TRENDS IN CODON AND AMINO ACID USAGE IN HUMAN PATHOGEN *TROPHERYMA WHIPPLEI*, THE ONLY KNOWN ACTINOBACTERIA WITH REDUCED GENOME

SABYASACHI DAS, SANDIP PAUL and CHITRA DUTTA

Bioinformatics Centre, Indian Institute of Chemical Biology, Kolkata –700032, India

The factors governing codon and amino acid usages in the predicted protein-coding sequences of *Tropheryma whipplei* TW08/27 and Twist genomes have been analyzed. Multivariate analysis identifies the replicational-transcriptional selection coupled with DNA strand-specific asymmetric mutational bias as a major driving force behind the significant inter-strand variations in synonymous codon usage patterns in *T. whipplei* genes, while a residual intra-strand synonymous codon bias is imparted by a selection force operating at the level of translation. The strand-specific mutational pressure has little influence on the amino acid usage, for which the mean hydropathy level and aromaticity are the major sources of variation, both having nearly equal impact. In spite of the intracellular life-style, the amino acid usage in highly expressed gene products of *T. whipplei* follows the cost-minimization hypothesis. Both the genomes under study are characterized by the presence of two distinct groups of membrane-associated genes, products of which exhibit significant differences in primary and potential secondary structures as well as in the propensity of protein disorder.

1. Introduction

Whipple's disease is a rare multisystemic bacterial infection caused by an intracellular pathogenic actinobacteria *Tropheryma whipplei*.[1] The complete genome sequences of two different strains of this human pathogen, *T. whipplei* TW08/27 and *T. whipplei* Twist, reveal several atypical characteristics of the organism.[1,2] First, their genome sizes are small (<1 Mb) and G+C-content (46% approx.) are low, as compared to other actinomycete species, which have, in general, genome sizes ranging from 1 million bp to 8 million bp and high G+C contents.[3] Second, the genomes bear all the traits of strictly host-adapted organisms, including pronounced deficiencies in energy metabolism and a lack of key biosynthetic pathways.[1,2] Third, both the genomes exhibit a great deal of genetic variability, mostly directed towards the changes in cell-surface proteins, indicating that immune evasion and host interaction play an important role in their lifestyle. Such atypical characteristics of *T. whipplei* reflect the possible existence of special selection forces operative at the genome and/or proteome levels, unveiling of which calls for an in-depth analysis of the trends in codon and amino acid usage in the organism.

Synonymous codon usage in most of the unicellular organisms are primarily governed by directional mutational bias and translational selection,[4,5] though several other factors such as context-dependence,[6] replicational-transcriptional selection,[7-9] protein hydropathy,[10] ecological niches [11] etc. may also have significant influences. The

amino acid usage in microbial organisms may also be influenced by several factors like hydrophobicity, expressivity and aromaticity of the respective proteins,[12] cost minimization and conservation of GC-rich amino acids in highly expressed gene-products etc.[9,13,14] Multivariate analyses carried out in the present study indicate that the codon and amino acid usage in this human pathogen might be a consequence of a complex balance between replicational-transcriptional selection, translational control and other physico-chemical properties of the gene-products. The study, apart from providing an insight into the underlying selection pressures operative at the gene/protein level of *T. whipplei*, may also offer a better understanding of evolution of this host-adapted microorganism.

2. Methods and Materials

2.1 Sequence Retrieval

All protein-coding sequences of *T. whipplei* TW08/27 and Twist genomes were extracted from NCBI GenBank. In order to reduce sampling errors, the annotated genes with less than 100 codons were excluded from the analysis. The presumed duplicates, transposes genes and the genes with internal stop codons, untranslatable codons were also excluded. Finally 729 sequences for *T. whipplei* TW08/27 and 734 sequences for *T. whipplei* Twist were selected for analysis.

2.2 Sequence analysis for identifying trends in codon and amino acid usage

The genes present in leading and lagging strand were isolated on the basis of reported location of oriC in *T. whipplei* Twist by Raoult et al. (2003).[2] Based on the change in AT-skew signal using Oriloc and the conservation of dnaA-dnaN-recF gene cluster, the oriC region in *T. whipplei* TW08/27 is assumed to be located at 0 KB. The terminus is assumed at the second inflexion in AT-skew and thus allowed us to locate each coding sequence that present either in the leading or lagging strands of replication.

To identify the major factors shaping variation in relative synonymous codon usage (RSCU) and relative amino acid usage (RAAU) among *T. whipplei* genes, we applied correspondence analyses (COA) using CODONW 1.4.2. GC_{1+2} (G+C content at first and second codon positions), GC_{3S} and GT_{3S} (the frequency of G+C and G+T respectively at synonymous third codon positions) were calculated for each coding sequences in both strains of *T. whipplei*. Parameters like total number of occurrence of each codon, RSCU,[4] codon adaptive index (CAI),[4] RAAU, average hydrophobicity (Gravy score),[15] aromaticity [12] and average size/complexity quotient [16] of encoded proteins were also calculated to find out the factors influencing codon and amino acid usage. To examine the nucleotide substitution patterns, we estimated pairwise synonymous divergences (d_S) as well as non-synonymous divergences (d_N) between the Orthologous genes of *T. whipplei* TW08/27 and *T. whipplei* Twist using the MEGA program (version 2.1), as described by Nei and Gojobori (1986).[17] In order to detect the differences between the

two classes of genes, if any, codon and amino acid abundance were compared in 2 x 2 contingency tables. Linear regression analysis was used to find out the significance of association between the positions of sequences along major axes of COA and biological variables using STATISTICA (Version 6.0). The prediction of protein secondary structure was performed using GOR IV algorithm[18] from Expasy proteomics server and the disordered regions within proteins were predicted using GlobPlot (http://globplot.embl.de).[19]

3. Results and Discussion

3.1 Asymmetrical mutational bias, coupled with replicational-transcriptional selection on synonymous codon usage

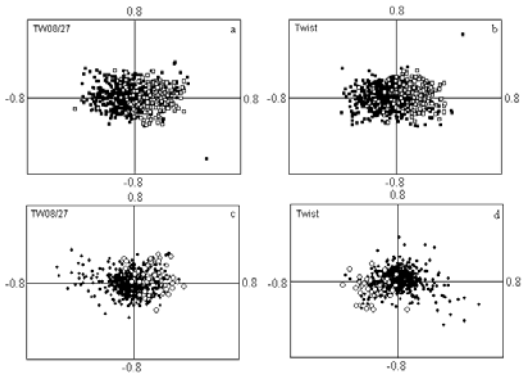

Figure 1. Position of genes along the first two principal axes generated by correspondence analysis (COA) on relative synonymous codon usage (RSCU) values from (a) *T. whipplei* TW08/27 (b) *T. whipplei* Twist genome and separately from leading strand of (c) TW08/27 and (d) Twist genome. The filled quadrangle and open quadrangle for figure a and b represent the genes transcribed in the leading and the lagging strands of replication respectively whereas for the figure c and d, the filled circles represent the genes transcribed only in the leading and the open circles represent the highly expressed genes of that strand.

Figure 1a and 1b show the position of the genes on the plane defined by the first and second major axes generated by COA on RSCU values of coding sequences in TW08/27 and *Twist* genomes respectively. The first principal axis accounts for 9.23% and 9.19%, while the second axis accounts for 5.15 % and 5.07 % of total variability for TW08/27 and *Twist* genomes respectively. In each plot, the genes transcribing in the leading and the lagging strands of replication are segregated in two discrete clusters with little overlap along the axis 1. Similar scatter plots with two distinct clusters of points were observed earlier in case of genomes with pronounced strand-specific mutational bias.[7-9] The chi square test on occurrences of different codons in two replicating strands showed that there are 21 G-ending/U-ending codons, usages of which significantly increase ($p<0.01$) in the leading strand genes, whereas usage of 24 and 25 C-ending/A-ending codons are significantly higher in the lagging strand of the TW08/27 and Twist genome respectively (data not shown). Thus, in both species under study, the primary cause of variation in synonymous codon usage is whether a gene is transcribed in the same direction as replication, or opposite to it. To our knowledge, this is for the first time the strong strand-specific biases is being reported to be the prime selection force behind the

synonymous codon usage in any Actinobacteria. Furthermore, in both strains of *Tropheryma*, the number of predicted coding sequences is significantly higher in the leading strand (73.8 % for TW08/27 and 74.1 % for Twist genome) than in the lagging strand and the distribution of highly expressed genes (*i.e.* ribosomal proteins, transcription and translation processing factors etc.) are also significantly skewed, most of the potential highly expressed genes (>70%) being transcribed from the leading strand. All these observations indicated that the replicational-transcriptional selection coupled with asymmetrical mutational bias is the primary cause of intra-genomic variations in codon usage pattern in *T. whipplei*.

3.2 Evidences for translational selection in T. whipplei

In order to examine the possible effect of translational selection, if any, on codon selection by the highly expressed genes of *T. whipplei*, we have performed a COA on RSCU values of the genes transcribed from the leading strand of replication only (as most of the highly expressed genes (>70%) are located in leading strand). Most of the potential highly expressed genes (*i.e.* genes encoding ribosomal proteins, transcription translation processing factors, heat-shock proteins etc.) are clustered at one extreme of the axis 1, which represents about 8% of total variance for both strains of *T. whipplei* (Fig. 1c,d). The first axis of COA on RSCU values of leading strand genes exhibits strong significant correlation ($r = 0.79$ and -0.77 at $p<0.0001$ for TW08/27 and Twist genome respectively) with the CAI values of genes. Therefore, the position of genes along the first axis of COA might be related to expressivity. In order to ascertain the influence of translation selection on highly expressed genes, we compared the estimated d_S values between 50 putative highly and 50 putative lowly expressed orthologous genes between TW08/27 and Twist genomes taken from the pooled data of the two extremes of axis1. The mean d_S value for high-expression genes (*i.e.* ribosomal proteins, transcription and translation processing factors, heat-shock proteins etc.) is 0.0049, which is significantly lower (t-test, $p<10^{-4}$) than that of low-expression genes (mean $d_S = 0.0058$) indicating that highly expressed genes have lower divergence at synonymous sites than lowly expressed genes. Therefore, it may be concluded that apart from the strand-specific mutational bias, a selection, putatively operating at the level of translation, also might have influenced the synonymous codon choice in genes of these two strains of *T. whipplei*.

3.3 Major sources of variation in proteome composition

To investigate whether the DNA strand specific mutational biases have any impact on amino acid usage in *T. whipplei* gene-products, we have performed COA on relative amino acid usage (RAAU) of the encoded proteins (Fig.2). There is no clear segregation of the proteins encoded by the leading and lagging strands genes. When the cumulative amino acid usage of encoded gene products in two strands are compared separately, only

two amino acids encoded by G+U-rich codons (Phe and Val) are significantly over represented ($p<0.001$) in the leading strand and three residues encoded by A+C-rich codons (Lys, Asn and Thr) are more abundant ($p<0.001$) in the lagging strand.

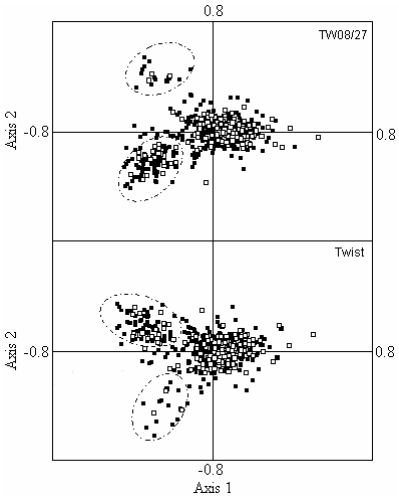

Figure 2. Position of each gene along the two major axes of variation generated by correspondence analysis on RAAU of encoded gene products for (a) *T. whipplei* TW08/27and (b) *T. whipplei* Twist. The filled quadrangle and open quadrangle represent the genes transcribed in the leading and lagging strands of replication respectively. Large and small dashed-line ovals represent the large and small clusters of genes encoding membrane-associated proteins.

In both strains of *T. whipplei*, the first three axes generated by COA on amino acid usage explain about 42% of total variability. The first and second major axes are strongly correlated both with hydrophobicity as well as with the aromaticity of encoded proteins (Table 1), implying hydropathy and aromaticity to be the major factors for amino acid variation in *T. whipplei* proteins. It is worth noting that there are two distinct clusters of proteins near the left end of the axis 1 (Fig. 2). A careful investigation reveals that the small cluster contained the genes for membrane associated proteins including WiSP family members and few hypothetical proteins, whereas genes coding for integral membrane proteins, several transporters, subunits of cytochrome C etc. are present in the large cluster. Although the genes present in both the clusters are mainly membrane-associated proteins, the amino acid usage profile and the propensities for formation of secondary structures are distinct from one another (Table 2). Proteins of these two clusters also differ in content of potential disordered structures. Disordered regions in proteins can be predicted by the lack of regular secondary structures whereas ordered regions (often termed globular) typically contain regular secondary structures packed into a compact globule.[19,20] As the probable coil forming regions are significantly higher in small cluster proteins (Table 2), disordered structures are more commonly found in the proteins of small cluster as compared to those comprising of the large cluster (Fig. 3). Recent investigations have indicated that disordered structures are usually more favored by proteins involved in regulatory functions and binding of various ligands.[20,21] Therefore, it may be presumed that the proteins in small cluster, which might play important roles in interactions with the host and/or immune evasion,[1,2] would be over represented by disordered structures. Members of the other cluster containing less disordered regions and exhibiting higher propensities for alpha-helical regions are

primarily involved in transport and other membrane-associated processes. Compositions of the membrane-associated proteins of these two clusters are, therefore, influenced by evolutionary selective pressures resulting in a fine coordination between function, structure and stability.

Table 1. Non-parametric tests of association between the first three axes of COA on RAAU and multiple parameters of encoded proteins

	T. whipplei TW08/27			*T. whipplei* Twist		
	Variability explained(%)	Source of variation	Correlation coefficient*(r)	Variability explained(%)	Source of variation	Correlation coefficient*(r
1st Axis	18.9	Gravy Score Aromaticity	-0.83 -0.68	18.4	Gravy Score Aromaticity	-0.87 -0.68
2nd Axis	14.8	Gravy Score Aromaticity	-0.59 -0.45	14.1	Gravy Score Aromaticity	0.46 0.38
3rd Axis	8.8	Size/complexity GC_{12}	-0.79 0.65	8.8	Size/complexity GC_{12}	-0.74 0.66

* All correlations are significant at $p < 0.0001$.

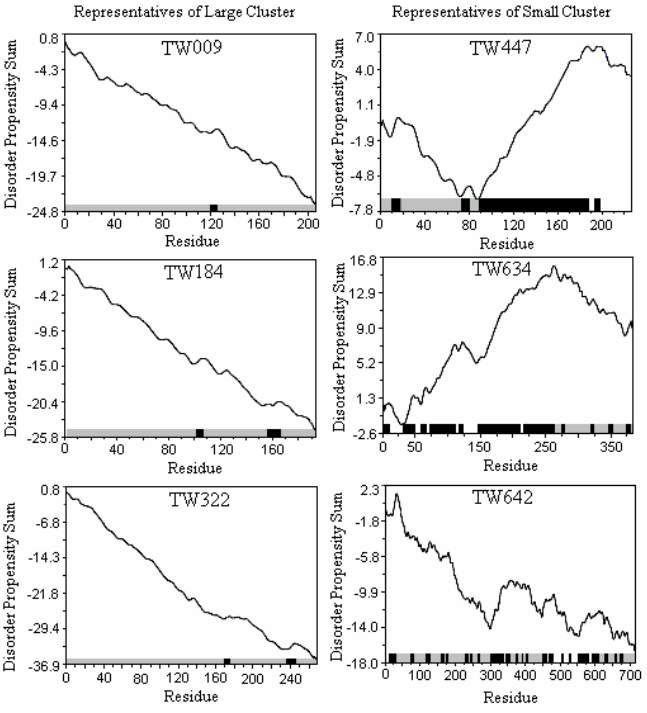

Figure 3. GlobPlot of the membrane associated proteins encoded by genes taken from large and small clusters generated by COA on RAAU for *T. whipplei* TW08/27 genome. The plots of the left and right panels are the representatives of large and small clusters respectively. Black colour indicates the disorder regions (lack of regular secondary structures) and gray colour indicates the ordered (globular) regions that typically contain regular secondary structures.

Another factor that largely influences the variations in amino acid usage is gene expressivity, as indicated by the presence of most of the potential highly expressed genes

near the negative extreme end of the axis 3. A strong negative correlation of this axis with the average size/complexity quotient of the encoded proteins (Table 1) suggests that the highly expressed genes in *T. whipplei* have a tendency of avoiding the heavier residues including the aromatic ones in spite of their obligatory intracellular lifestyle. This supports the cost minimization hypothesis.[13] This is probably a genome-level adaptation to the host environment, as utilization of the less expensive and smaller residues by the highly expressed genes can minimize the energy exhaustion of the host and help them thereby to maintain the sustained infection, having least chance of elimination by the host.

Table 2. Amino acid usage and mean value of potential secondary structures of membrane associated proteins in the large and small clusters for *T. whipplei* TW08/27 and *T. whipplei* Twist genome

Amino acids and Predicted Secondary Structures	***T. whipplei*** TW08/27		*T. whipplei* Twist	
	Large cluster (%)	Small cluster (%)	Large cluster (%)	Small cluster (%)
Phe	8.17*	3.17	7.87*	3.32
Leu	14.36*	8.24	14.46*	8.07
Ser	7.54	10.62+	7.66	10.27+
Tyr	2.97	4.70+	2.96	4.70+
Cys	1.07*	0.59	1.16*	0.59
Trp	1.81*	0.68	1.77*	0.79
Pro	4.27	8.07+	4.36	7.92+
His	1.33	1.73	1.39	1.47
Gln	1.83	3.20+	1.96	3.20+
Arg	4.56*	3.14	4.82*	3.18
Ile	8.76*	4.76	8.46*	4.77
Met	2.25*	0.50	2.18*	0.56
Thr	4.67	13.78+	4.67	13.93+
Asn	2.54	3.57+	2.57	3.44+
Lys	2.80	4.08+	2.83	3.97+
Val	9.29	8.74	9.25	8.99
Ala	9.43*	6.41	9.27*	6.43
Asp	2.33	3.67+	2.43	3.63+
Glu	1.97	2.10	2.08	2.16
Gly	7.71	8.09	7.49	8.46
Alpha helix	32.17*	9.44	31.92*	9.93
Beta sheet	24.29	29.95+	24.96	29.62+
Random coil	43.54	60.61+	43.12	60.45+

Values marked with *or + are significantly ($p < 0.001$) more frequent in large or small cluster gene products respectively.

3.4 Relation between gene expression, protein conservation and amino acid usage

To understand the possible effect of gene expression on amino acid usage, we compared the average RAAU profiles of putative highly and lowly expressed gene-products of *T. whipplei* (Fig.4). It is interesting to note that except Pro, all other residues encoded by GC-rich codons (Gly, Ala and Arg) are significantly over represented ($p<0.001$) in the

highly expressed gene-products, while usages of the residues encoded by all AU-rich codons (Phe, Tyr, Met, Ile and Asn) except Lys, are significantly higher ($p<0.001$) in lowly expressed gene-products. Increased usage of GC-rich or decreased usage of AU-rich codons in highly expressed genes of *T. whipplei* is supported by the significant positive correlation of GC_{1+2} with the coordinates of axis 3 of COA on RAAU (Table 1).

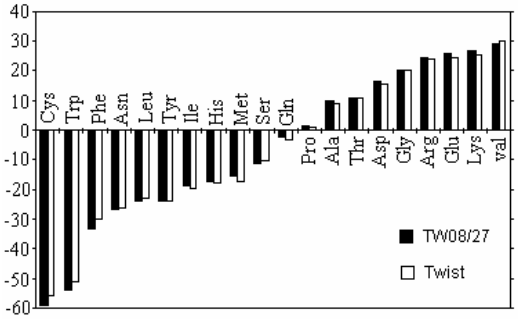

Figure 4. Difference in relative amino acid usage of highly and lowly expressed genes of *T. whipplei* TW08/27 (filled bars) and *T. whipplei* Twist (open bars). The differences were derived as R_{HL} = [{(Freq.H/Freq.L)-1} x 100].

Both the genomes under study have relatively low G+C-content (46% approx.), as compared to other actinomycete species and hence, a question arises: why do the highly expressed genes of these species exhibit higher usage of the residues encoded by G+C-rich codons? This might have happened for either of the following two reasons: (i) the highly expressed genes of *T. whipplei* are more conserved at the amino acid level than their lowly expressed counterparts and hence, they have retained a GC-richer composition which is closer to their putative ancestral state, or (ii) all genes are undergoing substitutions at a comparable rate irrespective of their level of expression, but due to some functional advantages, the highly expressed genes exhibit a positive selection in favour of the GC-rich residues. In order to examine which of these two possibilities is more likely to be true, we have compared the estimated pairwise non-synonymous divergences (d_N) between all orthologous of putative highly and lowly expressed genes. The mean $d_N = 0.021$ for highly expressed genes are significantly lower (t-test, $p<10^{-4}$) than that of lowly expressed genes ($d_N = 0.057$). It may be inferred therefore that in *T. whipplei*, the overall protein conservation in highly expressed genes is much higher and it also influences the discrimination in amino acid usage between highly and lowly expressed genes. The enhanced usage of the residues encoded by GC-rich codons in the highly expressed gene-products might have been possible due to more conservation at the amino acid level in highly expressed genes than the low-expression genes and hence, they retain GC-rich compositions which are closer to their putative ancestral state.

In summary, the present study reveals that the patterns of synonymous codon and amino acid usage in *T. whipplei* are the result of several factors. The replicational-transcriptional selection coupled with asymmetric mutational bias is the primary cause of intra-genomic variations in codon usage pattern. There is a residual intra-strand bias in

synonymous codon usage by the highly expressed genes possibly due to the presence of translational selection. However, the influence of strand-specific mutational pressure is not so pronounced at the level of amino acid usage. The hydrophobicity and aromaticity seem to be the major sources of variation, both having nearly equal influence in amino acid usage. In spite of the intracellular life-style, the amino acid preferences in highly expressed gene products of *T. whipplei* follow the cost-minimization hypothesis. Another interesting finding is that the products of the highly expressed genes prefer to use the residues encoded by GC-rich codons, although the *T. whipplei* genomes, on an average, has only 46% G+C-content. The analysis presented here indicates that this might be due to greater conservation of a relatively GC-rich ancestral state in the highly expressed genes. Even the energetic cost of amino acid residues [22] might play a significant role in retaining the GC-rich residues in highly expressed genes. The study also shed lights on the diverse compositional and structural characteristics of two groups of membrane-associated proteins that might play distinct roles in host interactions.

References

1. S. D. Bentley, M. Maiwald, L.D. Murphy, et al. Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet*, 361:637-644, 2003.
2. D. Raoult, H. Ogata, S. Audic, C. Robert, K. Suhre, M. Drancourt and J. M. Claverie. *Tropheryma whipplei* Twist: a human pathogenic Actinobacteria with a reduced genome. *Genome Res*, 13:1800-1809, 2003.
3. S. Casjens. The diverse and dynamic structure of bacterial genomes. *Annu Rev Genet*, 32:339-377, 1998.
4. P. M. Sharp and W. H. Li. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15:1281-1295, 1987.
5. A. Pan, C. Dutta, and J. DAS. Codon usage in highly expressed genes of *Haemophilus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. *Gene*, 215:405-413, 1998.
6. O. G. Berg and P. J. Silva. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. *Nucleic Acids Res*, 25:1397-1404, 1997.
7. J.O. McInerney. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA*, 95:10698-10703, 1998.
8. H. Romero, A. Zavala, H. Musto. 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, 2000.
9. S. Das, S. Paul, S. Chatterjee and C. Dutta. Codon and Amino Acid Usage in Two Major Human Pathogens of Genus *Bartonella* – Optimization Between Replication-Transcriptional Selection, Translational Control and Cost Minimization. *DNA Res.*, 12:91–102, 2005a.
10. A. B. de Miranda, F. Alvarez-Valin, K. Jabbari, W. M. Degrave and G. Bernardi. Gene expression, amino acid conservation, and hydrophobicity are the main factors

shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *J. Mol. Evol.*, 50:45-55, 2000.

11. G. A. Singer and D. A. Hickey. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*, 317:39-47, 2003.
12. J. R. Lobry and C. Gautier. 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, 1994.
13. H. Seligmann. Cost-minimization of amino acid usage. *J Mol Evol*, 56:151-161, 2003.
14. S. Das, A. Pan, S. Paul, and C. Dutta. Comparative Analyses of Codon and Amino Acid Usage in Symbiotic Island and Core Genome in Nitrogen-Fixing Symbiotic Bacterium Bradyrhizobium japonicum. *J. Biomol. Struct. Dyn.*, In Press., 2005b.
15. J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 157:105-132, 1982.
16. M. J. Dufton. Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins? *J Theor Biol*, 187:165-173, 1997.
17. M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3:418-426, 1986.
18. J. Garnier, J. F. Gibrat and B. Robson. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.*, 266:540-553, 1996.
19. R. Linding, R. B. Russell, V. Neduva and T. J. Gibson. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, 31:3701-3708, 2003.
20. M. Fuxreiter, I. Simon, P. Friedrich and P. Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol*, 338:1015-1026, 2004.
21. A. L. Fink. Natively unfolded proteins. Curr Opin Struct Biol, 15; 35-41, 2005.
22. H. Akashi and T. Gojobori. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci USA*, 99:3695-3700, 2002.