Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 22, Issue Number 2, (2004)

Investigation on the Causes of Codon and Amino Acid Usages Variation Between Thermophilic *Aquifex aeolicus* and Mesophilic *Bacillus subtilis*

http://www.jbsdonline.com

S. Basak
T. Banerjee
S.K. Gupta
T. C. Ghosh*

Bioinformatics Centre
Bose Institute, P 1/12
C.I.T. Scheme VII M
Kolkata 700 054, India

Abstract

Base composition, codon usages and amino acid usages have been analyzed by taking 529 orthologous sequences of *Aquifex aeolicus* and *Bacillus subtilis,* having different optimal growth temperatures. These two bacteria do not have significant difference in overall GC composition, but $GC_{(1+2)}$ and GC_3 levels were found to vary significantly. Significant increments in purine content and GC_3 composition have been observed in the coding sequences of *Aquifex aeolicus* than its *Bacillus subtilis* counterparts. Correspondence analyses on codon and amino acid usages reveal that variation in base composition actually influences their codon and amino acid usages. Two selection pressures acting on the nucleotide level (GC_3 and purine enrichment), causes variation in the amino acid usage differently in different protein secondary structures. Our results suggest that adaptation of amino acid usages in coil structure of *Aquifex aeolicus* proteins is under the control of both purine increment and GC_3 composition, whereas the adaptation of the amino acids in the helical region of thermophilic bacteria is strongly influenced by the purine content. Evolutionary perspectives concerning the temperature adaptation of DNA and protein molecules of these two bacteria have been discussed on the basis of these results.

Introduction

Microorganisms can be broadly divided into three categories namely, psychrophiles, mesophiles and thermophiles according to their optimal growth temperatures. Microorganisms that grow below 20 ºC are known as psychrophiles and that grow above 55 ºC are called thermophiles and the rest are known as mesophiles. All macromolecules of such bacteria must be stable and functional at their respective growth temperature. Much research has been carried out to understand the basic underlying principles in this regard. Bernardi proposed thermal adaptation hypothesis, which gives a link between GC content and temperature if there exists considerable variation of guanine (G) and cytosine (C) content between species (1, 2). G:C pairs are thermally more stable than A:T pairs, G:C pairs being connected by three hydrogen bonds and A:T pairs by two (3). However, the thermal adaptation hypothesis has failed to make any correlation between genomic GC and growth temperature in many cases (4-6). In recent years it has been reported (7) that dinucleotide composition of DNA are positively correlated with OGT of organisms, irrespective of their physiological relatedness. On the other hand, it was also reported that genomic G+C contents have a dramatic effect on the average amino-acid composition of the encoded proteins (8) and subsequently the role of natural selection in changing the amino acids according to their DNA base composition has been discarded (9). On the other hand D'Onofrio *et al*. (10) by analyzing a large number of bacterial genomes demonstrated that changes in GC_3 composition is significantly related to the structural and functional changes of the encoded protein, which completely goes against neutralist view as proposed by Gu *et al*. (11).

Phone: +91-33-2334 6626
Fax: +91-33-2334 3886
Email: tapash@bic.boseinst.ernet.in

Asymmetrical substitution pattern in orthologous proteins has been reported from organisms having different optimal growth temperature but sharing similar GC composition and these results gave an indication of which amino acid substitutions were adaptive at different temperature (12, 13). Even when mesophiles and thermophiles have the same genomic G+C content, it would not be possible to interpret all asymmetrical substitution patterns between them as evidence for thermal adaptation, because there are other processes besides changes in G+C content (14, 15). A lot of research works have been carried out in selecting the amino acid usages according to the growth temperature of an organism, but unfortunately very few authors have studied the relationships between the synonymous codon usage and their growth temperature. Very recently, it has been shown that there is a consistent difference in the pattern of synonymous codon usage between thermophilic and mesophilic prokaryotes (16, 17) and there is strong evidence that this difference is the result of selection linked to thermophily (18). The selection for codon and amino acid usages at high temperature and its relation to the base composition has been discussed elsewhere (18).

In the present study, detailed comparative analyses on base composition, codon and amino acid usages have been performed on 529 orthologous sequences of two eubacteria namely, *Aquifex aeolicus* and *Bacillus subtilis* both having similar genomic G+C contents but different optimal growth temperatures. At this point it is to be mentioned that the inter-species variability of (G+C) content in weakly selected positions is close to the intra-species variability in these two eubacteria (19) and since the polymorphisms in (G+C) content is about 5%, it is quite sensible for evolutionary studies with these two organisms. The goal of our study is to understand the effect of base composition in selecting the codon and amino acid usages in these two organisms. Our results suggest that base composition has a strong effect in selecting the codon and amino acid usage in these two bacteria. Our results also suggest that purine enrichment in the coil and helix region not only stabilizes the nucleic acid but also influences the amino acid composition according to the growth temperature. Additionally, GC_3 composition influences the amino acid usage significantly in coil structure.

Materials and Methods

Five hundred and twenty nine orthologous gene pairs of *Aquifex aeolicus* and *Bacillus subtilis* were extracted from Cluster of Orthologous Genes (COG) database, available at NCBI (http://www.ncbi.nlm.nih.gov/COG). CLUSTALW was used to align the protein sequences. Codon based alignment program developed by us was used to align the DNA sequences. The secondary structures corresponding to each of the sequences were downloaded from Consensus Secondary Structure Prediction Program available at http://npsa-pbil.ibcp.fr (20). The consensus in the prediction of secondary structure is defined as at least five algorithms out of eight used yielding the same predicted structure. Correspondence analysis available in CodonW program was used to investigate the major trend in codon and amino acid usage variation among the genes (21). The Students *t test* was used to evaluate the significance of the pair-wise differences in nucleotide as well as amino acid composition.

Results & Discussion

Overall Compositional Analysis

The average levels of $GC_{(1+2)}$ and GC_3 of 529 orthologous genes of both *Aquifex aeolicus* and *Bacillus subtilis* shown in Table I indicates that there are significant differences in $GC_{(1+2)}$ and GC_3 levels between these two organisms in spite of the fact that these two organisms do not have significant difference in overall GC composition. Indeed, a GC_3 level of *Aquifex aeolicus* is significantly higher than the GC_3 levels of *Bacillus subtilis* ($P < 2.72 \times 10^{-52}$), whereas opposite is true for

$GC_{(1+2)}$ levels ($P < 4.18 \times 10^{-50}$). Differences in $GC_{(1+2)}$ levels between the two organisms indicate that the occurrences of amino acids must be different between the orthologous sequences of these two bacteria. It is well known that majority of the substitutions at the first two codon positions results in non-synonymous substitution and it was also argued that first and second codon positions undergo strong purifying selection, whereas majority of nucleotide substitutions at the third codon positions are synonymous and is under weak purifying selection (22-24). Therefore, one can reasonably expect that the large differences in GC_3 levels between these two bacteria might have very little role in differentiating the amino acids in these two organisms. In order to investigate the role of GC_3 in influencing the evolution of encoded proteins we further classified the genes as invariant (where the identical amino acid appears in the corresponding aligned sites) and variant (where non-identical amino acid appears in the corresponding aligned sites) portions. To do this we aligned the orthologous sequence pairs and then corresponding nucleotide sequences of invariant as well as variant portions were retrieved. The GC_3 levels at the three different codon positions both at variant and invariant positions (data not shown) indicates that GC levels at both first and third

Table I
Base compositions of *Aquifex aeolicus* and *Bacillus subtilis* genes.

		A.aeolicus	*B.subtilis*	p-value
OVERALL				
	GC	43.58	44.29	n.s.
	GC_{1+2}	42.34	45.59	4.18×10^{-50}
	GC_3	48.71	42.56	2.72×10^{-52}
	A_1	33.02	29.77	3.72×10^{-36}
	A_2	34.13	32.95	1.03×10^{-03}
	A_3	30.36	28.23	8.47×10^{-13}
	G_1	35.64	35.97	n.s.
	G_2	15.50	15.50	n.s.
	G_3	24.40	21.75	6.05×10^{-18}
HELIX				
	GC_{1+2}	38.18	43.20	1.03×10^{-53}
	GC_3	48.94	42.09	5.63×10^{-40}
	A_1	34.25	28.75	4.74×10^{-42}
	A_2	38.83	35.98	6.23×10^{-08}
	A_3	32.76	31.02	1.00×10^{-04}
	G_1	34.78	36.14	2.33×10^{-03}
	G_2	10.83	10.49	n.s.
	G_3	28.02	25.25	5.34×10^{-10}
COIL				
	GC_{1+2}	50.99	52.46	2.52×10^{-07}
	GC_3	49.72	43.14	2.17×10^{-40}
	A_1	31.66	29.46	3.26×10^{-09}
	A_2	35.55	35.65	n.s.
	A_3	28.58	28.27	n.s.
	G_1	37.48	37.35	n.s.
	G_2	23.50	23.26	n.s.
	G_3	23.02	19.86	1.73×10^{-18}
STRAND				
	GC_{1+2}	33.17	33.82	n.s.
	GC_3	45.55	43.17	9.62×10^{-04}
	A_1	32.65	33.31	n.s.
	A_2	18.13	16.98	2.49×10^{-02}
	A_3	28.59	21.06	6.96×10^{-29}
	G_1	34.43	32.83	1.00×10^{-02}
	G_2	8.61	7.97	n.s.
	G_3	18.70	18.98	n.s.

codon positions at invariant sites are significantly different suggesting that the differences in GC levels at the first and third codon positions between the two organisms are nothing but the manifestations of synonymous nucleotide substitution. In the variant portion it was observed that GC composition of *Aquifex aeolicus* genes at the first and second codon positions were found to be significantly lower than the *Bacillus subtilis* counterparts and GC_3 levels of *Aquifex aeolicus* genes were found to be significantly higher than *Bacillus subtilis*. The difference in GC_2 levels in the variant region is expected as it is mainly responsible for the non-synonymous substitutions, but GC_1 and GC_3 differences are not always responsible for a synonymous substitution. In fact, data of amino acid substitution matrix based on pair-wise alignment of orthologous sequences indicates that large proportions of amino acids are substituted by those amino acids, which are interchangeable only by changing bases at the third positions of codons (data not shown).

Overall Codon Usage Analysis

Multivariate statistical analysis has been widely used to study the codon usage variation among the genes in different organisms. Correspondence analysis is one of the multivariate statistical technique in which the data are plotted in a multidimensional space of 59 axes (excluding Met, Trp and stop codons) and then it determines the most prominent axes contributing the codon usage variation among the genes. The CA was performed using the RSCU data in order to minimize the effects of amino acid composition. CA analysis on RSCU values detected one major trend and accounted for 26.53% of the total variation and none of the other axes accounted more than 9.26% of the total variation. Thus it can be said that there are single major explanatory axis on the synonymous codon usage variation among the genes in this dataset. The positions of the genes along the first and second major axes produced by CA on RSCU values are shown in Figure 1 and it was observed that the genes can be separated along the first major axis according to the genome type. At this point it is worthwhile to mention that since the two bacteria have different growth temperatures it can reasonably be argued that growth temperatures have a profound effect in separating the genes along the first major axis as observed by Lynn *et al*. (17). It was also observed that the positions of the genes along the first major axis is positively correlated with GC_3 ($r = 0.545$, $p < 0.01$), indicating that GC_3 levels are relatively higher in *Aquifex aeolicus* genes than *Bacillus subtilis* genes. To see how the different codons are contributing towards codon usage variation among the genes in the first major axis we have plotted the distribution of codons on the first two major axes, which is shown in Figure 2. From the Figure 2 it is evident that almost all the codons are equally contributing towards the codon usage variation among the genes along the first major axis. The analysis of relative synonymous codon usage may not detect any constraint imposed by amino acid composition. To examine if amino acid compositions exert

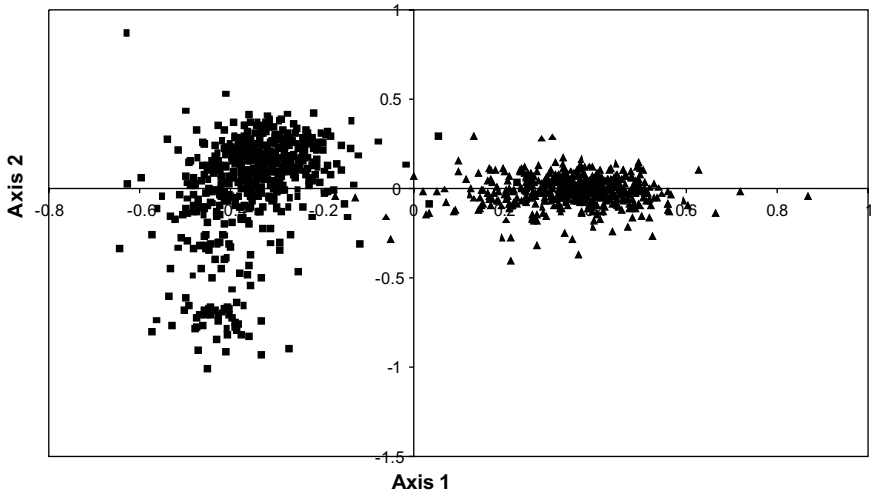

Figure 1: Positions of genes along the first two major axes in the correspondence analysis based on overall RSCU values. Squares represent *B. subtilis* genes; triangles represent *A. aeolicus* genes.

any constraint on synonymous codon usage we also performed CA on simple codon count. It was observed that CA on codon count accounted for 28.50% and 7.13% of the total variation on the first and second major axes respectively. The positions of genes along the first and second major axes produced by CA on codon counts (shown in Fig. 3) are very similar to the figure produced by CA on RSCU values (Fig. 2). Thus it is evident that amino acid composition does not exert any constraint in separating the genes according to their synonymous codon usage.

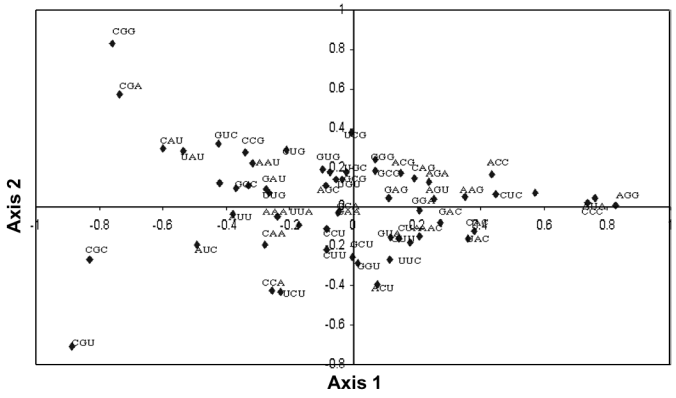

Figure 2: Distribution of synonymous codons along the first and second major axes of the correspondence analysis on RSCU values.

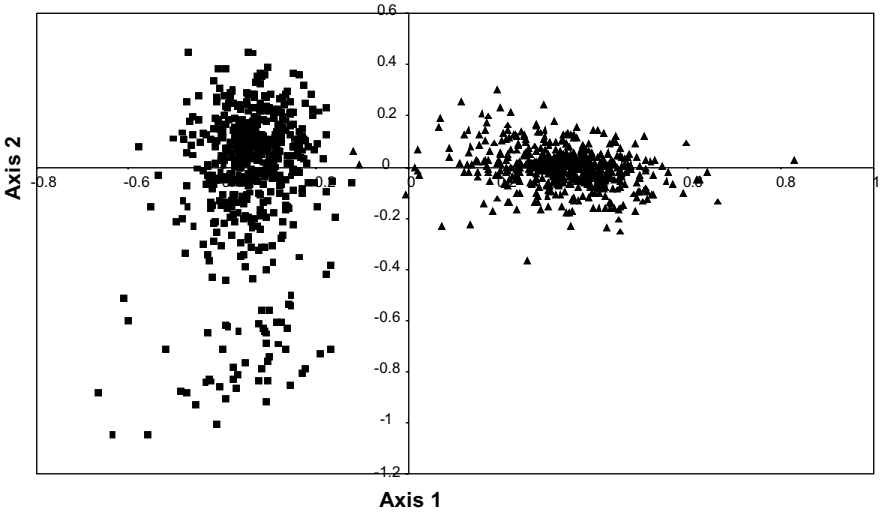

Figure 3: Positions of genes along the first two major axes in the correspondence analysis based on codon counts. Squares represent *B. subtilis* genes; triangles represent *A. aeolicus* genes.

Overall Amino Acid Usage Analysis

Correspondence analysis was carried out on the amino acid composition of the encoded proteins with an aim to understand the possible functional adaptations of the encoded proteins in these two organisms having different optimal growth temperatures. Correspondence analysis on relative amino acid usages actually plot the data in a multidimensional space of 20 axes and then it determines the most prominent axes contributing the relative amino acid usage variation among the genes. CA on relative amino acid usages accounted for 16.98% and 13.77% of the total variation in protein amino acid content respectively. When the positions of the genes against the first two major axes are plotted (Fig. 4) it was observed that the genes along the horizontal axis are highly correlated with the hydrophobicity of the encoded proteins ($r = -0.837$, $p < 0.01$). We have not found any significant correlation between the positions of the genes along the first major axis produced by CA on amino acid usage with GC_3. These results indicate that GC_3 levels have practically no effect in differentiating the genes according to the amino acid usages along the first major axis. In the second major axis the genes are separated according to the growth temperature of the organisms and it was also observed that the positions of the genes along the second major axis is significantly corre-

lated with GC_3 ($r = 0.276$, $p < 0.01$) as well as with $GC_{(1+2)}$ ($r = -0.697$, $p < 0.01$). This trend suggests that both GC_3 and $GC_{(1+2)}$ levels can explain variation along the second major axis according to the amino acid usages.

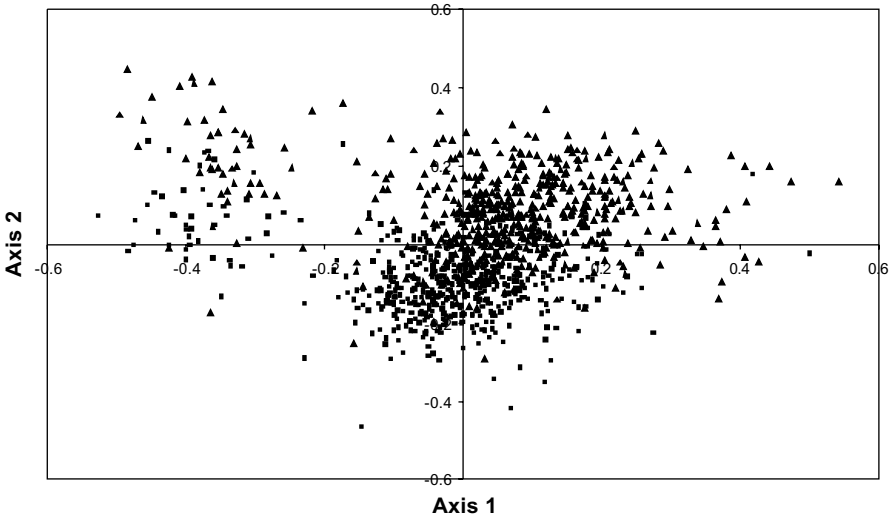

Figure 4: Positions of genes along the first two major axes in the correspondence analysis based on amino acid usage. Squares represent *B. subtilis* genes; triangles represent *A. aeolicus* genes.

It should be noted that CA on overall relative amino acid usage discriminate the genes according to the growth temperature of the organisms along the second major axis whereas CA on overall RSCU or simple codon count differentiates the genes according to the growth temperature of the organisms along the first major axis. Since the variant portions contain the information about the sequence divergence we have performed CA on relative amino acid usages of variant portions only. Figure 5 shows the positions of the genes along the first two major axes. The first and second major axes accounted for 19.65% and 13.15% of the total variation on the relative amino acid usages in this data set respectively. Thus it may be inferred that amino acid composition of invariant portions imposes some constraints on the overall relative amino acid composition, which is responsible in preventing the genes to be discriminated according to their growth temperature on the first major axis. On the other hand overall codon usage variation between the two organisms is strong enough to separate the genes along the first major axis according to their growth temperature. It was also observed that the positions of the genes along the first major axis are significantly correlated with the GC_3 levels of the variant portions of the genes. This result suggests that GC_3 levels can explain variation in amino acid usage along the first major axis according to the growth temperature of the organisms. These results prompted us to make a detailed study to investigate the influences of base composition on the evolution of encoded proteins in different secondary structures of proteins.

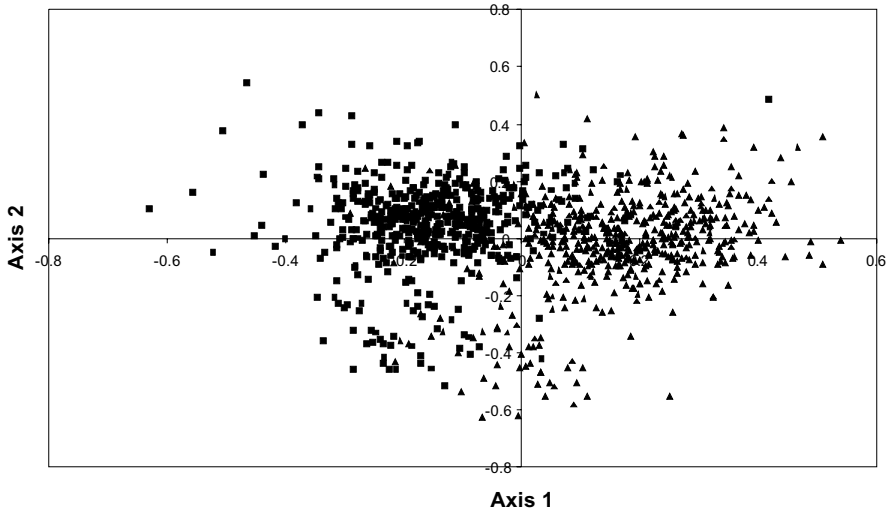

Figure 5: Positions of genes along the first two major axes in the correspondence analysis based on amino acid usage in the variant portion only. Squares represent *B. subtilis* genes; triangles represent *A. aeolicus* genes.

Table I shows the average levels of $GC_{(1+2)}$ and GC_3 in three different predicted secondary structures of the two bacteria and the corresponding p values of their differences.

From the Table I it is evident that in helix structure of thermophilic bacteria, $GC_{(1+2)}$ levels are much lower than its mesophilic counterparts, whereas the opposite is true for the GC_3 levels. The average levels of their differences for both GC_3 and $GC_{(1+2)}$ are quite high and their p values have almost in the same order of magnitude. Since $GC_{(1+2)}$ decreases in the thermophilic organism, so one might expect amino acids whose occurrences are higher in thermophilic organism must be AT-rich. But no conclusion can be drawn whether the increase in GC_3 levels have any role in changing amino acids from mesophilic to thermophilic bacteria.

In coil region it was observed that the differences in average levels of GC_3 is much higher than the differences in average levels of $GC_{(1+2)}$. The pair-wise comparison of GC_3 levels between the two bacteria shows that p value is much lower ($p < 2.71 \times 10^{-40}$) than their $GC_{(1+2)}$ levels ($p < 2.53 \times 10^{-7}$). Since a large number of amino acids whose occurrences are statistically significant in coil structure, the influence of GC_3 in changing one amino acid to another can not be ruled out.

In strand structure the average level of $GC_{(1+2)}$ is does not differ significantly. The levels of GC_3 though differ significantly ($p < 1.09 \times 10^{-3}$), the differences in average levels of GC_3 is not so high as in helix and coil structures.

Correspondence Analysis on Three Different Secondary Structures

Correspondence analyses on RSCU values have been carried out on the helix and coil structures. Correspondence analysis on strand structure could not be performed due to its very small length. It was observed that in both helix and coil structures the genes were separated according to their growth temperatures along first major axis as observed in overall codon usage analyses for the two bacteria.

CA on relative amino acid usages in helix and coil structures gave the same trend as observed in the overall amino acid usages. It was also observed that $GC_{(1+2)}$ levels are significantly correlated with the positions of the genes along the first major axis in both helix ($r = 0.239$; $P < 0.01$) and coil structure ($r = 0.354$; $P < 0.01$). It is also interesting to note that GC_3 levels are significantly correlated with the positions of the genes along the first major axis only in coil structure, but not in the helix structure. These results suggest that the differences in GC_3 levels between the two bacteria have practically no effect in differentiating the genes according to their amino acid usage in the helix structure. Since we have taken two eubacteria having almost similar GC composition, so GC directional pressure is not effective in this study. Undoubtedly, temperature is the most important selection pressure that has to be considered in mesophilic-thermophilic comparison.

Considering the nature of the most important selection pressure in our study, one might be tempted to see whether this selection pressure is equally effective at the protein level as well as nucleotide level. In our study we have seen that ten amino acids (Phe, Leu, Ile, Val, Pro, Tyr, Lys, Glu, Trp, Arg) have been increased in thermophilic *Aquifex aeolicus* (Table II). Surprisingly, the thermolabile amino acids (Asn, Gln, Met, Cys) have been avoided at higher temperature and we can say that this avoidance is due to the selection at higher temperature. So selection pressure is acting on the amino acid level, which supports the results of Kreil *et al*. (25). On the other hand, data of amino acid substitution matrix based on pair-wise alignment of orthologous sequences indicates that a large proportion of lysine in *B. subtilis* has been converted into arginine in *A. aeolicus*. Identical results were observed by

comparing the genomes of *Corynebacterium efficiens* and *Corynebacterium glutamicum* and substitution of lysine to arginine has been described to contribute the protein stability at high temperature (26) and it has been argued that the mechanism of thermo stabilization depends on the resonance stabilization effect of arginine residues (27). Another important substitution is from serine in *B. subtilis* to alanine in *A. aeolicus,* which has earlier been described to stabilize the protein by increasing the hydrophobic interaction (26, 28).

Table II
Amino acid composition of *A.aeolicus* and *B.subtilis* genes.

Amino acid	*A.aeolicus*	*B.subtilis*	p-value
Phe	4.59	3.88	5.38×10^{-08}
Leu	10.12	9.17	8.17×10^{-08}
Ile	7.56	7.28	3.85×10^{-02}
Met	2.09	2.78	2.51×10^{-22}
Val	8.42	7.48	4.64×10^{-13}
Ser	4.56	5.68	1.51×10^{-24}
Pro	4.09	3.59	1.71×10^{-08}
Thr	4.17	5.40	1.55×10^{-34}
Ala	6.21	8.02	9.62×10^{-33}
Tyr	3.64	2.87	5.98×10^{-20}
His	1.63	2.28	2.09×10^{-19}
Gln	2.11	3.60	3.36×10^{-50}
Asn	3.30	3.81	3.90×10^{-08}
Lys	9.58	7.32	3.09×10^{-36}
Asp	4.21	5.27	6.09×10^{-22}
Glu	9.64	7.77	7.93×10^{-23}
Cys	0.84	0.89	n.s.
Trp	0.82	0.71	1.24×10^{-02}
Arg	5.18	4.71	1.24×10^{-03}
Gly	7.15	7.40	n.s.

Although, *B. subtilis* and *A. aeolicus* have same genomic (G+C) composition, all the preferred amino acids in thermophiles cannot be interpreted as evidence for thermal adaptation, because apart from the changes in (G+C) compositions, there are other processes, which could cause differences in amino acid usages (15). McDonald *et al*. (12) showed significant variation of nine amino acids (Glu, Ile, Lys, Met, Asn, Gln, Arg, Ser and Thr) between thermophile and mesophile in the archaeal genus *Methanococcus* and surprisingly these nine amino acids were also varying significantly in our study (Table II). From these results it can reasonably be argued that differences of those nine amino acids between thermophilic and mesophilic comparisons might be due to the adaptation of thermophilic proteins at higher temperature. However, there are other nine amino acids, which were found to vary between *B. subtilis*, and *A. aeolicus* and the differences of these nine amino acids might be due to the several environmental factors other than temperature.

Bio-energetic costs of different amino acids may also have significant role in asymmetrical substitution patterns for various pairs of amino acids. Amino acids with lower bio-energetic costs are favored over functionally equivalent amino acids (14). The relative bio-energetic costs of different amino acids, which were found to vary among different species depends on (i) availability for uptake of each amino acid in the environment, (ii) the biosynthetic pathways used to synthesize each amino acid and (iii) the abundance of raw materials in the environment for biosynthesis (15). For example, the thermophile *A. aeolicus* is a chemolithoautotroph (*i.e*., it uses an inorganic carbon source for biosynthesis) (29). The mesophile *B. subtilis* generally lives in soil, water resources and in association with plants (30). Therefore at aligned sites in protein sequences where two or more amino acids are functionally equivalent, the one that is most abundant in its environment would be favored in *B. subtilis*, while the one with the lowest cost of biosynthesis would be favored in *A. aeolicus*.

At the nucleotide level it has been observed that the coding sequences of thermophiles are relatively rich in purines (31) and it was also argued that purine can serve an adaptive function in promoting RNA stability, possibly by stabilizing tertiary structures (32-34). In our study, we also found an increase in the purine content at higher temperature. We have observed about 3% increase in purine content in the thermophilic bacteria. Moreover, we also obtained a significant correlation between the positions of the genes along the second major axis produced by correspondence analysis on amino acid usages with GC_3 and purine contents.

Singer & Hickey (18) rightly pointed out that purine enrichment cannot provide a complete explanation for changes in amino acid frequencies; and our study clearly demonstrate that along with purine, GC_3 is also a major factor for changes in amino acid frequencies.

Although it has been known that the frequencies of some codons and amino acids correlate with nucleotide content, one aspect has remained unclear: correlations could exist because selection for a particular codon or amino-acid usage produces a particular nucleotide content, or because mutation towards a particular nucleotide content determines codon and amino acid usage. Or these two hypotheses are not mutually exclusive and may both play a role.

Singer & Hickey (18) rejects the hypothesis that the differences at the amino acid level are the reflections of underlying differences at the nucleotide contents. Our results with these two bacteria do not agree with the conclusion made by Singer and Hickey (18) as we got significant correlation between the positions of the genes along the second major axis with GC_3 composition as well as purine content.

In the same article Singer and Hickey (18) considered the alternative hypothesis, *i.e.*, the selection at the amino acid level, could explain the differences at the nucleotide level. They rejected this alternative hypothesis based on the results obtained from correspondence analysis on RSCU and discriminant function analysis and assuming that third codon positions have little relationships to the coding capacity. Our results are also in agreement with the suggestion made by Singer and Hickey. In our study CA on RSCU values shows that the genes can be separated along the first major axis according to the growth temperature of the organisms. CA on codon counts is very similar to the figure produced by CA on RSCU values. Thus it is evident that amino acid composition does not exert any constraint in separating the genes according to their synonymous codon usage.

In summary, our study reveals that selection pressure is acting on both the amino-acid level as well as on nucleotide level and this action of selection pressure has been observed in all the three protein secondary structures. Purine and GC_3 enrichment are pervasive and occurs in all three secondary structures. GC_3 composition influences the amino acid usage only in the coil structure but not in helix structure. Regardless of the structural unit at which the selection takes place, a positive outcome of our study reveals that natural selection maximizes the biological function of both protein and DNA at high growth temperature.

Acknowledgements

Authors are thankful to the Department of Biotechnology, Government of India, for the financial help.

References and Footnotes

1. G. Bernardi and G. Bernardi. *J. Mol. Evol. 24*, 1-11 (1986).
2. G. Bernardi. *Gene 241*, 3-17 (2000).
3. A. Wada and A. Suyama. *Prog. Biophys. Mol. Biol. 47*, 113-157 (1986).
4. N. Galtier and J. R. Lobry. *J. Mol. Evol. 44*, 632-636 (1997).
5. A. Muto and S. Osawa. *Proc. Natl. Acad. Sci. 84*, 166-169 (1987).

6. L. D. Hurst and A. R. Merchant. *Proc. R. Soc. Lond. 268*, 493-497 (2001).
7. H. Nakashima, S. Fukuchi and K. Nishikawa. *J. Biochem. 133*, 507-513 (2003).
8. N. Sueoka. *Proc. Natl. Acad. Sci. 47*, 1141-1149 (1961).
9. G. A. C. Singer and D. A. Hickey. *Mol. Biol. Evol. 17*, 1581-1588 (2000).
10. G. D'Onofrio, K. Jabbari, H. Musto, and G. Bernardi. *Gene 238*, 3-14 (1999).
11. X. Gu, D. Hewett-Emmett, and W. H. Li. *Genetica 103*, 383-391 (1998).
12. J. H. McDonald, A. M. Grasso, and L. K. Rejto. *Mol. Biol. Evol. 16*, 1785-1790 (1999).
13. P. J. Haney, J. H. Badger, G. L. Buldak, C. I. Reich, C. R. Woose, and G. J. Olsen. *Proc. Natl. Acad. Sci. 96*, 3578-3583 (1999).
14. C. L. Craig and R. S. Weber. *Mol. Biol. Evol. 15*, 774-776 (1998).
15. J. H. McDonald. *Mol. Biol, Evol. 18*, 741-749 (2001).
16. S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, T. Ikemura, *et al*. *Gene 276*, 89-99 (2001).
17. D. J. Lynn, G. A. C. Singer, and D. A. Hickey. *Nucleic Acids Res. 30*, 4272-4277 (2002).
18. G. A. C. Singer and D. A. Hickey. *Gene 317*, 39-47 (2003).
19. J. R. Lobry and N. Sueoka. *Genome Biology 3*, 1-14 (2002).
20. G. Deleage, C. Blanchet, and C. Geourjon. *Biochimie 79*, 681-686 (1997).
21. M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, UK (1984).
22. S. Majumder, S. K. Gupta, V. S. Sundararajan, and T. C. Ghosh. *Biochem. Biophys. Res. Commun. 266*, 66-71 (1999).
23. S. K. Gupta, S. Majumder, T. K. Bhattacharya and T. C. Ghosh. *Biochem. Biophys. Res. Commun. 269*, 692-696, (2000).
24. W. Gu, T. Zhou, J. Ma, X. Sun and Z. Lu. *BioSystems 73*, 89-97 (2004).
25. D. P. Kreil and C. A. Ouzounis. *Nucleic Acids Res. 29*, 1608-1615 (2001).
26. Y. Nishio, Y. Nakamura, Y. Kawarabayasi, Y. Usuda, E. Kimura, S. Sugimoto, K. Matsui, A. Yamagishi, H. Kikuchi, K. Ikeo, and T. Gojobori. *Genome Research 13*, 1572-1579 (2003).
27. C. Vieille and G. J. Zeikus. *Microbiol. Mol. Biol. Rev. 65*, 1-43 (2001).
28. W. R. Taylor. *J. Theor. Biol. 119*, 205-218 (1986).
29. G. Deckert, P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huber, R. A. Feldman, J. M. Short, G. J. Olsen, and R. V. Swanson. *Nature 392*, 353-358 (1998).
30. F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, *et. al. Nature 390*, 249-256 (1997).
31. P. J. Lao and D. R. Forsdyke. *Genome Research 10*, 228-236 (2000).
32. E. Schultes, P. T. Hraber, and T. H. LaBean. *RNA 3*, 792-806 (1997).
33. H. C. Wang and D. A. Hickey. *Nucleic Acids Res. 11*, 2501-2507 (2002).
34. J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, A. A. Szewezak, C. E. Kundrot, T. R. Cech, and J. A. Doudna. *Science 273*, 1996-1999 (1996).

Date Received: May 12 2004

Communicated by the Editor Ramaswamy H Sarma