Journal of Advanced Bioinformatics Applications and Research
ISSN 0976-2604.Online ISSN 2278–6007
Vol 7, Issue1, 2016, pp20-31
http://www.bipublication.com

Research Article

Exploring the diversities of codon and amino acid usage patterns in type strain *Corynebacterium urealyticum* DSM 7109: A Bioinformatics perspective

Shilpee Pal[1], Ayan Roy[2], Arnab Sen[2], Keshab C Mondal[1], and Pradeep Kumar Das Mohapatra*[1]

[1]Bioinformatics Infrastructure Facility Centre, Department of Microbiology, Vidyasagar University, Midnapore 721 102, India
[2]Department of Botany, School of Life Sciences, University of North Bengal, Darjeeling 734 013, India
*Corresponding author: Email: pkdmvu@gmail.com,
Tel: +91 3222 276554 (extn. 477); Fax: +91 3222 276554 (extn. 477)

ABSTRACT:

Corynebacterium urealyticum DSM 7109, an opportunistic pathogen retains a strong urease activity that hydrolyses urea and causes hyperammonuria. Overall codon usage study revealed its evolution as well as the expressivity of genes which regulate pathogenic gene expression. GC compositional constraint was present in this pathogen. Correspondence analysis (CA) on codon and amino acid usage revealed the effect of translational selection, hydrophobicity, aromaticity, protein biosynthesis cost on codon and amino acid usage bias in PHX (predicted highly expressed) genes. Higher frequency of C ending preferred codons as well as optimal codons was also an evidence of translational selection pressure on codon usage variation. Strand specific replicational selection was also present to avoid collisions between polymerases during PHX gene expression. Nucleotide substitution rate revealed PHX genes were less diverging in their synonymous positions thus conserving optimal codon usage, though synonymous substitution was shaping genomic evolution of the studied organism. Urease coding genes clustered along with PHX genes on CA plots and their expression as well as frequency of optimal codon usage were higher than other protein coding genes. Thus the present study reveals codon and amino acid usage variation as well as genomic evolution of *C. urealyticum* genes, which may help in further investigation of their pathogenesis in addition with host-pathogen interaction.

Keywords:relative synonymous codon usage, correspondence analysis, highly expressed gene, translational selection, nucleotide substitution, replicational selection

[I] INTRODUCTION

Codon usage bias performs a major role towards molecular evolution at DNA level of living organisms. According to codon degeneracy rule, most amino acids are coded by more than one codon which are called synonymous codons. According to "genome hypothesis", frequency of synonymous codon usage pattern is usually varies species to species. Moreover, synonymous codons differ by nucleotide in third position of codons and are conserved among genes [1]. Codon usage variation is a consequence of many important features like nucleotide compositional constraint [2], mutational bias [3], gene length [4], tRNA abundance [5, 6], codon–anticodon interaction [7], gene expression level [8, 9], amino acid conservation [10], replicational-transcriptional selection [11], synonymous and asynonymous nucleotide substitutions of orthologous genes [12] etc. In some prokaryotes such as, *Escherichia coli* [13], the occurrence of codon usage is directly proportional to the corresponding anticodon frequencies and the preferred codons in highly expressed genes are

predicted by the presence of most abundant tRNAs. In microbial organisms, the frequency of amino acid usage is also affected by several physiochemical properties such as hydrophobicity and aromaticity of respective proteins [14], as well as the ecological niches of that organism [15]. Other factors such as protein secondary structure [16, 17, 18, 19] and mRNA folding [20, 21] are also responsible for codon usage variation in genes. Thus analysis of codon usage data would reveal the fundamental cause of molecular evolution at gene level of an organism.

Corynebacterium urealyticum DSM 7109 is a gram positive bacillus with a strong urease (urea amidohydrolase, EC 3.5.1.5) activity [22], slow growing opportunistic pathogen [23]. It causes urinary tract infection by hydrolysing urea to ammonia thus increases renal pH and generates renal disease, renal stones, etc. [24]. In most cases, *C. urealyticum* becomes antibiotic resistant by rising urine pH. The complete genome sequence of *C. urealyticum* DSM 7109 has already been determined and deposited in public database [25] but sequence based analysis yet not been done. Genome size of *C. urealyticum* DSM 7109 is also small which may be the effect of bacterial life style in a specific niche, responsible for their pathogenic activities. In this study, codon usage as well as amino acid usage variations of *C. urealyticum* DSM 7109 has been analysed by using multivariate statistical analysis, optimal codon usage study, gene expression study, correlation analyses, nucleotide substitution rate, etc. to recognize the molecular evolution strategy of the genome.

[II] METHODS

2.1. Sequence retrieval

Total 2024 protein coding gene sequences of *C. urealyticum* DSM 7109 were retrieved from IMG-JGI database (https://img.jgi.doe.gov/).

2.2. Codon and amino acid usage pattern analysis

To analyse the codon and amino acid usage patterns in *C. urealyticum* DSM 7109genes, several parameters such as gene length, the percentage of nucleotide compositions at third position like A_{3s}, T_{3s}, G_{3s} and C_{3s}, total amount of GC as well as total GC compositions at the third position of codons (GC_{3s}), hydrophobicity, aromaticity were calculated for each protein coding gene. CodonW [26] software was used to accomplish said principle. Biosynthesis cost of proteins was calculated by using DAMBE [27] software.

2.3. Detection of the trend in codon usage pattern

CodonW [26] was also used to determine heterogeneity of codon usage pattern such as effective number of nucleotides (ENc), characteristics of synonymous codon usage like, RSCU (relative synonymous codon usage), frequency of optimal codons (Fop), etc.

Expected curve of GC_3-ENc plot was determined by the formula:

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

where, $s = GC_{3s}$.

The continuous curve interpreted relationship between ENc and GC_3 under H_0 (no selection)[28].

RSCU of codon 'i' was defined as, $RSCU_i = (Obs_i/ Exp_i)$, where, Obs_i= observed number of occurrences of codon i and Exp_i= expected number of occurrences of same codon [29].

Fop was calculated by the formula:

$$\text{Optimal Fop} = \frac{\text{The number of "optimal" codons}}{\text{Sum of the numbers of "optimal" and "non-optimal" codons}}$$

codons were recognized by the most abundant isoacceptingtRNAs [13].

2.4. Gene expression study

Codon adaptation index (CAI) measurement is a broadly used technique to study gene expression level in both prokaryotes and eukaryotes [26]. CAI was calculated by the formula:

$$CAI = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_{c(k)}$$

where, L is the number of codon in a gene, $w_{c(k)}$ is relative adaptiveness of k^{th} codon [30].

In this study, CAI was calculated according to the equation of Sharp and Li [31] in CAI Calculator 2 (http: //www.evolvingcode./net/codon/) by considering ribosomal protein coding genes as a reference set of genes. After sorting the genes according to their CAI values, top 10% genes were predicted as highly expressed genes (PHX) and bottom 10% genes were predicted as lowly expressed genes (PLX).

2.5. Localizing the genes on leading and lagging strand

To determine the leading and lagging strand of replication, origin of replication and termination sites, GC skew was calculated using GenSkew: Genomic nucleotide skew application (http://genskew.csb.univie.ac.at/) by taking a 200 kb window size and a step size of 3kb. Total number of genes in both strands was calculated by developing PERL scripts.

2.6. Multivariate statistical analysis

Correspondence analysis was performed using CodonW [26]. It is a geometric approach for data analysis[32], which was carried out to detect codon and amino acid usage variations among the studied genes. A multidimensional space of 59 axes was created for CA on RSCU, whereas, for CA on RAAU (relative amino acid usage), a multidimensional space of 20 axes was developed, where each axis was explaining a decreasing amount of variation [33]. It was carried out on simple codon count and amino acid frequencies.

2.7. Nucleotide substitution rate calculation

Nucleotide substitution rate was analysed by calculating (non-synonymous substitutions per non-synonymous substitutions per time period) and Ks (synonymous substitutions per synonymous substitutions per time period) between orthologous genes of *C. glycinophilum*and *C. urealyticum.* Orthologous genes were predicted by using Reciprocal Best Blast Hit (RBBH) approach with an identity level of ≥50% and an E value of 1e 10 and with at least 50% alignment score in local BLASTP program (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/). Ka and Ks were calculated by using Codeml program in PAML package (version 4.5) (http://abacus.gene.ucl.ac.uk/software/paml.html) [34]. Spearman's rho correlation coefficient tests between codon and amino acid usage indices were performed by SPSS statistical software (version 19).

[III] RESULTS

3.1. Codon usage analysis

Overall codon usage of *C. urealyticum* DSM 7109 genes (excluding ATG for Met and TGG for Trp) and relative synonymous codon usage (RSCU) shown in Table 1. It was observed that, among 27 preferred codons 55.56% codons was C ending, 37.04% was G ending and only 3.70% of each was A and T ending. Moreover, among 19 optimal codons 73.68% codons was C ending and rest was G ending. A or T ending codons was not used as optimal codons in *C. urealyticum* genes.

3.2. Heterogeneity of codon usage among genes

Protein coding genes of *C. urealyticum* biased towards high GC content ranging from 39% to 73.70% (data not shown) with a mean value of 64.58% and standard deviation of 4.02%. GC_3 content was ranging from 37.20% to 94.20% with an average value of 80.20% and standard deviation of 8.42% and ENc values were ranging from 20.82% to 61% with a mean value of 37.75% and standard deviation of 7.12%. Genes with higher GC_3 content showed lower ENc values (Table 2). GC_3-ENc plot (Figure 1) also depicted that ENc decreased with corresponding increase in GC_3 and considerable number of genes were lying below the expected curve towards GC_3 dense area. PHX as well as urease coding genes deposited in same region towards GC_3 rich region in GC_3-ENc plot. Expression level of protein coding genes was not so high (Table 2) and was ranging from 0.079 to 0.928 (Figure 2) with an average of 0.474 and standard deviation of 0.14. Majority of genes displayed CAI values from 0.35 to 0.55. Urease coding genes showed higher expression level than other protein coding genes but not as high as ribosomal protein coding genes. Frequency of optimal codons was

higher in urease coding genes with higher GC_3 content (GC_3>0.80) (Table 3). Hydrophobicity of these genes was lower.

3.3. Multivariate statistical analysis on codon and amino acid usages and correlation tests

Correspondence analysis (CA) was performed on RSCU values as well as amino acid usage to investigate the major possible trend responsible for codon and amino acid usage bias in *C. urealyticum* genes. CA on RSCU (Figure 3a) accounted for f1 (18.92%) and f2 (8.81%) as first (axis 1) and second (axis 2) major axes of the total variations in data respectively. Here, protein coding genes clustered towards origin of the axes. PHX and PLX genes distinctly deposited on the plot, urease coding genes deposited along with PHX genes. Correlation tests showed significant correlations between axis 1 and C_3, G_3, GC, GC_3, ENc, CAI and Fop (Table 4). Significant positive correlation was noticed between GC_3 and CAI (0.704, p<0.01), whereas a significant negative correlation was observed between GC_3 and ENc (-0.775, p<0.01). CAI also significantly negatively correlated with ENc (-0.878, p<0.01). All the above observations were an evidence of the presence of translational selection of codons in gene expression of *C. urealyticum*. C_3 significantly positively correlated with CAI (0.736, p<0.01). Fop also significantly positively correlated with GC_3 (0.569, p<0.01) and with CAI (0.774, p<0.01). Significant positive correlation was observed between Fop and C_3 (0.649, p<0.01), which supported the former prediction.

CA on amino acid usage (Figure 3b) accounted for 18.02% (axis 1) and 12.45% (axis 2) respectively. In this plot, protein coding genes also deposited towards the centre of the two axes. Here PHX and hydrophobic protein coding genes deposited separately, and urease coding genes deposited along with PHX genesas in CA on RSCU before. It revealed PHX genes were not including hydrophobic protein coding genes. Moreover, correlation tests showed axis 1 and axis 2 significantly correlated with CAI, gravy, aromaticity and mean cost of proteins (Table 5). CAI significantly negatively correlated with hydrophobicity (-0.081, p<0.01) as well as with aromaticity (-0.060, p<0.01). Significant positive correlation was noticed between hydrophobicity and aromaticity of genes (0.080, p<0.01). Protein biosynthesis cost strongly negatively correlated with CAI (-0.129, p<0.01), whereas hydrophobicity and aromaticity significantly positively correlated with protein synthesis cost (0.047, p<0.05) and (0.821, p<0.01) respectively.

3.4. Location and expression of genes on leading and lagging strand

Among both complementary strands, leading strand was enriched in positive GC skew values, represented abundance of G over C. The Z-curve (Figure 4) indicated the origin and terminus of genome, where, X and Y axes denoted chromosomal location (in base pair) and GC skew values correspondingly. Gene expression of most of the protein coding genes was from 0.40 to 0.65 and leading strand contained 55.25% of the said genes (Table 6). About 55% protein coding genes with CAI>0.65 were residing on leading strand of replication. Proportion of ribosomal protein coding genes on leading strand touched 73.33% and all urease coding genes were coded by leading strand of replication. Thus, amount of genes located on leading strand was higher than lagging strand of replication in *C. urealyticum*.

Table-1. Overall RSCU values and codon count of genes.

Amino Acid	Codon	Codon Count	RSCU
Phe	UUU	4482	0.05
	UUC*	10874	1.42
Leu	UUA	2510	0.28
	UUG	7430	0.84
	CUU	8907	1.00
	CUC*	14100	1.59

	CUA	3947	0.44
	CUG*	16433	1.85
Ile	AUU	3975	0.63
	AUC*	12298	1.93
	AUA	2797	0.44
Val	GUU	7996	0.78
	GUC*	13702	1.34
	GUA	4871	0.47
	GUG	14476	1.41
Tyr	UAU	2846	0.74
	UAC*	4890	1.26
His	CAU	9449	0.76
	CAC*	15301	1.24
Gln	CAA	7904	0.66
	CAG*	16203	1.34
Asn	AAU	3881	0.63
	AAC*	8352	1.37
Lys	AAA	4802	0.68
	AAG*	9333	1.32
Asp	GAU	12081	0.92
	GAC*	14103	1.08
Glu	GAA	11172	0.86
	GAG*	14691	1.14
Ser	UCU	8139	0.64
	UCC*	15249	1.20
	UCA	10189	0.80
	UCG	19875	1.56
	AGU	5951	0.47
	AGC	16811	1.32
Pro	CCU	14211	0.76
	CCC	17466	0.93
	CCA	16801	0.90
	CCG*	26412	1.41
Thr	ACU	6167	0.56
	ACC*	16391	1.50
	ACA	6768	0.62
	ACG	14354	1.31
Ala	GCU	16651	0.78
	GCC*	25457	1.19
	GCA	15658	0.73
	GCG	27484	1.29
Cys	UGU	6481	0.59
	UGC*	15368	1.41
Arg	CGU	14009	0.77
	CGC*	27210	1.50
	CGA	19339	1.07
	CGG	25549	1.41
	AGA	8268	0.46
	AGG	14389	0.79
Gly	GGU	15268	0.84
	GGC*	24936	1.37
	GGA	15398	0.85
	GGG	17142	0.94

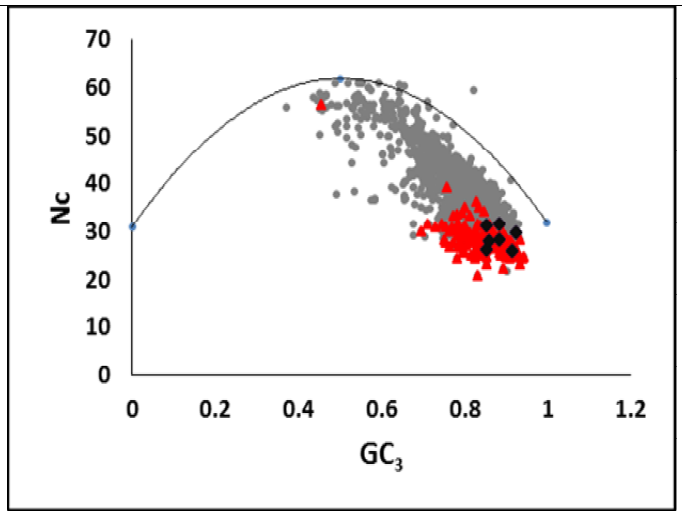

Fig: 1.GC_3-ENc plot genes. The continuous curve identifies the relationship between ENc and GC_3 under random codon usage. Grey circles indicate protein coding genes. Red triangles indicate PHX genes. Black circles indicate urease coding genes.

Preferred codons (RSCU>1) are in bold and optimal codons are with sign * (*p<0.01*)

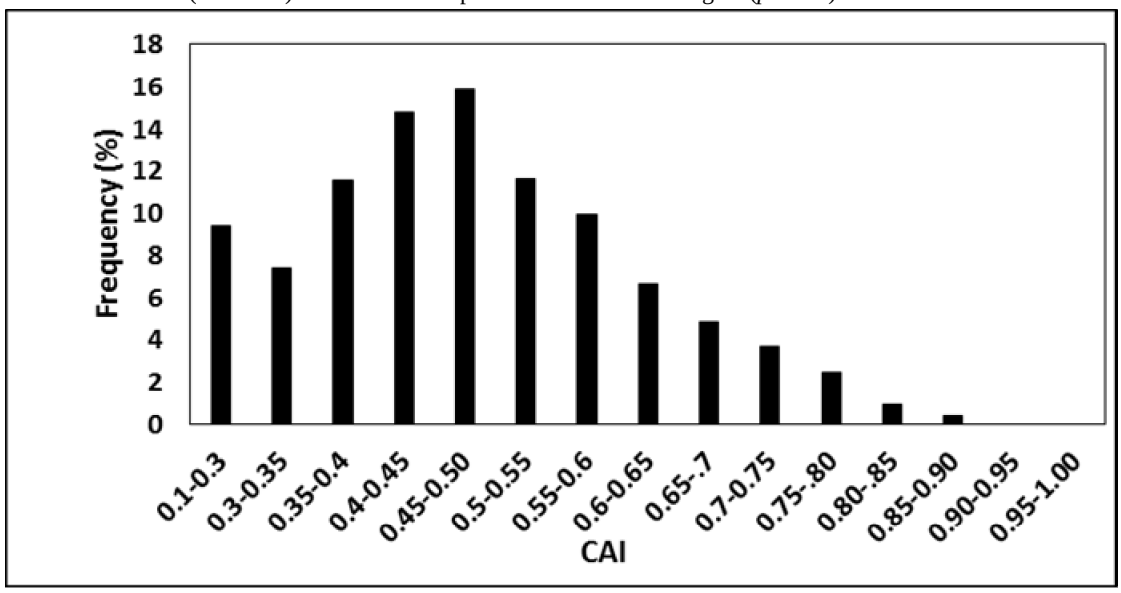

Fig: 2. Frequency distribution of CAI values for protein coding genes.

Table-2. Mean values of codon and amino acid usage indices.

Gene group	GC%	GC_3%	ENc	Fop	CAI
Protein coding genes	64.58±4.02	80.20±8.42	37.75±7.12	56.84±6.25	0.474±0.14
Ribosomal genes	62.43±2.64	79.40±5.08	31.30±4.70	63.49±5.57	0.709±0.09
Urease coding genes	63.97±1.72	88.18±2.86	28.79±2.26	64.71±5.65	0.687±0.08

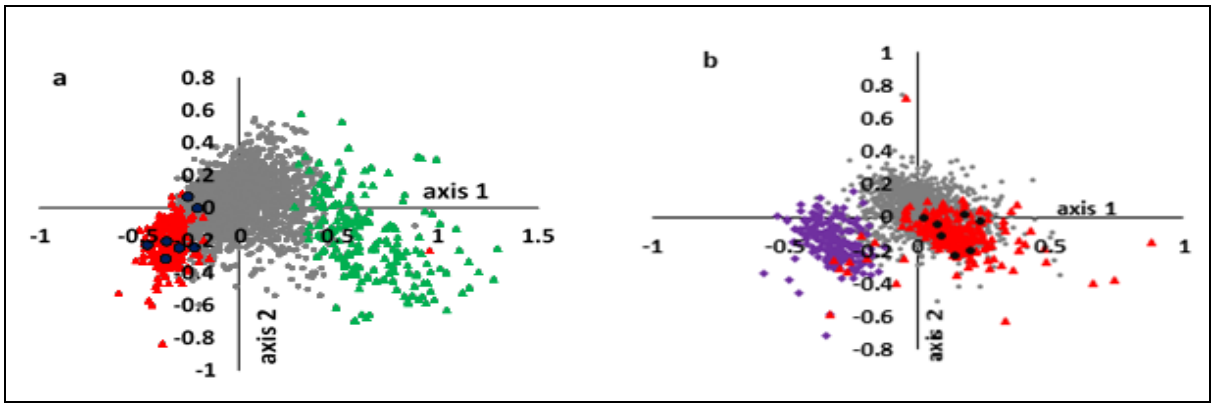

Fig: 3. Correspondence analysis on (a) RSCU values, where grey circles indicate protein coding genes,red and green triangles indicate PHX, PLX genes respectivelyand black circles indicate urease coding genes; and on (b) amino acid usage, where grey circles indicate protein coding genes;red triangles and violet diamonds indicate PHX genes and highly hydrophobic protein coding genes respectively. Urease coding genes are indicated in black circles.

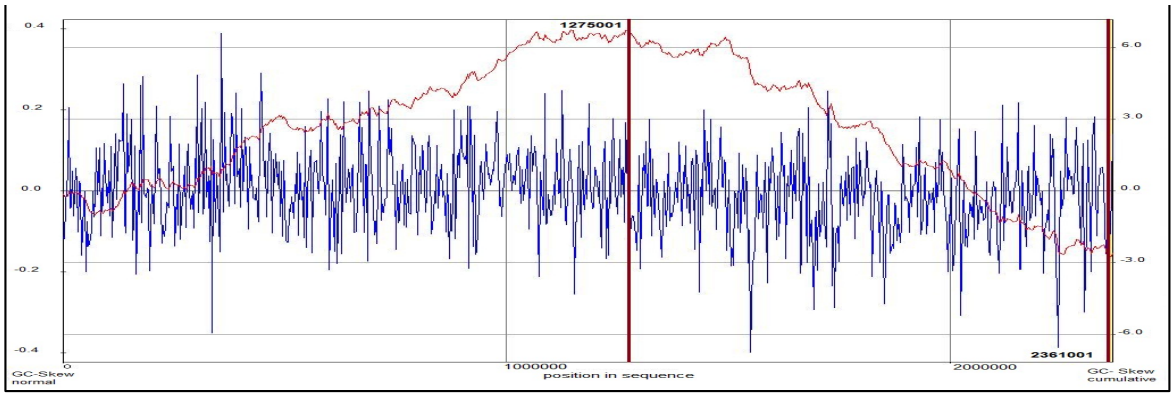

Fig: 4.GC skew identifies probable sites of origin of replication and terminus in *C. urealyticum* DSM 7109 genome.

Table-3. Codon usage indices of urease coding genes.

IMG ID	Description	GC_3	ENc	CAI	Fop	Gravy
641674198	Urease gamma subunit	0.85	29.96	0.613	0.537	-0.369
641674199	Urease beta subunit	0.81	28.25	0.738	0.716	-0.532
641674200	Urease alpha subunit	0.84	28.91	0.736	0.682	-0.248
641674201	Urease accessory protein	0.91	28.22	0.713	0.643	-0.570
641674202	Urease accessory protein	0.91	30.22	0.596	0.679	-0.264
641674203	Urease accessory protein	0.90	27.71	0.775	0.65	-0.130
641674204	Urease accessory protein	0.93	30.52	0.617	0.662	-0.328

Table-4. Correlations between codon usage indices.

	Axis1	Axis2	C_{3s}	G_{3s}	GC	GC_{3s}	ENc	CAI	Fop
Axis1	1	0.176**	-0.753**	-0.117**	-0.258**	-0.768**	0.899**	-0.983**	-0.761**
Axis2		1	-0.369**	0.569**	0.561**	0.241**	0.120**	-0.219**	-0.194**
C_{3s}			1	-0.323**	0.180**	0.656**	-0.732**	0.736**	0.649**
G_{3s}				1	0.228**	0.322**	-0.101**	0.077**	-0.008
GC					1	0.531**	-0.301**	0.220**	0.112**
GC_{3s}						1	-0.775**	0.704**	0.569**
ENc							1	-0.878**	-0.736**
CAI								1	0.774**
Fop									1

** Correlation is significant at the 0.01 level (2-tailed).

Table-5. Correlation between amino acid usage indices.

	Axis1	Axis2	CAI	Gravy	Aromo	Mean Cost
Axis1	1	-0.122**	0.320**	-0.832**	-0.171**	-0.118**
Axis2		1	-0.357**	-0.184**	-0.428**	-0.433**
CAI			1	-0.081**	-0.060**	-0.129**
Gravy				1	0.080**	0.047*
Aromo					1	0.821**
Mean Cost						1

**Correlation is significant at the 0.01 level (2-tailed) and
*Correlation is significant at the 0.05 level (2-tailed).

Table-6. Number of protein coding genes on leading and lagging strand of replication.

Gene group		Total No. of genes (n)	Leading strand		Lagging strand	
			(n)	Percentage	(n)	Percentage
Protein coding genes	CAI≤0.40	566	303	53.53	263	46.47
	0.40<CAI≤0.65	1191	658	55.25	533	44.75
	CAI>0.65	267	147	55.06	120	44.94
Ribosomal		60	44	73.33	16	26.67
Urease		7	7	100	0	0

3.5. Selection pressure on codon usage

It was observed that correlation between Ka and CAI of orthologous genes was more significantly negative (r = -0.633,p<0.01) than correlation between Ks and CAI (r = -0.532**,** p<0.01). When the orthologous genes were sorted according to their Ks values, lowest Ks was possessed by highly expressed genes. Fop of orthologous genes was significantly negatively correlated with Ks (r = -0.446, p<0.01).

[IV] DISCUSSION

Species-specific codon usage as well as preference forsynonymous codons in prokaryotes are consequences of various important features such as nucleotide compositional constraint, mutational bias and translational selection of codons, etc. *C. urealyticum* genes biased towards high GC content with a little variation around mean value except small regions. High proportion of C_3 and G_3 than frequency of A_3 and T_3 ending codons suggested greater stability of C and G ending codons in contrast with A and T ending. Thus it could be assumed that compositional constraint was affecting codon usage variation in *C. urealyticum* genes [35]. Higher GC_3 content genes corresponded to lower ENc and a significant negative correlation between them showed strong influence of compositional constraint on codon usage bias among genes. In GC_3-ENc plot majority of points deposited below the expected curve towards GC_3 rich region, while points must be deposited on continuous curve if only compositional constrain was responsible for codon usage variation in genes [28]. So, apart from compositional constraint other factors like mutational bias, natural selection and translational selection were also affecting codon usage variation in *C. urealyticum* genes.

Correspondence analysis is a graphical illustration of multi-way contingency table consisting several rows and columns. It can generate multidimensional planes by calculating fractions corresponding to the number of columns [36]. CA on RSCU of *C. urealyticum* genes showed that fraction of first major axis (18.92%) was larger than second major axis (8.81%), which was indicating that axis 1 was primary factor for codon usage variations. Significant correlation between axis 1 and CAI indicated that codon usage bias in *C. urealyticum* genes was considerably affected by

gene expression. Strong negative correlation between CAI and ENc held the former prediction. Most of the genes concentrated towards origin of axes revealed that codon usage bias in *C. urealyticum* genes were homogeneous to some extent. Deposition of PHX and PLX genes along axis 1 revealed synonymous codon usage of these two types of genes were different and clustering strategy of PHX genes showed conservation of synonymous codons among PHX genes. Moreover, significant positive correlations between CAI and GC, GC_3, G_3 and Fop suggested that translational selection of codons was present in *C. urealyticum* genes [37]. Significant positive correlation between CAI and C_3 revealed C ending codon usage bias in PHX genes. Moreover, optimal codons of *C. urealyticum* genes was also biased towards C ending over G ending triplets. According to Moriyama and Powell RNY (R- Purine, N-any nucleotide base and Y-pyrimidine) codons were more beneficial for translation of genes. CA on amino acid usage revealed amino acid conservation among genes. PHX genes of *C. urealyticum* distinctly deposited from hydrophobic protein coding genes. Moreover, significant negative correlation between CAI and GRAVY as well as aromaticity revealed unwillingness of PHX genes to compriseof hydrophobic and aromatic protein coding genes as biosynthesis cost of hydrophobic as well as aromatic proteins are higher [14]. Protein synthesis cost was also inversely proportional to gene expression level, which revealed that PHX genes of *C. urealyticum* were very economical towards their expression. Similar expression strategy was also observed in Bifidobacterial PHX genes [38]. Nucleotide substitution rate (Ka and Ks) can estimate the nature of selection pressure, which affects genomic evolution of organisms [39]. In *C. urealyticum,* synonymous positions ofPHX genes wereless diverging and helped in stabilizing optimal codon usage in PHXgenes.

In silico detection of the origin of replication and terminus sites in a genome had already been well established by several experimental corroborations [40] and in present study, Z-curve distinguished the mentioned sites in *C. urealyticum* genome (Figure 4). According to Chargaff and his colleagues [41], nucleotide distribution between two complementary DNA strands is dissimilar. Furthermore, amount of genes is also different on both leading and lagging strands of replication [42]. In this study existence of greater number of genes as well as PHX genes and urease protein coding genes on leading strand than lagging strand of replication revealed that leading strand was enriched with essential genes than nonessential ones in *C. urealyticum*like other prokaryotes[43]. Lafay*et al*. has been discussed about 'replicational selection', which permits the occurrence of gene enrichment on leading strand to avoid collision between polymerases during replication and transcription at the same time in prokaryotes [44].

Urease coding genes of *C. urealyticum* play an important role in their pathogenicity [25]. In the present study urease coding genes showed higher expression level and optimal codon usage. They clustered along with PHX genes and were also reluctant to use hydrophobic amino acids, thus revealed translational selection pressure on codon and amino acid usage bias in urease coding genes of *C. urealyticum*. Replicational selection was also present in these gene expression. These genes are very important in amino acid transport and metabolism of *C. urealyticum*. In urinary tract of patients with urological diseases the pathogen takes the opportunity to break down urea by urease and make urine alkaline which further leads to form renal stones [45]. Thus codon usage study may further help to understand molecular basis of pathogenesis and basis of host-pathogen interactions in *C. urealyticum*.

[V] CONCLUSION

Codon usage variation among *C. urealyticum* DSM 7109 genes was mostly influenced by translational selection as well as natural selection. Biosynthesis cost of proteins as well as their hydrophobicity, aromaticity were also

regulating codon and amino acid usage variations in highly expressed genes. Strand specific gene expression was present in genes. Overall gene expression of *C. urealyticum*was moderate, whereas urease coding genes showed a higher degree of expression and codon usage bias towards GC rich codons. Synonymous positions were diverging less in highly expressed genes of *C. urealyticum*, thus conserved optimal codon usage in these genes. More study would be helpful to understand the gradual evolution of *C. urealyticum* DSM 7109genes and their pathogenicity.

ACKNOWLEDGEMENT

The Department of Biotechnology (DBT), Government of India, New Delhi is acknowledged gratefully for creation of Bioinformatics Infrastructure Facility Centre at Vidyasagar University, Midnapore, West Bengal, India.

REFERENCES

1. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M and Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, 43-74.
2. Romero, H., Zavala, A and Musto, H. (2000a). Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoebahistolytica*. Gene. 242, 307-311.
3. Karlin, S and Mrazek, J. (1996). What drives codon choices in human genes? J. Mol. Biol. 262, 459-472.
4. Moriyama, E.N. and Powell, J.R. (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26, 3188-3193.
5. Duret, L. (2000). tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16, 287-289.
6. Percudani, R., Pavesi, A and Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J. Mol. Biol. 268, 322-330.
7. Shi, X.F., Huang, J.F., Liang, C.R., Liu, S.Q., Xie, J and Liu, C.Q. (2001). Is there a close relationship between synonymous codon bias and codon–anticodon binding strength in human genes? Chinese Sci. Bulletin. 12, 1015-1019.
8. Romero, H., Zavala, A., Musto, H and Bernardi, G. (2003). The influence of translational selection on codon usage in fishes from the family *Cyprinidae*. Gene. 317, 141-147.
9. Sharp, P.M., and Li, W.H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28-38.
10. Romero, H., Zavala, A and Musto, H. (2000b). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. 28, 2084-2090.
11. Das, S., Paul, S., Chatterjee, S and Dutta, C. (2005). Codon and amino acid usage in two major human pathogens of genus *Bartonella*--optimization between replicational-transcriptional selection translational control and cost minimization. DNA Res. 12, 91-102.
12. Liu, G., Wu, J., Yang, H and Bao, Q. (2010). Codon usage patterns in *Corynebacterium glutamicum*: mutational bias natural selection and amino acid Conservation. Comp. Functional Genom. 1-7.
13. Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389-409.
14. Lobry, J.R. and Gautier, C. (1994). Hydrophobicity expressivity and aromaticity are the major trends of amino-

acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174-3180.
15. Singer G.A. and Hickey D.A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage amino acid composition and nucleotide content. Gene. 317, 39-47.
16. Oresic, M. and Shalloway, D. (1998). Specific correlations between relative synonymous codon usage and protein secondary structure. J. Mol. Biol. 281, 31-48.
17. Xie, T. and Ding, D. (1998). The relationship between synonymous codon usage and protein structure. FEBS Letters. 434, 93-96.
18. Gupta, S.K., Majumdar, S., Bhattacharya, T.K. and Ghosh, T.C. (2000). Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophysic. Res. Commun. 269, 692-696.
19. Gu, W., Zhou, T., Ma, J., Sun, X and Lu, Z. (2004). The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapien.*Biosyst. 73, 89-97.
20. Chamary, J.V., Parmley, J.L. and Hurst, L.D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics. 7, 98-108.
21. Kahali, B., Basak, S and Ghosh, T.C. (2007). Reinvestigating the codon and amino acid usage of *S. cerevisiae* genome: a new insight from protein secondary structure analysis. Biochem. Biophysic. Res. Commun. 354, 693-699.
22. Soriano, F., Ponte, C., Santamaria, M., Castilla, C and Roblas, R.F. (1986). In Vitro and In Vivo study of stone formation by *Corynebacterium* group D2 (*Corynebacterium urealyticum*). J. Clinical Microbiol. 23, 691-694.
23. Elad, D., Aizenberg, I., Shamir, M., Soriano, F and Shlomovitz, S (1992). Isolation of *Corynebacterium* Group D2 from two dogs with urinary tract infections. J. Clinical Microbiol. 30, 1167-1169.
24. Soriano, F., Aquado, J.M., Ponte, C., Fernandez-Roblas, R and Rodriguez, J.L. (1990). Urinary tract infection caused by *Corynebacterium* group D2: report of 82 cases and review. Rev. Infectious Diseases. 12, 1019-1034.
25. Tauch, A., Trost, E., Tilker, A., Ludewig, U., Schneiker, S., Goesmann, A., Arnold, W and Bekel, T. (2008). The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. J. Biotechnol. 31, 11-21.
26. John, F and Peden. (1999). Analysis of Codon Usage. Ph. D. Thesis The University of Nottingham, United Kingdom.
27. Xia, X. (2013). DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol. 30, 1720-1728.
28. Wright, F. (1990). The 'effective number of codons' used in a gene. Gene. 87, 23-29.
29. Liangwei, L., Linmin, W., Zhang, Z., Suya, W and Hongge, C. (2012). Effect of codon message on xylanase thermal activity. Rev. Infectious Diseases. 287, 27183-27188.
30. Soohyun, L. and Seyeon, W. (2010). Relative codon adaptation index a sensitive measure of codon usage bias. Evol. Bioinformatics. 6, 47-55.
31. Sharp, P.M. and Li, W.H. (1987). The codon adaptation index- a measure of directional synonymous codon usage bias and its potential application. Nucleic Acids Res. 15, 1281-1295.
32. Goldstein, R. (1991). Statistical computing software reviews. The American Statistician. 45, 305-311.
33. Benzecri, J.P. (1992). Correspondence Analysis Handbook. Marcel Dekker, New York.
34. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Appl. BioSci. 13, 555-556.

35. Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F and Osawa, S. (1987). The ribosomal protein gene cluster of *Mycoplasma capricolum*. Mol. Gen. Genet. 210, 314-322.
36. Nenadic, O. and Greenacre, M. (2007). Correspondence analysis in R with two- and three-dimensional graphics: The CA package. J. Stat. Software. 20, 3.
37. Sharp, P.M., Stenico, M., Peden, J.F. and Lloyd, A.T. (1993). Codon usage: mutational bias translational selection or both? Biochem. Soc. Trans. 21, 835-841.
38. Roy, A., Mukhopadhyay, S., Sarkar, I and Sen, A. (2015). Comparative investigation of the various determinants that influence the codon and amino acid usage patterns in the genus *Bifidobacterium*. World J. Microbiol. Biotechnol. 31, 959-981.
39. Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M and Nielsen, R. (2007). Genes under positive selection in *Escherichia coli*. Genome Res. 17, 1336-1343.
40. Sernova, N.V. and Gelfand, M.S. (2008). Identification of replication origins in prokaryotic genomes. Brief Bioinform. 9, 376-391.
41. Rudner, R., Karkas, J.D. and Chargaff, E. (1968). Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc. Natl. Acad. Sci. USA. 60, 921-922.
42. Xizeng, M., Han, Z., Yanbin, Y and Ying, X. (2012). The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res. 40, 8210-8218.
43. Rocha, E.P. and Danchin, A. (2003). Essentiality not expressiveness drives gene-strand bias in bacteria. Nat. Genet. 34, 377-378.
44. Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M and Wolfe, K.H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand specific mutational biases. Nucleic Acids Res. 27, 1642-1649.
45. Soriano, F and Tauch, A. (2008). Microbiological and clinical features of *Corynebacterium urealyticum*: urinary tract stones and genomics as the Rosetta Stone. Clin. Microbiol. Infect. 14, 632-643.