codon bias index (CBI) and most of the GC ending codons were positively correlated with GC bias as indicated by GC3. That mutation pressure and natural selection affect the codon usage pattern were revealed by correspondence analysis (COA) and neutrality plot. Moreover, the neutrality plot further suggested that the role of natural selection is higher than mutation pressure on SPANX genes. **Conclusions:** The codon usage bias in SPANX genes is not very high and the role of natural selection dominates over mutation pressure in the codon usage of human SPANX genes.

Keywords: SPANX gene, codon usage, synonymous codon, natural selection, mutation pressure

Background:

The genes of *SPANX* family (sperm protein associated with the nucleus in the X chromosome) located in a cluster on Xq27 chromosome encode the protein products that are expressed in germ cells, non gametogenic tissue as well as several tumors **[1].** *SPANX* genes encode small unfolded proteins of approximately 100 amino acid residues and these resemble with the high mobility group A (*HMGA*) proteins to some extent which are involved in the formation of different nucleoprotein complexes. They can form dimers and complex with other proteins resembling the HMGA proteins **[2].** SPANX proteins are linked with the nuclear envelope in transformed mammalian cells, similar to the one in human spermatozoa. *SPANX* genes emerge to have evolved under strong positive selection, parallel to genes associated with reproduction **[3]**. They consist of two subfamilies *SPANX-A/D* and *SPANX-N.* SPANX-A/D proteins are found within the cytoplasm associated with the nuclear envelope in the mature spermatids **[4]**. *SPANX-A/D* genes map within segmental duplications that are the regions involved in genomic rearrangements resulting in an abnormally high level of structural polymorphisms. SPANX A1 serves as a biochemical marker to study unique structures in spermatozoa. Accordingly, the *SPANX-B* and the *SPANX-C* genes were shown to be present in variable copy number (ranging from one to >11) in the normal population. *SPANX-A/D* genes help in spermatogenesis but their expression was not found in nongametogenic tissue. Analysis of *SPANX* gene homologs (nonhuman primates) showed that *SPANX-A/D* genes arose nearby 7 million years ago and followed expansion in hominids **[3].** The *SPANX-N* gene subfamily found in all mammals gave rise to the *SPANX-A/D* subfamilies in the hominoid lineage. The *SPANX N* (N1, N2, N3 and N4) are mapped 1.3 Mb away from the cluster of *SPANX-A/D* gene and *SPANX-N5* is located on the short arm of the X chromosome at Xp11**[5].**

It is well known that genetic code consists of 64 codons out of which 61 encode 20 standard amino acids but the remaining three codons encode termination signals (UAA, UAG, and UGA). The usage of synonymous codons is different in the genes of an organism and also among other organisms. Unequal usage of synonymous codons is called ''codon usage bias''. Codon usage bias is an intricate evolutionary phenomenon, and exists in diverse organisms, from prokaryotes to unicellular and multicellular eukaryotes. The usage of synonymous preferred codons is a unique property of a genome [6]. Generally, mutational pressure and natural selection have been reported to be the two vital factors contributing to synonymous codon usage discrepancy among genes of an organism [7]. However, mutation in the synonymous codon generally occurs in the third base position without varying the primary sequence of the protein product. In some organisms, mutation pressure plays a central role in influencing the pattern of synonymous codon usage with extremely high A, T, G or C content. Further the processes of DNA replication, transcription, gene structure, and environmental conditions significantly influence codon usage pattern [8]. The alteration of synonymous codon usage pattern is a skill for reengineering genomes from the nucleotide level to the mega base scale [9]. Codon usage bias has practical implications in mRNA translation, new gene discovery, design of transgenes, and studies of molecular biology and evolution.

Analysis of codon usage pattern is a key tool for understanding the molecular mechanism of codon distribution. The present study was undertaken to elucidate the compositional features and codon usage pattern in SPANX genes in human. Our analysis has given a novel insight into the codon usage patterns of SPANX genes that would assist in better understanding of the synonymous codon usage pattern as well as the factors influencing it.

Methodology:

Coding sequence data

Using accession numbers different variants of SPANX genes were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). Only those coding sequences (cds) were considered for analyses which are exact multiples of three bases with proper start and stop codon. The accession numbers of 46 cds are shown in **Table 1 (see supplementary material).**

Indices of codon usage bias

Relative synonymous codon usage (RSCU) was calculated for the 59 synonymous codons for exploring the pattern of codon usage in the translation of amino acids. RSCU >1.6 indicated that codons were over-represented while the RSCU values >1.0 indicated that the codon is more frequently used [10]. The formula used to estimate RSCU is as follows

$$RSCUij = \frac{Xij}{\frac{1}{ni}\sum_{j=1}^{ni} Xij}$$

where, X_{ij} is the frequency of occurrence of the j^{th} codon for i^{th} amino acid (any X_{ij} with a value of zero is arbitrarily assigned a value of 0.5) and n_i is the number of codons for the i^{th} amino acid (i^{th} codon family).

The codon adaptation index (CAI) was used to estimate the degree of gene expression level of a single gene. The CAI value ranged between 0 and 1.0, and high value of CAI indicates high gene expression [11]. The CAI is calculated as

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

where ωk is the relative adaptiveness of the kth codon and L is the number of synonymous codons in the gene.

The codon bias index (CBI) measures the extent to which preferred codons are used in a gene. The formula used to calculate CBI is as follows

CBI=Nopt−NranNtot−Nran

Where N_{opt} is the number of preferred optimal codons, N_{tot} is the total number of codons, and $N_{ran}$is the expected number of optimal codons if random codon assignments were made for each amino acid [12]. GRAVY (Grand Average of Hydropathicity) values are the sum of the hydropathy values of all the amino acids in the encoded protein of the gene divided by the number of residues in the sequence [13]. Aromo stands for aromaticity and refers to the frequency of aromatic amino acids (Phe, Tyr, Trp) in the translated gene product [14].

The frequency of overall A,T,G,C and their frequency at third codon position , overall GC content and GC contents at first, second and third (GC1, GC2, GC3) position were calculated using a perl script. GC3s was used as a good marker for compositional constraint bias.

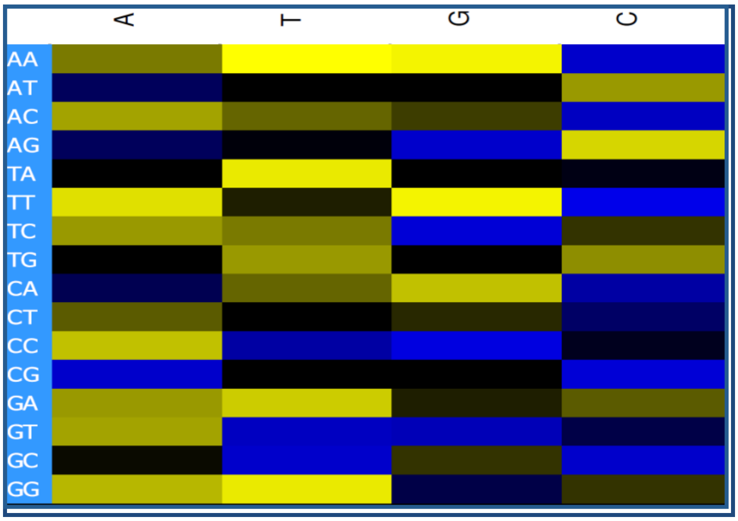

Figure 1: Heat maps of correlation coefficient values for codon usage vs GC3 for human SPANX gene. The color and intensity indicates type and degree of correlation: blue indicates positive, yellow negative. Black fields are stop and non-degenerate codons (tryptophan and methionine).

Analysis tools

Codon usage parameters and compositional dynamics were calculated (excluding the codons for Met, Trp, and the termination codons) using the Perl script developed by corresponding author SC. Correspondence analysis (COA) is a multivariate statistical analysis used to analyse the variation in codon usage pattern using XLSTAT. Correspondence analysis uses RSCU value and its axes 1 and 2 contribute to total variation. Correlation and regression analysis were carried out by using the multi-analysis software SPSS 21.0.

Result & Discussion:

Codon usage bias can be affected by the overall nucleotide composition of genomes [15]. Therefore, we first analyzed the compositional features of coding sequences from different variants of SPANX genes. It is observed from the **Table 2 (see supplementary material)**, nucleobase A and G3 were the highest, with average values of 116.78 and 32.69 respectively whereas nucleobase T and T3 were the lowest, with average values of 53.54 and 17 indicating that the variants in SPANX gene might use mostly A ending codons and less T ending codons. The average GC and AT % were 48.15 and 51.85 respectively and the gene is AT rich. These results suggest that compositional constraints might affect the codon usage pattern in SPANX gene supporting the result of Hoda *et.al.* in the codon usage pattern in human albumin superfamily **[16].** The average value of CBI used as a parameter of codon usage bias was 0.3273, which suggested that the codon usage bias was low and maintained a stable level which was similar to the findings of Huda *et.al.* **[16].** Zhang *et.al.* reported that the codon usage bias was low in TTSuV2 virus using effective number of codons (ENC) as CUB parameter **[8].**

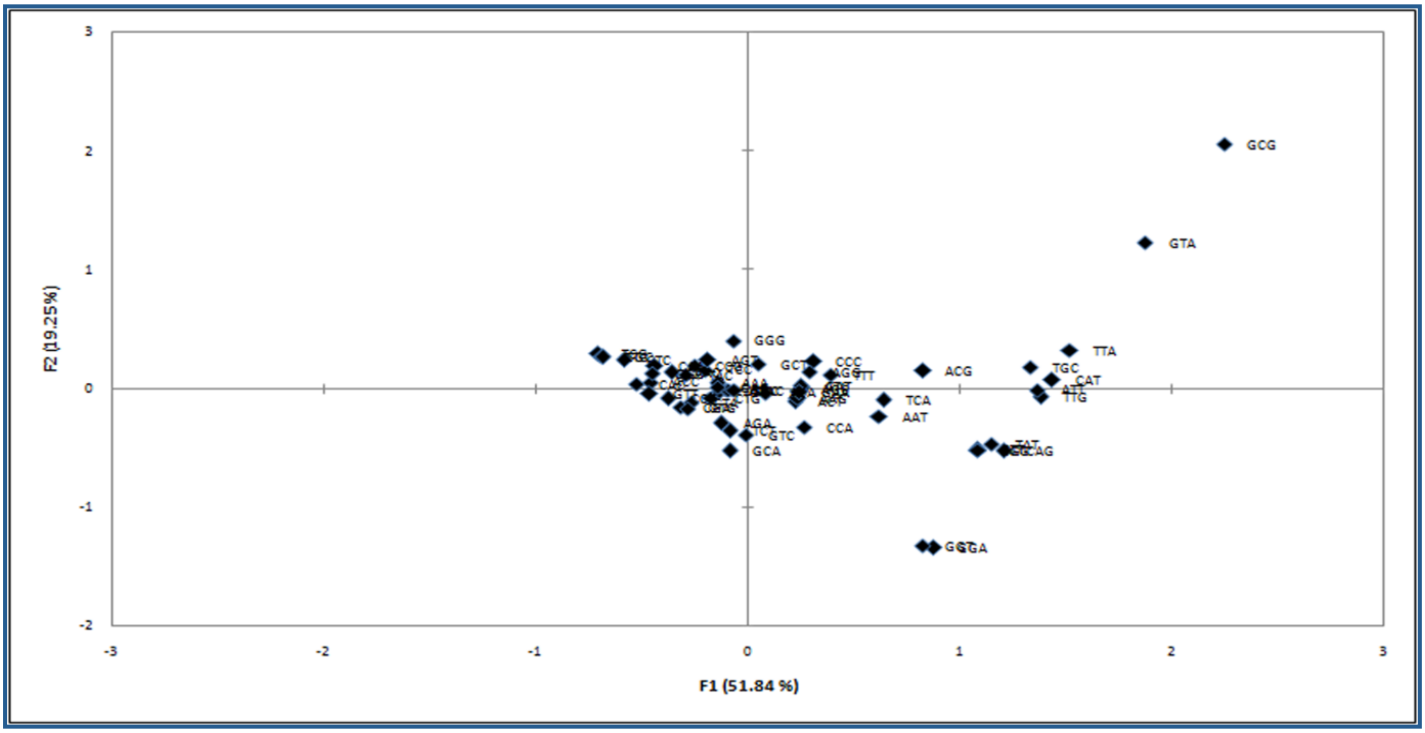

Figure 2: Correspondence analysis of RSCU value for the SPANX genes. Distribution of the 46 genes in SPANX on the plane corresponding to the coordinates on the first and second principal axes was shown.

In order to investigate the codon usage pattern of SPANX gene, we correlated codon usage with GC3 content. **Figure 1** shows the heat map of the correlation coefficients between codon usage and GC bias in human SPANX gene. In our analysis, most of the G- and C-ending codons were positively correlated with GC3, and most of the A- and T-ending codons were negatively correlated with GC3. However, twelve G/C ending codons namely ACG, TTG, CAG, CTG, GAG, GCG, ATC, AGC, TCC, TGC, GAC and GGC showed a negative correlation between codon usage and GC3 whereas eight A/T ending codons, ATA, AGA, CAA, CGA, AGT, CCT, GTT and GCT showed a positive correlation with GC3. This indicates that twelve G/C ending codons will show decreasing usage with increasing GC bias as indicated by GC3 and eight A/T ending codons will show increasing usage with increasing GC3 bias. Palidwor et.al reported that GC ending codons were mostly positively correlated with GC3 and AT ending codons were mostly negatively correlated with GC3 in codon usage pattern in prokaryotes, plants and human thus supporting our results **[17].**

To investigate the variation of codon usage in the SPANX genes, correspondence analysis (COA) was performed based on the RSCU values of each gene (**Figure 2).** The COA of different variants of SPANX gene detected the first principal component (F1'), which could account for 51.84 % of the total synonymous codon usage variation, whereas the second principal component (F2') accounted for 19.25 % of the total variation. Again, several significant correlations were observed between the two principal axes and nucleotide contents **Table 3 (see supplementary material).** Axis 1(F1) showed a significant positive correlation with A, T, A3, T3 but showed significant negative correlation with C3, GC, GC1, GC2, GC3, Gravy and CBI. Axis 2 (F2) of COA showed significant positive correlation with GC2, GC3, ARO, Gravy and CBI while significant negative correlation with A, T, G, C, A3, T3, G3, C3, GC1, CAI and Laa. Our analysis suggests that mutational pressure and natural selection might have played a major role in shaping the dynamics of codon usage patterns within different variants of SPANX gene supporting the finding of Wei **[18].**

A neutrality plot was drawn to estimate the magnitude of natural selection against mutation pressure in the codon usage pattern of SPANX gene. Neutrality plot is the regression analysis of G12 (GC12 average of GC1 and GC2) on GC3. The points in the neutrality plot are not diagonally distributed and the values of GC3 are in a narrow distribution, indicating that GC12 and GC3 are definitely not due to the mutational bias **(Figure 3).** On the other hand, the regression curve (green line) tended to slope towards the horizontal axis. The regression coefficient of GC12 on GC3 in SPANX genes is 0.242, indicating the relative neutrality is 24.20 % while the relative constraint is 75.80 % for GC3. This result suggests natural selection played a major role while mutation pressure played a minor role in shaping the codon usage pattern in SPANX gene. Jia *et.al.* also found that natural selection played a prominent role in codon usage pattern in *Bombyx mori* **[19].** We also found similar result.

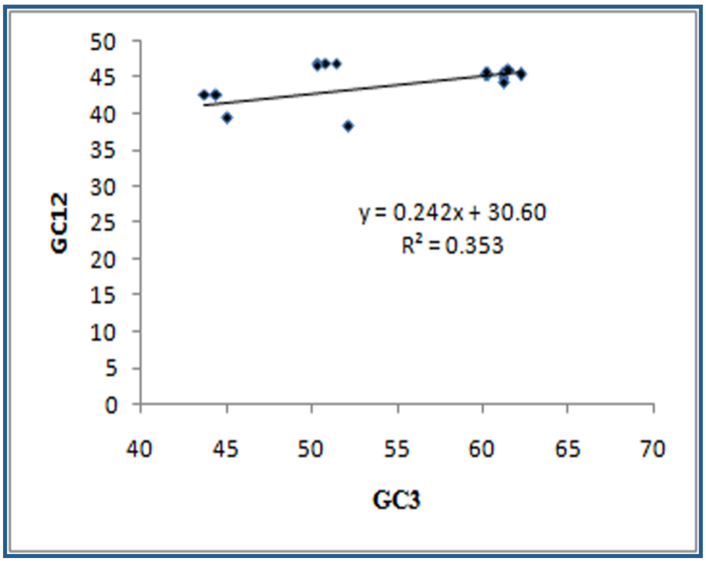

Figure 3: Neutrality plot analysis of the GC12 (GC12 stands for the average value of GC content in the first and second position of the codons) and GC content at the third codon position (GC3) for the coding sequence of SPANX genes. The solid line is the linear regression of GC12 against GC3. Y=0.242x+30.60, R2=0.353.

Conclusions:

The codon usage bias is not very high in SPANX genes. The overall GC content is low and the gene is AT rich. Natural selection is the major determining factor in shaping the pattern of codon usage in different variants of SPANX gene rather than mutation pressure.

Acknowledgement:

We are thankful to Assam University, Silchar, Assam, India for providing the necessary infra structural facilities in carrying out this research work. There is no conflict on interest in this research work.

References:

[1] Zendman AJ *et al. Gene* 2003 **309**: 125 [PMID: 12758128]
[2] Berman HM *et al. Nucleic Acids Res.* 2000 **28:** 235 [PMID: 10592235]
[3] Kouprina N *et al. Proc Natl Acad Sci USA* 2004 **101:** 3077 [PMID: 14973187]
[4] Kouprina N *et al. PLoS One* 2007 **2**: e359 [PMID: 17406683]
[5] Zendman AJ *et al. Cancer Res.* 1999 **59**: 6223 [PMID: 10626816]
[6] Grantham R *et al. Nucleic Acids Res.* 1981 **9**: r43 [PMID: 7208352]
[7] Stenico M *et al. Nucleic Acids Res.* 1994 **22:** 2437 [PMID: 8041603]
[8] Zhang Z *et al. Arch virol*. 2013 **158**: 145 [PMID: 23011310]
[9] Shi SL *et al. Virus Genes* 2013 **46:** 10 [PMID: 22996735]
[10] Sharp PM & Li WH, *Nucleic Acids Res.* 1986 **14**: 7737 [PMID: 3534792]
[11] Carbone A *et al.* B*ioinformatics* 2003 **19:** 2005 [PMID: 14594704]
[12] Sur *et al. Indian Journal of Biotechnology* 2007 **6:** 321
[13] Kyte J & Doolittle RF, *J Mol Biol*. 1982 **157**: 105 [PMID: 7108955]
[14] Lobry JR & Gautier C, *Nucleic Acids Res.* 1994 **22:** 3174 [PMID:8065933]
[15] Jenkins GM & Holmes EC, *Virus Res.* 2003 **92:** 1 [PMID: 12606071]
[16] Mirsafian H *et al. Scientific World Journa*l 2014 **2014:** 639682 [PMID: 24707212]
[17] Palidwor GA *et al. PLoS One.* 2010 **5:** e13431 [PMID: 21048949]
[18] Wei L *et al BMC Evol Biol. 2014* **14**: 262 [PMID: 25515024]
[19] Jia X *et al. BMC Genomics*. 2015 **16**: 356 [PMID: 25943559]

Edited by P Kangueane

Citation: **Choudhury & Chakraborty**, Bioinformation 11(10): 454-459 (2015)

Supplementary material:

Table 1: Cds no, accession no and gene name

Cds No	Accession No	Homo sapiens SPANX mRNA complete cds
cds1	GI:614458155	SPANX family, member N3 (SPANXN3)
cds2	GI:608601809	Sperm protein associated with the nucleus, X-linked, family member A1 (SPANXA1)
cds3	GI:608601808	SPANX family, member A2 (SPANXA2)
cds4	GI:84783510	SPANX-N3 locus variant 2 mRNA
cds5	GI:84783508	SPANX-N3 locus variant 1
cds6	GI:84783506	SPANX-N4 locus variant 2
cds7	GI:84783504	SPANX-N4 locus variant 1
cds8	GI:84783494	SPANX-N1 locus variant 4
cds9	GI:84783492	SPANX-N1 locus variant 3
cds10	GI:84783490	SPANX-N1 locus variant 2
cds11	GI:84783488	SPANX-N1 locus variant 1
cds12	GI:84783482	SPANX-N2 locus variant 5
cds13	GI:84783480	SPANX-N2 locus variant 4
cds14	GI:84783478	SPANX-N3 locus variant 3
cds15	GI:84783476	SPANX-N2 locus variant 2
cds16	GI:84783474	SPANX-N2 locus variant 1
cds17	GI:62860707	Isolate control15 SPANX-A2 (SPANXA2)
cds18	GI:62860705	Isolate control14 SPANX-A2 (SPANXA2)
cds19	GI:62860703	Isolate control13 SPANX-A2 (SPANXA2)
cds20	GI:62860701	Isolate control12 SPANX-A2 (SPANXA2)
cds21	GI:62860699	Isolate control11 SPANX-A2 (SPANXA2)
cds22	GI:62860697	Isolate control10 SPANX-A2 (SPANXA2)
cds23	GI:62860695	Isolate control9 SPANX-A2 (SPANXA2)
cds24	GI:62860693	Isolate control8 SPANX-A2 (SPANXA2)
cds25	GI:62860691	Isolate control7 SPANX-A2 (SPANXA2)
cds26	GI:62860689	Isolate control6 SPANX-A2 (SPANXA2)
cds27	GI:62860687	Isolate control5 SPANX-A2 (SPANXA2)
cds28	GI:62860685	Isolate control4 SPANX-A2 (SPANXA2)
cds29	GI:62860683	Isolate control3 SPANX-A2 (SPANXA2)
cds30	GI:62860681	Isolate control2 SPANX-A2 (SPANXA2)
cds31	GI:62860679	Isolate control1 SPANX-A2 (SPANXA2)
cds32	GI:6808525	Nuclear-associated protein SPAN-Xb (SPANX)
cds33	GI:6808523	Nuclear-associated protein SPAN-Xa (SPANX)
cds34	GI:13507166	SPAN-Xd
cds35	GI:187952644	SPANX family, member B1 (cDNA clone MGC:169156 IMAGE:9021533)
cds36	GI:187951712	SPANX family, member B1 (cDNA clone MGC:169159 IMAGE:9021536)
cds37	GI:126632046	SPANX family, member D (cDNA clone MGC:161912 IMAGE:40119568)
cds38	GI:120660241	SPANX family, member N4 (cDNA clone MGC:163377 IMAGE:40146536)
cds39	GI:120659907	SPANX family, member N4 (cDNA clone MGC:163375 IMAGE:40146534)
cds40	GI:115528465	SPANX family, member D(cDNA clone MGC:150331 IMAGE:40119569)
cds41	GI:74355481	SPANX family, member D(cDNA clone MGC:119769 IMAGE:40013988)
cds42	GI:38541646	SPANX family, member E (cDNA clone MGC:71908 IMAGE:4047937)
cds43	GI:38541204	SPANX family, member N3(cDNA clone MGC:72116 IMAGE:6618011)
cds44	GI:32450696	SPANX family, member C (cDNA clone MGC:61861 IMAGE:6648369),
cds45	GI:21759804	SPANX family, member B1 (cDNA clone MGC:26207 IMAGE:4824918)
cds46	GI:13529244	SPANX family, member E (cDNA clone MGC:12501 IMAGE:3935644)

Table 2: Compositional constraints and CBI

CDS No	A	T	G	C	A3	T3	G3	C3	GC%	AT%	CBI
cds1	165	77	101	83	51	28	33	30	43.19	56.8	0.2
cds2	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds3	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds4	165	77	101	83	51	28	33	30	43.19	56.8	0.2
cds5	165	77	102	82	51	28	34	29	43.19	56.8	0.22
cds6	126	50	65	59	32	23	26	19	41.33	58.7	0.22
cds7	127	49	65	59	33	22	26	19	41.33	58.7	0.22
cds8	86	39	50	44	22	13	23	15	42.92	57.1	0.34
cds9	86	39	50	44	22	13	23	15	42.92	57.1	0.34
cds10	86	39	50	44	22	13	23	15	42.92	57.1	0.34
cds11	86	39	50	44	22	13	23	15	42.92	57.1	0.34
cds12	195	86	143	119	62	27	51	41	48.25	51.7	0.37
cds13	194	86	143	120	61	27	51	42	48.43	51.6	0.39
cds14	195	87	143	118	62	28	51	40	48.07	51.9	0.36
cds15	195	87	143	118	62	28	51	40	48.07	51.9	0.36
cds16	195	88	142	118	62	28	50	41	47.88	52.1	0.38
cds17	98	48	75	73	25	14	31	28	50.34	49.7	0.36

cds18	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds19	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds20	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds21	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds22	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds23	96	49	78	71	25	13	32	28	50.68	49.3	0.34
cds24	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds25	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds26	99	45	78	72	25	13	32	28	51.02	49	0.39
cds27	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds28	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds29	97	49	76	72	25	13	32	28	50.34	49.7	0.3
cds30	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds31	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds32	110	42	80	80	27	13	33	31	51.28	48.7	0.35
cds33	98	48	75	73	25	14	31	28	50.34	49.7	0.36
cds34	98	49	76	71	25	13	32	28	50	50	0.33
cds35	110	42	79	81	27	13	33	31	51.28	48.7	0.34
cds36	110	42	79	81	27	13	33	31	51.28	48.7	0.34
cds37	98	49	76	71	25	13	32	28	50	50	0.33
cds38	127	49	65	59	33	22	26	19	41.33	58.7	0.22
cds39	127	49	65	59	33	22	26	19	41.33	58.7	0.22
cds40	98	49	76	71	25	13	32	28	50	50	0.33
cds41	98	49	76	71	25	13	32	28	50	50	0.33
cds42	97	47	78	72	25	12	33	28	51.02	49	0.32
cds43	165	78	102	81	51	29	34	28	42.96	57	0.22
cds44	99	46	78	71	26	13	33	26	50.68	49.3	0.35
cds45	110	42	80	80	27	13	33	31	51.28	48.7	0.35
cds46	97	47	78	72	25	12	33	28	51.02	49	0.32
Mean	116.783	53.5435	82.4565	75.3261	31.8696	17	32.6957	27.8043	48.1567	51.8565	0.32739

Table 3: Correlation between first two principal axes of COA and index of total genes' codon usage and synonymous codon usage bias

	A	T	G	C	A3	T3	G3	C3	GC	CAI	Gravy	Laa	CBI	GC1	GC2	GC3	ARO
F1	0.42**	0.34*	0.12	-0.08	0.44**	0.54**	0	-0.29*	-.84**	0.13	-0.76**	0.24	-0.36*	-0.44**	-0.68**	-0.80**	.13
F2	-0.76**	-0.66**	-0.65**	-0.63**	-.070**	-0.75**	-0.58**	-0.51**	.19	-0.75**	0.53**	-0.72**	0.35*	-0.41**	0.47**	0.50**	.55**

Note: ** p < 0.01, * p < 0.05.