Int. J. Chem. Sci.: 10(4), 2012, 2225-2236
ISSN 0972-768X
www.sadgurupublications.com

ASSESSMENT OF CODON USAGE PATTERN IN NITROGEN-FIXING UNCULTURED MARINE CYANOBACTERIUM UCYN-A

RATNA PRABHA[a], DHANANJAYA P. SINGH[*a] and ARUN K. SHARMA

National Bureau of Agriculturally Important Microorganisms, Indian Council of Agricultural Research,
Kushmaur, MAUNATH BHANJAN – 275101 (U.P.) INDIA
[a]Department of Biotechnology, Mewar University, Gangrar, CHITTORGARH (Raj.) INDIA

ABSTRACT

In recent years, certain cyanobacteria are reported to lack functional PS II of photosynthesis. These unicellular Nitrogen-fixing cyanobacteria termed as UCYN-A are yet not cultured in laboratory and are major contributor to the nitrogen input of marine ecosystem. Genome sequence of one such cyanobacterium is available, for which codon selection pattern is analysed using various in-silico and statistical techniques. It is identified that T and/or A ending codons were predominant. Compositional mutational bias was observed to be one of the major factors responsible for codon bias among the genes. A strong positive correlation between effective number of codons (Nc) and GC_{3S} content was also observed showing that the codon usage was affected by gene nucleotide composition. Translational selection and gene length were observed to have a minor role in influencing the codon usage variation among the genes. Multivariate statistical analysis identified that there is not any major explanatory axis in this organism. A set of 16 codons were determined as the 'optimal codons' using χ^2 test ($P < 0.01$). Thus, above analysis reveals that compositional constraints plays major role in selecting codons in UCYN-A genome which is also affected by translational selection, although in low amount.

Key words: Codon, Nitrogen fixation, Cyanobacterium UCYN-A.

INTRODUCTION

Cyanobacteria are the prokaryote, which trap solar energy, utilizes it for performing photosynthetic activity and nitrogen fixation on earth. They represents diverse group with multiple habitats from aquatic (marine and freshwater) environments to terrestrial and

[*]Author for correspondence; E-mail: dpsfarm@rediffmail.com, Ph.: +91-547-2530080, Fax.: +91-547-2530358

symbiotic habitats[1,2]. Cyanobacteria UCYN-A is a globally distributed species of nitrogen-fixing marine cyanobacteria. It drives attention of scientists worldwide, for their novel metabolism that lacks the oxygen producing PS II complex of the photosynthetic apparatus along with the enzymes of the Calvin and tricarboxylic acid cycles, and amino acid biosynthesis[3,4]. It has the smallest genome size in comparison to other cyanobacteria, which are sequenced till date (Genome Database, NCBI, Sept. 2012). Genome sequence which is available for this cyanobacterium was analysed to assess its codon usage pattern. *In-silico* analysis and statistical methods were implemented to determine the codon selection patterns in genes of cyanobacterium UCYN-A to establish differential level of gene expressions in the genomes. Present study aims to predict whether this organism follows any specific pattern of codon usage or not. Characterizing the codon usage pattern in a specific organism is useful due to several reasons as it can help in understanding the basics of molecular biology of that organism in practical and theoretical terms, both[5].

Protein translation is an universal phenomenon and the genetic code implied for the same is degenerate as multiple codons can code for one amino acid. There are 64 codons to represent 20 standard amino acids and the translation termination signal[6]. There are synonymous codons, which codes for a single amino acid and are well conserved over most species with few exceptions[7,8]. Five synonymous families (SF), designated by SF types 1, 2, 3, 4 and 6 are available for any gene using the universal code where the SF type implies about available codon choice for members of that particular family[9].

Biased usage of alternative synonymous codon usages (SCU) can be observed in most protein-coding genes which are non-random and species-specific. Codon usage pattern is affected by various factors, major among them are mutational bias, GC content bias, translation efficiency, natural selection, gene length, tRNA abundance and up to some extent horizontal gene transfer[10-13]. Even significant variation is observed in codon usage bias among different genes within the same organism[14-16].

EXPERIMENTAL

Genome sequence

Available genome sequence of cyanobacterium UCYN-A was taken from Genome database of NCBI (http://www.ncbi.nlm.nih.gov/genome/). It has a genome size of 1.44 Mb and consists of 1241 genes, out of which 1199 are protein coding. Genes sequence less than 80 codons, hypothetical and/or with intermediate termination codons were excluded for the data analysis, for minimization of sampling errors. Final dataset consist of 858 gene sequences.

Measures of Codon selection pattern

For all genes, Relative Synonymous Codon Usage (RSCU)[17], GC_{3s} (Frequency of GC at the synonymous third position), A_3, T_3, G_3, C_3 (Frequency of A, T, G, C at third position), codon adaptation index (CAI)[18] and Effective number of codons (Nc)[9] was calculated. The programs General Codon Usage Analysis i.e. GCUA[19] and CodonW version 1.4.2 (http://codonw.sourceforge.net/) were employed for calculating the codon usage indices.

Statistical methods

Multivariate statistical analysis was carried out to identify possible source of variation. In this, all the genes are plotted in a 59-dimensional hyperspace according to their usage of 59 informative codons (excluding Met, Trp, and stop codons). Correspondence Analysis of RSCU values was carried out to reduce dimensionality of visualization space of genes and determination of major sources of variation among synonymous codons. Furthermore, bivariate correlation analysis was carried out to identify correlation between various indices. SPSS 16.0 was implemented for statistical analysis.

RESULTS AND DISCUSSION

Codon bias, base composition and RSCU

UCYN-A is a GC poor genome (GC content 31.10%). Most preferentially used codons in UCYN-A are T-ended or A-ended codons in which 12 are T-ended and 6 are A-ended (Table 1). Codons ending with T (U) and/or A are predominately used by UCYN-A than the codons ending with G and/or C. It is expected that due to compositional constraint T- and/or A-ending codons were preferentially used in this genome but the reason for codon selection pattern is not very clear because the RSCU values are not sufficient to provide complete picture of factors governing codon selection pattern of any organism[10,20]. RSCU is defined as the ratio of the observed frequency of codons to the expected frequency if all the synonymous codons for those amino acids are used equally. RSCU values greater than 1.0 indicate that the corresponding codons are used more frequently than the expected frequency whereas the reverse is true for RSCU value less than 1.0[17].

Table 1: Overall codon usage data of cyanobacterium UCYN-A genes

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Ala	GCU	9423	1.98	Ser	AGU	5416	1.47
	GCC	1715	0.36		AGC	1824	0.5
	GCA	6849	1.44		UCU	8374	2.28
	GCG	1009	0.21		UCC	1140	0.31
Asp	GAU	12339	1.66		UCA	4560	1.24
	GAC	2550	0.34		UCG	726	0.2
Glu	GAA	15960	1.62	Tyr	UAU	8541	1.61
	GAG	3778	0.38		UAC	2042	0.39
Gly	GGU	6663	1.31	Gln	CAA	10280	1.61
	GGC	2029	0.4		CAG	2529	0.39
	GGA	9679	1.9	Asn	AAU	13513	1.59
	GGG	1960	0.39		AAC	3485	0.41
Cys	UGU	2646	1.55	Pro	CCU	7081	2.29
	UGC	776	0.45		CCC	887	0.29
Phe	UUU	10743	1.66		CCA	3882	1.25
	UUC	2233	0.34		CCG	530	0.17
His	CAU	4961	1.66	Val	GUU	9404	1.98
	CAC	1017	0.34		GUC	1784	0.38
Ile	AUU	16714	1.75		GUA	6691	1.41
	AUC	3378	0.35		GUG	1124	0.24
	AUA	8569	0.9	Thr	ACU	8385	2
Lys	AAA	17615	1.61		ACC	1610	0.38
	AAG	4300	0.39		ACA	5949	1.42
Leu	UUA	17480	3.06		ACG	864	0.21
	UUG	4148	0.73	Arg	CGU	4419	1.98
	CUU	5633	0.99		CGC	1178	0.53
	CUC	1009	0.18		CGA	1579	0.71
	CUA	5012	0.88		CGG	331	0.15
	CUG	1023	0.18		AGA	4919	2.21
Ter	UAA	586	0		AGG	958	0.43
	UGA	90	0				
	UAG	182	0				

AA represents amino acid; N is the number of codons; RSCU represents relative synonymous codon usage.

Role of various factors in codon usage bias

Nc Plot

Nc and GC_{3s} have been widely used to detect the codon usage variation among the genes[21,11]. Nc can take values from 20 to 61 when only one codon or all synonyms in equal frequencies were used per amino acid, respectively[9]. The sequences in which Nc values are < 30 are highly expressed and more biased while those with > 55 are poorly expressed genes[22,11]. Wright[9] suggested that genes, whose codon choice is constrained only by a G + C mutational bias, will lie on or just below the curve of the predicted value in the Nc plot (a plot of Nc versus GC_{3s}). In order to understand the codon usage variation across the genome of UCYN-A, Nc and GC_{3s} values were calculated. Nc value in the UCYN-A ranges from 30.93 to 61 with a mean value of 41.10 and standard deviation (SD) of 3.77. Nc value of all genes of UCYN-A are much higher (> 40), thus the codon usage bias in UCYN-A genome is low and there is a lesser variation in codon usage pattern among different UCYN-A genes (SD = 3.77).

In Nc plot (Fig. 1), majority of genes fall on or just below the expected curve towards AT rich regions which certainly originate from extreme compositional constraints. This is also evident from the fact that AT rich genes have lower Nc values. However, considerable numbers of points show deviation from the expected curve, especially those with low Nc values suggesting that majority of genes have an additional codon usage bias apart from compositional bias. Furthermore, most of the genes fall within a restricted cloud at a relatively narrow range of GC_{3s} between 0.1 to 0.2 with large variation of Nc values ranging between 30 to 45 (Fig. 1).

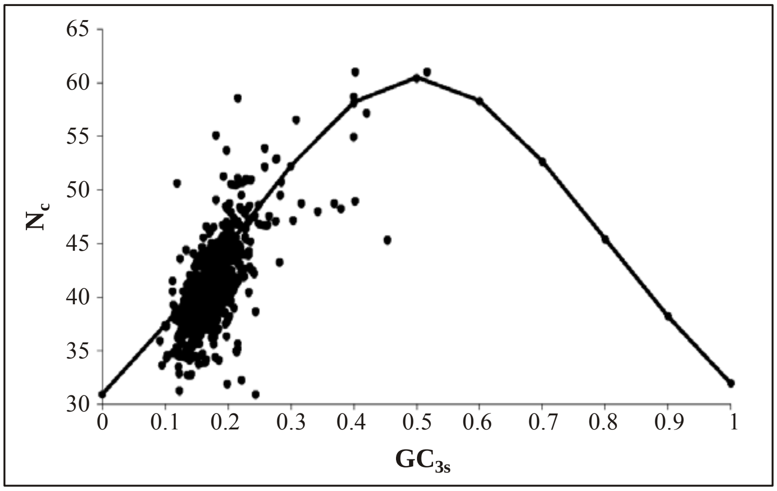

Fig 1: Nc plot of cyanobacterium UCYN-A

This suggests that translational selection is also responsible for codon bias among the genes. However, strong influence of compositional constraints on codon usage bias in the genes could be understood from the presence of significant positive correlation between GC3s and N_c ($r = -0.674$, $P < 0.001$).

Continuous curve represents the expected curve between GC_{3s} and Nc under random codon usage.

Selective use of base at third codon position

In order to examine the base composition variation among different genes, the base composition of different protein-coding genes was calculated and correlation of the frequencies of A, T, G and C at the third position against Nc values was estimated as shown in Table 2. The frequency of T and A at the third codon position increased with decreasing Nc values whereas those of G and C increased. It can be assumed that the influence of mutational bias of these genes is reflected in the choice of bases at the third position. This is expected since the optimal codons are in general, chosen in accordance with the mutational bias of these genes[23].

Table 2: Correlation coefficient values of A, T, G, C at third position with N_c

	T3s	C3s	A3s	G3s
Nc	-0.477**	0.524**	-0.249**	0.489**

**Represents significant correlation with probability, $P < 0.001$

Codon bias and gene length

Selection for translational accuracy is predicted to have a positive correlation between codon bias and gene length. From the plot drawn with gene length against Nc (Fig. 2), it is understood that shorter genes have much wider variance in Nc values and vice-versa for longer genes. Lower Nc values for longer genes may be due to the direct effect of translation time or to extra energy cost of proofreading associated with longer translating time.

Correlation analysis of gene length against Nc, GC_{3s} and axis 1 is shown in Table 3. A negative correlation was observed with gene length against Nc ($r = -0.086$, $P < 0.005$) revealing that gene length influences codon usage of these genes. Eyre-Walker[24] has reported that the selection for accuracy in protein translation is likely to be greater in longer genes because the cost of producing a protein is proportional to its length. In this study, the results of correlation analyses between gene length and the genes positions on axis 1 (r = 0.109, P < 0.001) showed significant correlation. The findings indicated that more biased genes with short length and lower GC_{3s} values are distributed at the left side of the first axis and vice-versa for longer genes. This suggests the role of gene length on codon selection pattern. While comparing with correlation coefficients of nucleotide composition, this is quite small and thus indicated compositional constraints as major factor of codon usage variation where the gene length seemed to play a minor role.

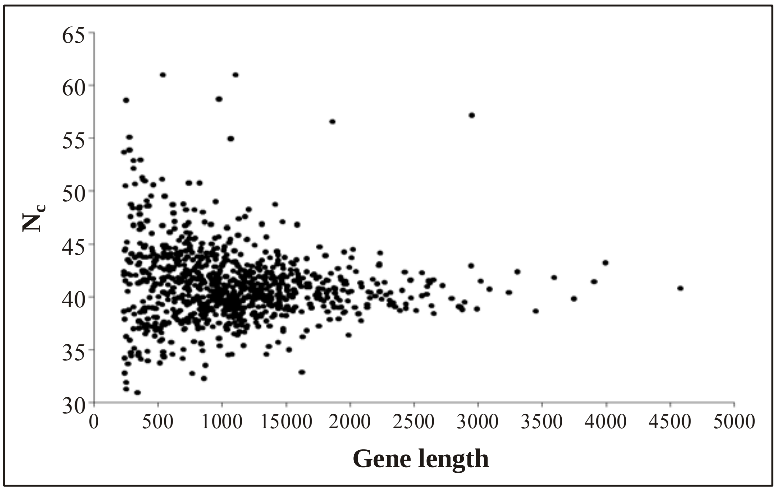

Fig 2: Plot of Nc versus gene length

Table 3: Correlation coefficient values gene length with axis 1, Nc and GC_{3s}

	Axis 1	Nc	GC_{3s}
Gene length	-0.109**	-0.086*	0.070*

**Represents significant correlation with probability P < 0.001;
*Represents significant correlation with probability P < 0.005

Correspondence analysis using RSCU values

To explore various factors that influence variation in codon usage among the genes, correspondence analysis was conducted on their RSCU values. In cyanobacterium UCYN-A, no major explanatory axis was identified. However, distribution of the genes along the first two major axes were considered because these accounted for 6.57% and 5.94% of the total variation, which is much greater when compared to rest of axes. Distribution of all the genes along these two axes i.e. Axis 1 and Axis 2 is shown in Fig. 3.

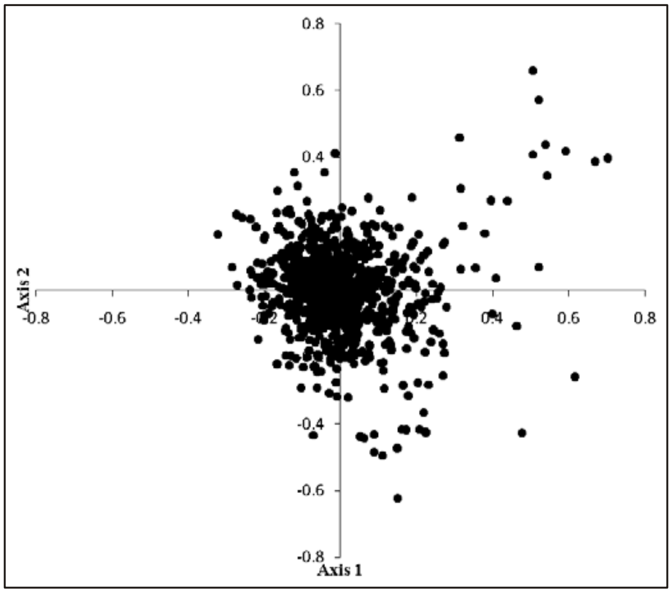

Fig. 3: Scatter plot of genes positions on Axis 1 and Axis 2

Bivariate correlation analysis was carried to identify the relation between various indices of codon usage variation with Axis 1 and Axis 2 (Table 4). The first major axis is correlated negatively with T_{3S} but positively with A_{3S} and G_{3S}. No significant correlation was observed between Axis 1 and C_{3S} (Table 4). Also, strong positive correlation exists between position of genes along the first axis with Nc and high degree of positive correlation with GC_{3s}. These findings suggest that highly biased genes, those ending with T are clustered on the negative side, whereas the codons ending in A and G predominate on the positive side of the first major axis.

Table 4: Correlation analysis of A, T, G, C, at third position, Nc values with axis 1 and axis 2

	N_c	A_{3s}	T_{3s}	G_{3s}	C_{3s}	GC_{3s}	GC
Axis 1	0.240^{**}	0.291^{**}	-0.493^{**}	0.432^{**}	0.004^{NS}	0.247^{**}	-0.314^{**}
Axis 2	0.607^{**}	-0.474^{**}	-0.269^{**}	0.447^{**}	0.644^{**}	0.723^{**}	0.437^{**}

*Represents significant correlation with probability,
$P < 0.001$ and ^{NS} represent Non-significant

Additionally, significant positive correlation is observed with Nc against GC_{3s} ($r = 0.674$, $P < 0.001$). It is found that highly expressed genes tend to use "T" or "A" at the synonymous positions as compared to lowly expressed genes and T-ending codons are preferred over A-ending codons in highly expressed genes. We therefore, speculate that in UCYN-A compositional mutation bias plays an important role in shaping the genome of these genes.

Translational optimal codons

Furthermore, to investigate the difference between high and low expressed genes, the codon usage variation between 10% of the genes located at the extreme right of major axis and 10% of the genes located towards the extreme left produced by CA using RSCU were compared. Chi square contingency test of the two groups was performed to estimate the optimal codons i.e. synonymous codons frequently used in highly expressed genes. Codons whose frequency of usage were significantly higher ($P < 0.01$) in highly expressed genes than the genes with low level of expression were determined as optimal codons. RSCU values for each codon for the highly and lowly expressed genes are shown in Table 5. The asterisk represents the codons whose occurrences are significantly higher in the genes situated on the extreme left side of axis 1, compared to the genes present on the extreme right of the first major axis. 16 codons were determined as the 'optimal codons', which were significantly more frequent among the highly expressed genes using χ^2 test at $P < 0.01$. Maximum of these have T at the third position. Out of 16 codons, there are 10 T-ending, 3 A-ending and rest 3 C ending.

Table 5: RSCU for the highly and lowly expressed genes highlighting translational optimal codons

AA	Codon	RSCU^1	N^1	RSCU^2	N^2	AA	Codon	RSCU^1	N^1	RSCU^2	N^2
Phe	UUU	1.42	(257)	1.51	(534)	Ser	UCU*	2.59	(248)	1.78	(264)
	UUC	0.58	(104)	0.49	(171)		UCC	0.49	(47)	0.47	(69)
Leu	UUA*	2.65	(346)	2.29	(599)		UCA	0.72	(69)	1.33	(197)
	UUG	0.47	(61)	1.02	(267)		UCG	0.09	(9)	0.37	(55)
	CUU*	1.61	(210)	0.86	(226)		AGU	1.6	(153)	1.27	(188)
	CUC	0.15	(20)	0.37	(98)		AGC	0.5	(48)	0.78	(116)
	CUA	0.97	(127)	0.96	(252)	Pro	CCU*	2.7	(265)	1.66	(145)
	CUG	0.14	(18)	0.49	(129)		CCC	0.23	(23)	0.49	(43)
Ile	AUU*	1.89	(412)	1.5	(615)		CCA	1.01	(99)	1.48	(129)
	AUC*	0.58	(127)	0.41	(169)		CCG	0.06	(6)	0.37	(32)
	AUA	0.52	(114)	1.08	(443)	Thr	ACU*	2.19	(285)	1.72	(252)

Cont…

AA	Codon	RSCU^1	N^1	RSCU^2	N^2	AA	Codon	RSCU^1	N^1	RSCU^2	N^2
Val	GUU*	2.16	(346)	1.64	(282)		ACC	0.59	(77)	0.41	(60)
	GUC	0.36	(57)	0.47	(81)		ACA	1.13	(147)	1.46	(214)
	GUA	1.38	(222)	1.35	(232)		ACG	0.08	(11)	0.41	(60)
	GUG	0.11	(17)	0.54	(93)	Ala	GCU*	2.29	(395)	1.65	(199)
Tyr	UAU	1.4	(195)	1.47	(398)		GCC	0.25	(43)	0.44	(53)
	UAC	0.6	(84)	0.53	(144)		GCA	1.32	(227)	1.61	(194)
His	CAU	1.6	(121)	1.5	(169)		GCG	0.14	(24)	0.31	(37)
	CAC	0.4	(30)	0.5	(56)	Arg	CGU*	3.89	(234)	0.38	(29)
Gln	CAA	1.55	(240)	1.44	(317)		CGC*	0.61	(37)	0.32	(24)
	CAG	0.45	(70)	0.56	(123)		CGA	0.47	(28)	0.53	(40)
Asn	AAU	1.36	(311)	1.51	(459)		CGG	0.15	(9)	0.25	(19)
	AAC*	0.64	(148)	0.49	(147)		AGA	0.86	(52)	3.26	(248)
Asp	GAU	1.63	(397)	1.56	(287)		AGG	0.02	(1)	1.27	(97)
	GAC	0.37	(89)	0.44	(82)	Lys	AAA*	1.64	(518)	1.46	(621)
Glu	GAA*	1.66	(536)	1.43	(371)		AAG	0.36	(115)	0.54	(232)
	GAG	0.34	(110)	0.57	(149)	Cys	UGU*	1.57	(85)	1.28	(132)
Gly	GGU*	1.69	(311)	1.19	(179)		UGC	0.43	(23)	0.72	(74)
	GGC	0.43	(80)	0.46	(69)	Trp	UGG	1	(89)	1	(216)
	GGA	1.62	(299)	1.85	(279)	TER	UAA	2.07	(29)	1.35	(218)
	GGG	0.26	(47)	0.5	(75)		UAG	0.71	(10)	0.98	(158)
Met	AUG	1	(233)	1	(262)		UGA	0.21	(3)	0.66	(107)

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

CONCLUSION

Compositional constraint is most dominant factor in codon selection pattern of the cyanobacterium UCYN-A. However, translational selection also plays a minor role in shaping the codon usage variation among the genes in this organism. This study reveals that

T/A-ending codons are preferred over G/C-ending codons in highly expressed genes. Number of codons in highly expressed genes was much higher than those in lowly expressed genes and gene length also played a minor role in codon usage bias. A set of sixteen codons were determined as the optimal codons which were significantly more frequent among the highly expressed genes in χ^2 test at $P < 0.01$. Maximum of these codons were T-ending, thus confirming compositional constraint as major factor in codon selection pattern of this organism.

ACKNOWLEDGEMENT

Financial support from Indian Council of Agricultural Research, India in the form of a “National Agricultural Innovation Project” (NAIP) entitled “Establishment of National Agricultural Bioinformatics Grid” (NABG) is gratefully acknowledged.

REFERENCES

1. R. W. Hess, Current Opinion in Microbiology, **14**, 5 (2011).
2. J. Mrázek, D. Bhaya, A. R. Grossman and S. Karlin, Nucleic Acids Research, **29**, 7 (2001).
3. H. J. Tripp, S. R. Bench, K. A. Turk, R. A. Foster, B. A. Desany, F. Niazi, J. P. Affourtit and J. P. Zehr, Nature, **464**, 7285 (2009).
4. H. Bothe, H. J. Tripp and J. P. Zehr, Arch Microbiol, **192**, 10 (2010).
5. A. T. Lloyd and P. M. Sharp, Nucleic Acids Research, **20**, 20 (1992).
6. S. Osawa and T. H. Jukes, J. Molecular Evolution, **28**, 4 (1989).
7. S. Osawa, T. H. Jukes, K. Watanabe and A. Muto, Microbiol. Rev., **56**, 1 (1992).
8. M. A. Santos, G. Moura, S. E. Massey and M. F. Tuite, Trends Genet, **20**, 2 (2004).
9. F. Wright, Gene **87**, 1 (1990).
10. S. K. Gupta, T. K. Bhattacharyya and T. C. Ghosh, Indian J. Biochem. Biophy., **39**, 1 (2002).
11. S. K. Gupta, T. K. Bhattacharyya and T. C. Ghosh, J. Biomolecular Structure and Dynamics, **21**, 4 (2004).
12. F. Supek and K. Vlahoviček, BMC Bioinformatics, **6**, 182 (2005).
13. X. F. Wan, D. Xu, A. Kleinhofs and J. Zhou, BMC Evolutionary Biol., **4**, 19 (2004).

14. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone and R. Mercier, Nucleic Acids Research, **9**, 1 (1981).

15. M. Gouy and C. Gautier, Nucleic Acids Research, **10**, 22 (1982).

16. T. Ikemura, Mol Biol. Evol., **2**, 1 (1985).

17. P. M. Sharp and W. H. Li, J. Mol. Evol., **24,** 1 (1986).

18. P. M. Sharp and W. H. Li, Nucleic Acids Research, **15**, 3 (1987).

19. J. O. McInerney, Bioinformatics, **14**, 4 (1998).

20. S. Hassan, V. Mahalingam and V. Kumar, Advances in Bioinformatics (2009).

21. K. Sahu, S. K. Gupta, T. C. Ghosh and S. Sau, J. Biochem. Molecular Biol., **37**, 4 (2004).

22. P. M. Sharp and E. Cowe, Yeast, **7**, 7 (1991).

23. A. Pan, C. Dutta and J. Das, Gene, **215**, 2 (1998).

24. A. Eyre-Walker, Mol. Biol. Evol., **13**, 6 (1996).

Revised : 09.11.2012 *Accepted : 12.11.2012*