

Synonymous codon usage in *Thermosynechococcus elongatus* (cyanobacteria) identifies the factors shaping codon usage variation

Ratna Prabha^{1,4}, Dhananjaya P Singh^{1*}, Shailendra K Gupta², Samir Farooqi³ & Anil Rai³

¹National Bureau of Agriculturally Important Microorganisms, Indian Council of Agricultural Research, Kushmaur, Maunath Bhanjan 275101, India; ²CSIR-Indian Institute of Toxicology Research, 80, Mahatma Gandhi Marg, Kaisarbagh, Lucknow 226001, India; ³Indian Agricultural Statistical Research Institute, Indian Council of Agricultural Research, Pusa, New Delhi 110 012, India; ⁴Department of Biotechnology, Faculty of Science and Technology, Mewar University, Gangrar, Chittorgarh, Rajasthan, India; Dhananjaya P Singh – Email: dpsfarm@rediffmail.com; *Corresponding author

Received May 28, 2012; Accepted June 7, 2012; Published July 06, 2012

Abstract:

Analysis of synonymous codon usage pattern in the genome of a thermophilic cyanobacterium, *Thermosynechococcus elongatus* BP-1 using multivariate statistical analysis revealed a single major explanatory axis accounting for codon usage variation in the organism. This axis is correlated with the GC content at third base of synonymous codons (GC3s) in correspondence analysis taking *T. elongatus* genes. A negative correlation was observed between effective number of codons i.e. Nc and GC3s. Results suggested a mutational bias as the major factor in shaping codon usage in this cyanobacterium. In comparison to the lowly expressed genes, highly expressed genes of this organism possess significantly higher proportion of pyrimidine-ending codons suggesting that besides, mutational bias, translational selection also influenced codon usage variation in *T. elongatus*. Correspondence analysis of relative synonymous codon usage (RSCU) with A, T, G, C at third positions (A3s, T3s, G3s, C3s, respectively) also supported this fact and expression levels of genes and gene length also influenced codon usage. A role of translational accuracy was identified in dictating the codon usage variation of this genome. Results indicated that although mutational bias is the major factor in shaping codon usage in *T. elongatus*, factors like translational selection, translational accuracy and gene expression level also influenced codon usage variation.

Key words: Thermophile, *Thermosynechococcus elongatus*, codon usage, cyanobacteria

Background:

Most amino acids in the organisms are coded by more than one triplet of nucleotides which are called synonymous codons that usually differ by single nucleotide in the third codon position but for some amino acids, it differs in the second position also. Codon usage is specially a genome strategy in which each genome has a particular codon usage signature that reflects particular evolutionary forces acting within that genome [1]. Codon usage pattern reflects a distinct pattern between [2-5] and within prokaryotic and eukaryotic organisms [6-9]. With time, different pattern in codon usage in bacterial genomes

indicates lateral gene transfer events and the analysis of codon bias helps establish the fact that the horizontal gene transfer is a major evolutionary force [10]. Variation in codon usage bias depends on various factors such as gene expression level [11, 12] with highly expressed genes showing greater codon usage bias than lowly expressed genes [13], amino acid composition [14], gene length [15] and tRNA abundance [16, 17]. Within the genomes which are the source of valuable biological information [18], codon usage pattern has both practical and theoretical importance in understanding the molecular basics of biology. This strategy can directly be utilized in molecular

characterization of species and can be used as a molecular marker for the characterization of molecular evolution. DNA primers and oligonucleotide probes can also be designed on the basis of this knowledge [19]. Natural selection acting through external environmental factors also influence genomic pattern of codon usage [20], variations in which can quantitatively predict the level of gene expressivity [21].

Among bacteria, primary factor responsible for codon usage variation is the differences in mutational biases. However, many species possess additional variation among genes that is consistent with the natural selection [22]. Among organisms, codon bias is maintained by a balance between selection, mutation and genetic drift [16, 23] where translational accuracy appears to be a major responsible factor as certain preferred codons are translated more accurately and efficiently [24]. *Thermosynechococcus elongatus* BP-1 is unicellular rod-shaped thermophilic cyanobacterium with an optimum growth temperature of 55°C. During the course of evolution, *T. elongatus* is branched very near to the origin of cyanobacteria [25]. Because of its ability to undergo natural transformation, this obligate photoautotroph has widely been used as a model organism for the study of photosynthesis. We analyzed codon usage pattern of whole genome of *T. elongatus* in order to understand the complete genetic organization and the factors shaping the codon usage pattern in this organism.

Methodology:

Sequence source and parameters

Complete genome and coding sequences of *T. elongatus* was obtained from NCBI (www.ncbi.nlm.nih.gov/genbank/genomes). Whole genome consists of 2476 genes. To minimize sampling error, only coding sequences (CDS) with correct initial and terminal codons and minimum 100 codons were considered. CDS with uncertain annotation or annotated as hypothetical were excluded by using PERL scripts.

Indices of codon usage bias

Frequency of 59 codons for 18 amino acids (except Met, Trp and stop codon) was examined by using different codon indices such as relative synonymous codon usage (RSCU) [26]; effective number of codons (Nc) [27]; codon adaptation index (CAI) [28]; frequency of G+C in a coding sequence (GC content); frequency of G+C at third position of codons (GC3s) and frequency of A, T, G, C at the synonymous 3rd positions of codons (A3, T3, G3, C3 respectively).

RSCU is the ratio of the observed frequency of codons to the expected frequency given that all the synonymous codons for the same amino acids are used equally. RSCU values >1 indicates more frequent use of corresponding codons than the expected frequency whereas the reverse is true for RSCU value <1 [26]. Nc measures the bias of synonymous codons and Nc ranges from 20 (when only one codon is used per amino acid) to 61 (when all codons are used in equal probability) [27].

CAI is used to estimate the extent of bias towards codons that were known to be preferred in highly expressed genes. CAI value ranges from 0 to 1.0, and a higher value indicates a stronger codon usage bias and higher expression level. CAI is a well accepted measure of gene expressivity [28].

These indices were calculated using GCUA [29] and CodonW version 1.4.2 [30].

Multivariate analysis

Correspondence analysis (CA) has been successfully used to investigate the variation of RSCU values among genes. CA plots all the genes in a 59-dimensional hyperspace according to their usage of the 59 informative codons, excluding Met, Trp, and termination codons. It identifies the axes that represent the most prominent factors contributing to the variation among genes. Statistical analysis was performed with SPSS version 16.0.

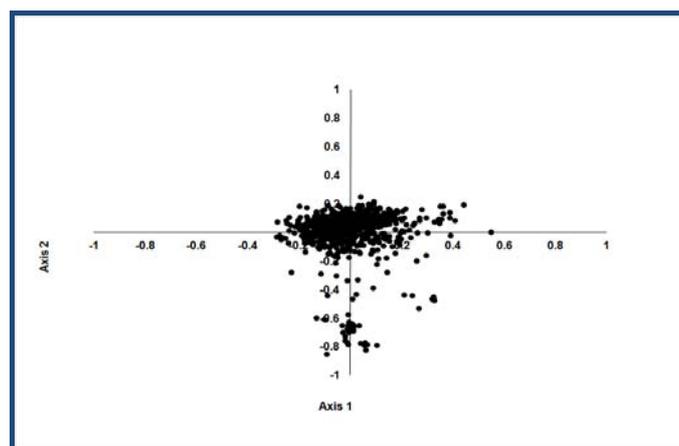


Figure 1: Position of *T. elongatus* genes along the two major axes of variation in the correspondence analysis on RSCU.

Discussion:

Overall codon usage analysis

T. elongatus genome contains moderately high (53%) GC content and therefore, due to compositional constraint, it is expected that there will be no or very little biasness in distribution of G-, C-, A- and T- ending codons. Among 18 most widely used codons, 8 are GC-ending (6 C-ending and 2 G-ending) whereas 10 codons are AT-ending (7 T- ending and 3 A-ending) **Table 1** (see supplementary material).

Nc and GC3s was used to further analyze the degree of heterogeneity in codon usage in *T. elongatus*. It was observed that Nc values for the genes range from 34.77 to 61 with a mean value of 49.17 and standard deviation of 3.65. This reflected a marked variation of codon usage in the genes. Analysis with GC3s values also confirmed the heterogeneity of synonymous codon usage among genes of *T. elongatus*. Strong influence of compositional constraints on codon usage bias in the genes could be understood from the presence of significant negative correlation ($r = -0.558$, $P < 0.01$) between GC3s and Nc. These results supported our hypothesis that compositional constraints positively influenced detection of codon usage variation in the genes in this organism.

Nc plot (plot of Nc vs GC3 content) reflected the determinants of the codon usage variation among genes in different organisms. Since Nc is constrained by G+C content, it is plotted against GC3s of the gene to investigate patterns of codon usage. It has been suggested that if GC3 content was the only determinant of the codon usage variation among the genes, the Nc value would fall on the continuous curve between Nc value and GC3 content [27]. In the Nc plot of *T. elongatus* (**Figure 1**),

majority of genes with higher GC3s have lower Nc value and except few genes, majority of genes lay far below the expected curve suggesting additional codon usage bias. This again indicated certain additional factors along with compositional constraint in shaping the codon usage bias of this genome. Translational selection therefore, also seems to affect codon usage because of wide range of Nc values for the same GC3 content.

Multivariate statistical analysis

To examine whether amino acid composition exerts any constraint on synonymous codon usage, we performed CA on RSCU values. The first axis accounted for 17.99% of total variation as compared to 7.77% for the second axis. A much higher value for the first axis in comparison to the second axis indicated a single major explanatory axis that accounted for the codon usage variation in the organism. Bivariate correlation analysis established the relationships between the first and second major axes with different codon usage indices and has led to a significant negative correlation between Nc and GC3s ($r = -0.558$, $P < 0.01$). We therefore, suggest that the genes located at the right side of the first axis with higher GC3s and GC content and lower ENC values reflect strong codon bias and this variations in codon usage is strongly correlated with nucleotide composition of the genes. The first major axis showed a significant positive correlation with C3s and G3s but is negatively correlated with A3s and T3s **Table 2 (see supplementary material)**. We showed that highly biased genes, those with G- and C- ending codons were clustered on the positive side whereas the codons ending with A and T predominate on the negative side of the first major axis (**Figure 1**). Significant negative correlation was observed between Nc and GC3s ($r = -0.558$, $P < 0.01$) in comparison to Nc and GC ($r = -0.394$). A significant positive correlation was also observed between Axis 1 and GC3s confirming that GC3s plays an important role in shaping codon usage bias in this genome. Moreover, when G3s and C3s are considered separately, the correlation coefficient exhibited by the positions of genes along with the first axis with C3s ($r = 0.542$, $P < 0.01$) is significantly larger than that with G3s ($r = 0.151$, $P < 0.01$), indicating that the contribution of C3s to the inter-genomic variation in overall GC3s content is greater than that of G3s. We therefore, speculate that the presence of compositional mutational bias shapes the codon usage bias in *T. elongatus* genome.

Translational optimal codons

To identify optimal codons which are preferred among the highly expressed sequences, we analyzed the codon usage in 10% of genes from both extremes of axis 1. Among highly biased genes, there were 20 codons with significantly higher usage (relative to synonyms) that could be inferred as translational optimal codons [7]. Among these 20 codons, there were 12 C-ending (60%), 6 G- (30% G) and 2 T-ending codons (10%) but no A-ending codon **Table 3 (see supplementary material)**. This is consistent with the correlation of Axis 1 with C3s ($r = 0.542$, $P < 0.01$). Most frequently used stop codon among highly expressed genes was UAG whereas among lowly expressed genes it was UGA.

The predominant nature of C-ending codons suggested that translational selection is also operating in codon usage variation in this organism. If the compositional constraints are the only

factor in shaping the codon usage variation among the genes, then A- and T-ending codons should have been present in good proportion but it is not there.

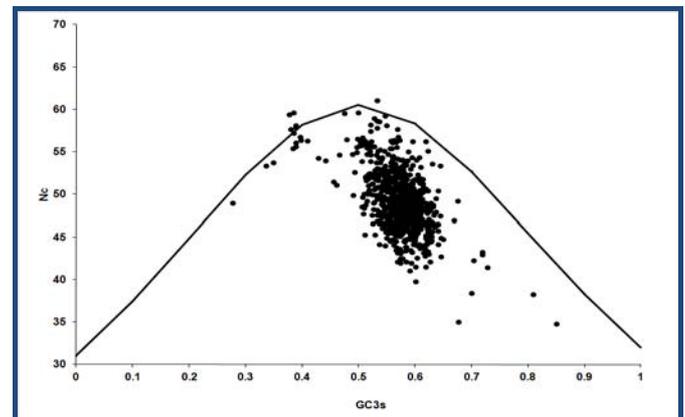


Figure 2: Nc plot for *T. elongatus* genes. The solid line indicates the expressed Nc value if codon bias is only due to GC3s.

Differential base usage in third codon position

The correlation of the frequencies of four bases in the third position against Nc values of all the genes has been checked **Table 4 (see supplementary material)**. We considered the highly biased genes having low Nc values as highly expressed and *vice versa*. In all the genes, the frequency of C at third codon position increased with decreasing Nc values, whereas frequencies of A and T increases with Nc. However, the frequency of G is not influenced in the third codon position.

Thus, the influence of mutational bias is reflected in the choice of bases in the third position. However, this is expected since the optimal codons are, in general, chosen in accordance with the mutational bias. It is expected that due to the translational selection, mutational bias appears to be more prominent in the third codon position of highly expressed genes [5].

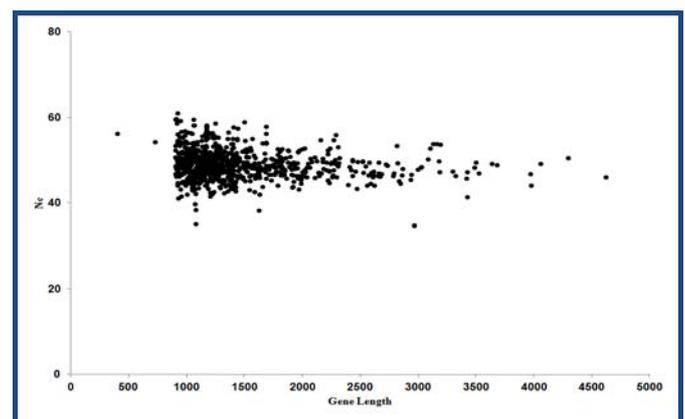


Figure 3: Plot of Nc vs Gene length for *T. elongatus* genes.

Relationship between codon bias and gene length

Selection for translational accuracy is predicted to have a positive correlation between codon bias and gene length [15]. From the plot drawn with gene length against Nc (**Figure 2**), it is clear that shorter genes have a much wider variance in Nc values and *vice versa* for longer genes. Lower Nc values of longer genes may be expected due to the direct effect of translation time on fitness or to the extra energy cost associated

with longer translational timing. Correlation analyses between gene length and gene position on axis 1, GC3s and Nc values ($r = 0.158, 0.170$ and -0.173 , respectively, $P < 0.01$) were all significantly correlated. The observed significant correlation in *T. elongatus* genome revealed that gene length influences codon usage of this organism. Gene length was negatively correlated with Nc ($r = -0.173$) but it showed positive correlation with GC3s ($r = 0.170$). Therefore, codon bias is lower in longer genes than in shorter ones (Figure 3). It is meaningful that up to certain extent, translational accuracy also plays a role in dictating the codon usage variation in this genome.

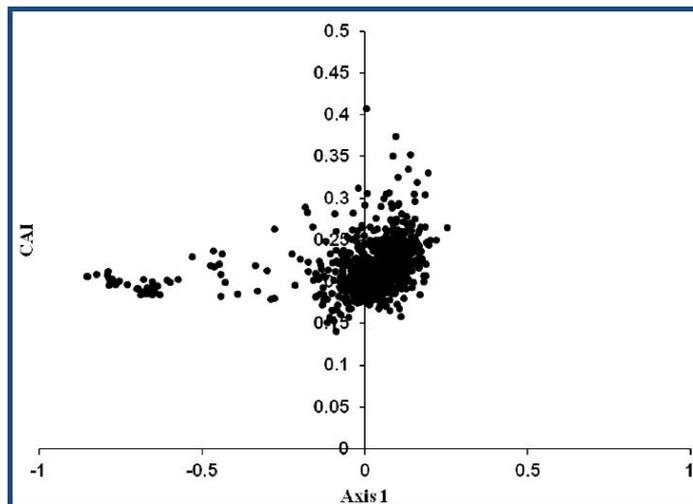


Figure 4: Scatter plot of CAI with Axis 1

CAI and codon usage variation: role of gene expressivity

Codon Adaptation Index (CAI) has been extensively used as a measure of gene expression level in organisms. Within *T. elongatus* genome, CAI showed positive correlation with the positions of genes along axis 1 and GC3s ($r = 0.274$ and $r = 0.229$ respectively, $P < 0.01$) (Figure 4) but was negatively correlated with Nc ($r = -0.389$). This has reflected the effect of gene expression level on codon selection pattern in *T. elongatus*. Data suggested that genes with higher expression level exhibiting a greater degree of codon usage bias were distributed at right side of the first axis and preferred to use G- or C-ending codons. This result indicated a role of expression levels of genes in codon usage variation along with compositional constraints and supported the argument of presence of additional factors along with compositional constraint in shaping the codon usage bias in this genome. Furthermore, CAI showed a positive correlation with C3s ($r = 0.605$, $P < 0.01$) whereas negative correlation with G3s ($r = -0.243$, $P < 0.01$) supporting our hypothesis that the influence of mutational bias is reflected in the choice of bases in the third codon position and thus, favor translational selection.

Conclusion:

Codon usage pattern was analyzed in *T. elongatus* genome to elucidate the factors responsible for variation in codon usage. From the analysis, it is clear that compositional constraint is not the only factor responsible for shaping the codon usage variation in this genome. In *T. elongatus*, there is a single major explanatory axis which accounted for codon usage variation. CA of *T. elongatus* genes revealed strong correlation of GC3s with this axis. In addition, a negative correlation was observed between Nc and GC3s. These results suggested that mutational

bias is the major factor in shaping codon usage in this organism. Among all the four nucleotides at third position, C3s is found to directly influence codon usage variation. Thus, the influence of mutational bias is reflected in the choice of bases in the third position and supported our hypothesis that compositional mutation bias possibly plays an important role in shaping the codon usage bias of this genome. The 3rd base of a codon is said to wobble and very often, changes in the 3rd base of a codon might not change the amino acid encoding reflecting that base compositional mutational bias could lead to different codon choice within the same protein sequence [23]. Highly expressed genes tend to use C- or G- at the synonymous positions compared with lowly expressed genes. Preference of C-ending codons in the highly expressed genes might be related to the translational efficiency of the genes as it has been reported that RNY (R- purine, N- any nucleotide base, and Y-pyrimidine) codons are more advantageous for translation [28]. It is interesting to note that 68% of the preferred triplets in the highly expressed genes of *T. elongatus* are pyrimidine ending codons suggesting that besides, mutational bias, translational selections are also operating extent in determining the codon usage variation in this organism. The interpretation of pyrimidine selection involves the energy level of the codon-anticodon interaction and is connected to translational fidelity [12]. Gene length was also found to influence codon usage and this fact supported the role of translational accuracy in dictating the codon usage variation of this genome. Gene expression level also influences codon selection pattern of this organism. Thus, it is imperative to note that mutational bias is the major factor in shaping codon usage in *T. elongatus* genome but the role of other factors such as translational selection and accuracy and gene expression level may not be excluded in codon usage in this organism.

Acknowledgement:

Financial support from Indian Council of Agricultural Research, New Delhi, India under National Agricultural Bioinformatics Grid project is gratefully acknowledged.

References:

- [1] Grantham R *et al.* *Nucleic Acids Res.* 1980 **8**: 49 [PMID: 6986610]
- [2] Sharp PM *et al.* *Nucleic Acids Res.* 1988 **16**: 17 [PMID: 3138659]
- [3] Zhou XP *et al.* *Biosystems.* 2005 **81**: 77 [PMID: 15917130]
- [4] Xu X *et al.* *J Zhejiang Univ Sci B.* 2008 **9**: 667 [PMID: 18763298]
- [5] Hassan S *et al.* *Advances in Bioinformatics.* 2010 **2009**: 316936. [PMID: 20150956]
- [6] Grantham R *et al.* *Nucleic Acids Res.* 1980 **8**: 1893 [PMID: 6159596]
- [7] Stenico M *et al.* *Nucleic Acids Res.* 1994 **22**: 2437 [PMID: 8041603]
- [8] Gupta SK *et al.* *J Biomol Struct Dyn.* 2004 **21**: 4 [PMID: 14692797]
- [9] Hsiao Y *et al.* *Comp Funct Genomics.* 2010 **2010**: 138538 [PMID: 21052492]
- [10] Chen SL *et al.* *Proc Natl Acad Sci.* 2004 **101**: 3480 [PMID: 14990797]
- [11] Grantham R *et al.* *Nucleic Acids Res.* 1981 **9**: 1 [PMID: 7208352]

- [12] Gouy M & Gautier C, *Nucleic Acids Res.* 1982 **10**: 22 [PMID: 6760125]
- [13] Roymondal U *et al.* *DNA Res.* 2009 **16**: 13 [PMID: 19131380]
- [14] Lobry JR & Gautier C, *Nucleic Acids Res.* 1994 **22**: 3174 [PMID: 8065933]
- [15] Eyre-Walker A, *Mol Biol Evol.* 1996 **13**: 864 [PMID: 8754221]
- [16] Bulmer M, *Genetics.* 1991 **149**: 897 [PMID: 1752426]
- [17] Ikemura T, *Mol Biol Evol.* 1985 **2**: 13 [PMID: 3916708]
- [18] Singh DP *et al.* *Am J Bioinform.* 2012 **1**: 10
- [19] Mitreva M *et al.* *Genome Biology.* 2006 **7**: 8
- [20] Lynn DJ *et al.* *Nucleic Acids Res.* 2002 **30**: 4272 [PMID: 12364606]
- [21] Supek F & Vlahovicek V, *BMC Bioinformatics.* 2005 **6**: 182 [PMID: 16029499]
- [22] Sharp PM *et al.* *Phil Trans R Soc.* 2010 **365**: 1203 [PMID: 20308095]
- [23] Bulmer M, *Nature.* 1987 **325**: 728 [PMID: 2434856]
- [24] Hershberg R & Petrov DA, *Annu Rev Genet.* 2008 **42**: 287 [PMID: 18983258]
- [25] Nakamura Y *et al.* *DNA Res.* 2002 **9**: 123 [PMID: 12240834]
- [26] Sharp PM & Li WH, *J Mol Biol.* 1986 **24**: 28 [PMID: 3104616]
- [27] Wright F, *Gene.* 1989 **87**: 23 [PMID: 2110097]
- [28] Sharp PM & Li W, *Nucleic Acids Res.* 1987 **15**: 3 [PMID: 3547335]
- [29] McInerney JO, *Bioinformatics.* 1998 **14**: 372 [PMID: 9632833]
- [30] <http://codonw.sourceforge.net/>

Edited by P Kanguane

Citation: Prabha *et al.* *Bioinformation* 8(13): 622-628 (2012)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

Supplementary material:

Table 1: Codon selection pattern in *Thermosynechococcus elongatus* BP-1 genes

AA	Codon	N	RSCU	AA	Codon	N	RSCU	
Phe	UUU	9544	1.42	Ser	UCU	2165	0.67	
	UUC	3898	0.58		UCC	4142	1.29	
Cys	UGU	2189	1.07		UCA	1670	0.52	
	UGC	1920	0.93		UCG	2215	0.69	
Tyr	UAU	5567	1.02		AGU	4846	1.5	
	UAC	5306	0.98		AGC	4300	1.33	
	UUA	3206	0.43		Pro	CCU	3745	0.69
Leu	CUU	5413	0.73			CCC	10636	1.97
	CUC	10228	1.38			CCA	3198	0.59
	CUA	4003	0.54			CCG	3979	0.74
His	CUG	12451	1.68	ACU		2793	0.52	
	CAU	3783	0.88	Thr		ACC	10173	1.89
Gln	CAC	4849	1.12		ACA	3410	0.64	
	CAA	11634	1.1		ACG	5103	0.95	
Ala	CAG	9594	0.9	Arg	CGU	3883	0.92	
	Asn	GCU	6793		0.71	CGC	11444	2.72
		GCC	17551		1.85	CGA	2223	0.53
		GCA	6386		0.67	CGG	6038	1.44
Lys	GCG	7318	0.77		AGA	701	0.17	
	AAU	5947	1.07		AGG	938	0.22	
Val	AAC	5173	0.93	Asp	GAU	12429	1.39	
	Glu	AAA	6752		1.06	GAC	5486	0.61
		AAG	6032		0.94	GGU	6874	0.97
Gly	GUU	5243	0.76	Gly	GGC	10644	1.5	
	GUC	7108	1.03		GGA	3343	0.47	
	GUA	2236	0.32		GGG	7580	1.07	
Ile	GUG	12995	1.88	Ile	AUU	14988	2.03	
	Glu	GAA	12375		1.08	AUC	6563	0.89
		GAG	10500		0.92	AUA	565	0.08

RSCU- relative synonymous codon usage calculated by the sum-up of all the genes together; N- number of codons; AA - amino acids

Table 2: Relationship between two major axes and codon usage indices identified through bivariate correlation analysis

	Nc	A3s	T3s	G3s	C3s	GC3s	GC	CAI
Axis 1	-.701**	-.739**	-.277**	.151**	.542**	.674**	.651**	.274**
Axis 2	-.077**	.021	-.279**	-.202**	.543**	.244**	-.114**	.582**

* Occurrences are significantly higher at P<0.01

Table 3: Relative synonymous codon usage for the highly and lowly expressed genes in *Thermosynechococcus elongatus* BP-1

AA	Codon	RSCU ^h	N ^h	RSCU ^l	N ^l	AA	Codon	RSCU ^h	N ^h	RSCU ^l	N ^l
Phe	UUU*	1.30	(508)	1.10	(305)	Glu	GAA	1.07	(849)	1.22	(458)
	UUC	0.70	(272)	0.90	(248)		GAG*	0.93	(738)	0.78	(291)
Leu	UUA	0.24	(95)	0.42	(103)	Ser	UCU	0.37	(70)	1.76	(325)
	UUG	1.05	(415)	2.18	(531)		UCC*	1.74	(328)	0.80	(147)
	CUU	0.60	(236)	0.99	(241)		UCA	0.33	(62)	0.65	(119)
	CUC*	1.71	(678)	0.61	(148)		UCG*	0.81	(152)	0.50	(93)
	CUA	0.46	(181)	0.52	(126)		AGU	1.22	(230)	1.33	(245)
	CUG*	1.94	(769)	1.27	(310)		AGC*	1.52	(286)	0.96	(177)
Ile	AUU*	2.03	(916)	1.36	(400)	Thr	ACU	0.27	(82)	0.93	(188)

	AUC	0.95	(429)	1.19	(349)		ACC*	2.38	(722)	1.10	(222)	
	AUA	0.01	(6)	0.45	(133)		ACA	0.44	(132)	1.17	(236)	
							ACG	0.91	(277)	0.81	(164)	
Val	GUU	0.52	(203)	1.43	(370)	Ala	GCU	0.56	(294)	1.09	(301)	
	GUC*	1.16	(452)	0.95	(245)		GCC*	2.19	(1149)	1.14	(314)	
	GUA	0.26	(102)	0.70	(180)		GCA	0.49	(255)	1.13	(313)	
	GUG*	2.06	(800)	0.92	(237)		GCG	0.77	(402)	0.64	(177)	
Tyr	UAU	0.87	(306)	1.14	(256)	Cys	UGU	0.85	(82)	1.27	(189)	
	UAC*	1.13	(400)	0.86	(194)		UGC*	1.15	(111)	0.73	(108)	
His	CAU	0.78	(189)	0.85	(160)		CGU*	0.82	(196)	0.54	(133)	
	CAC	1.22	(295)	1.15	(215)		CGC*	3.40	(812)	0.99	(245)	
Gln	CAA	1.08	(598)	1.07	(433)	Arg	CGA	0.24	(58)	1.11	(274)	
	CAG	0.92	(514)	0.93	(374)		CGG*	1.48	(354)	1.16	(286)	
Asp	GAU*	1.33	(753)	1.05	(414)		AGA	0.01	(2)	1.54	(380)	
	GAC	0.67	(382)	0.95	(372)		AGG	0.05	(11)	0.65	(160)	
Asn	AAU	0.95	(346)	0.87	(284)	Pro	CCU	0.48	(133)	1.05	(164)	
	AAC	1.05	(385)	1.13	(372)		CCC*	2.28	(625)	1.21	(189)	
Gly	GGU	0.92	(372)	1.21	(279)		CCA	0.35	(97)	1.22	(192)	
	GGC*	1.80	(729)	0.87	(200)		CCG*	0.88	(242)	0.52	(82)	
	GGA	0.30	(123)	1.09	(252)		Lys	AAA	0.93	(422)	1.11	(762)
	GGG	0.97	(393)	0.83	(191)			AAG*	1.07	(483)	0.89	(615)
Ter	UAA	0.92	12	0.08	1	Ter	UGA	0.38	5	2.00	26	
TR	UAG	1.69	22	0.92	12	Trp	UGG	1.00	185	1.00	364	

*codons putatively identified as optimal at $p < 0.01$; each group contains 10% of sequences at either extreme of the major axis generated by CA; AA-amino acid; N- number of codons; ^hhighly expressed genes; ^llowly expressed genes

Table 4: Correlation coefficient of Nc with T3s, C3s, G3s, A3s

Correlation	T3s	C3s	G3s	A3s
Nc	.210*	-.596*	-.002 ^{NS}	.567*

*significant relationships; NS: Non-significant; $P < 0.01$ for significant values