Bioinformation
www.bioinformation.net

open access
Hypothesis

Genomic adaptation of prokaryotic organisms at high temperature

Surajit Basak[1], Pamela Mukhopadhyay[2], Sanjib Kumar Gupta[2], Tapash Chandra Ghosh[2*]

[1]Biomedical Informatics Center, National Institute of Cholera and Enteric Diseases, P-33, C.I.T Road, Scheme-XM, Beliaghata, Kolkata-700010, India; [2]Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M -Kolkata 700054- India; Tapash Chandra Ghosh Tel.: +91-33-2355 6626; fax: +91-33-2355 3886 E-mail: tapash@bic.boseinst.ernet.in, *Corresponding author.

Received November, 5 2009; Accepted November, 15 2009; Published February 28, 2010

Abstract:
One of the central issues of evolutionary genomics is to find out the adaptive strategies of microorganisms to stabilize nucleic acid molecules under high temperature. Thermal adaptation hypothesis gives a link between G+C content and growth temperature if there is a considerable variation of guanine and cytosine content between species. However, there has been a long-standing debate regarding the correlations between genomic GC content and optimal growth temperature (Topt). We urged that adaptation to growth at high temperature requires a coordinated set of evolutionary changes affecting: (i) nucleic acid thermostability and (ii) stability of codon-anticodon interactions. Moreover, in *Bacillaceae* family we have demonstrated that a higher genomic GC level do not have any role in stabilizing mRNA secondary structure at high growth temperature. Comparative analysis between homologous sequences of thermophilic *Thermus thermophilus* and mesophilic *Deinococcus radiodurans* suggests that increased levels of GC contents in the coding sequence corresponding to strand structure of *Thermus thermophilus* genes have stabilizing effect on the mRNA secondary structure, whereas increased levels of GC contents in coding sequences corresponding to aperiodic structure have destabilizing effect on the mRNA secondary structure. In this perspective, a critical review of thermal adaptation hypothesis is further advocated.

Background:
For a long time, the central issue of evolutionary genomics was to find out the adaptive strategy of nucleic acid molecules towards different optimal growth temperatures (Topt). Heat-loving microorganisms are called thermophiles. They are the most studied extremophiles and are generally found in hot springs and deep sea vents. Thermophiles thrive at temperatures that would scald or kill most other types of life. Regular thermophiles are those that thrive at temperatures greater than 55° C. Hyperthermophiles grow optimally at above 80° C. On the other hand, microorganisms grow below 55° C are known as mesophiles.

All macromolecules such as DNA and RNA must be stable and functional at higher growth temperature. Thermal adaptation hypothesis argues in favor of a positive correlation between G+C content and growth temperature as G: C pair is thermally more stable than A: T pair **[1].** However, some authors have failed to make any correlation between GC composition and temperature **[2, 3].** Later, Musto et al. **[4]** claimed that positive correlation between genomic GC content and optimal growth temperature exists in most of the bacterial families they studied; nevertheless strong criticisms were raised against their work **[5, 6].**

Codon-anticodon interaction and its influence on codon usage strategy at high temperature
Considering the above contradicting views, we have proposed a new hypothesis [7] after considering the influence of codon-anticodon interaction on the correlation between genomic G+C content and T_{opt}. Our hypothesis is based on the fact that codon-anticodon interaction energy has a selective advantage on codon usage strategy for better translational efficiency at high growth temperature. We showed that the frequency of WWC and WWG groups of codons increase significantly, whereas the frequency of SSC and SSG groups of codons decrease significantly with growth temperature (W=A or T and S=G or C). It was previously noted **[8]** that occurrence of side-by-side GC base pairs are avoided in the highly expressed genes of yeast and such an avoidance has been interpreted just to eliminate very sticky codon-anticodon interaction so that translation can proceed smoothly. For thermophilic bacteria we proposed that the preference of WWC and WWG groups of codons at high growth temperature might be due to the maintenance of codon-anticodon interaction energy at an intermediate strength since no side-by-side GC base pairing occurs for these two codon groups. On the other hand, strong avoidance for GC-rich codon groups (SSC and SSG) at high growth temperature might be due to the influence of side-by-side GC base pairing to exceed the codon-anticodon interaction energy from intermediate strength. Thus codon-anticodon interactions might have some influence on the relationship between the genomic G+C content and T_{opt} to optimize translational efficiency. On the contrary, structural RNAs, being non-protein-coding genes, codon-anticodon interaction is in no way related to these genes and hence at high growth temperature structural RNA molecules might increase their GC composition in order to have greater thermal stability **[9].**

Methodology:
The complete genome sequences have been downloaded from ftp.ncbi.nlm.nih.gov/genbank/genomes and the coding sequences were extracted. Homologous sequences between *Thermus thermophilus* and *Deinococcus radiodurans* were identified by gapped BLASTP program **[10]** using cutoff of $e=1.0 \times 10^{-3}$. Hits less than 50% identity were removed from the dataset. The maximum gap size allowed between a pair of sequences is 5%. All hypothetical coding sequences, as well as genes having less than 100 codons were ignored. Gene pairs having size difference lower than or equal to 30 codons were retained. Finally 382 gene pairs were extracted for data analyses. The protein secondary structures corresponding to each of the sequences were downloaded from Consensus Secondary Structure Prediction Program available at http://npsa-pbil.ibcp.fr **[11]**. CodonW 1.4.2 (J. Peden, 200; http://www.molbiol.ox.ac.uk/cu/) was used to calculate the DNA base compositions and codon frequencies.

For each native mRNA sequence, 50 random sequences were generated using the randomization protocol, CodonShuffle **[12],** which randomly permutes synonymous codon in codon degenerate family preserving the exact count of each codon and order of encoded amino acids as in the original transcript. The mfold program (available at http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form4.cgi) was used to predict free-folding energies for mRNA sequences. The difference in the free energy of folding between the native sequence and the corresponding random sequences was measured by the Z-score, given by $Z\text{-score} = \{E_{native} - \langle E_{random} \rangle\}/STD$, where E_{native} denotes the folding free energy of native mRNA sequence, $\langle E_{random} \rangle$ denotes the average folding free energy over a large number of randomized sequences generated from the native sequence and STD denotes its standard deviation.

The nucleotide sequence of the protein secondary structure conserved between the two species was considered for further analysis of mRNA free folding energy contributed from each of the three protein secondary structures. To measure the contribution of each protein secondary structures on Z-score, only the fraction of the transcript encoding for a particular protein secondary structure was randomized and then the Z-score for the whole transcript was calculated. In this way, we measured $Z_{aperiodic,}$ Z_{helix} and Z_{strand} for part randomization of aperiodic, helix and strand structures respectively. Students *t test* was used to evaluate the significance of the pair-wise differences in the three protein secondary structures.

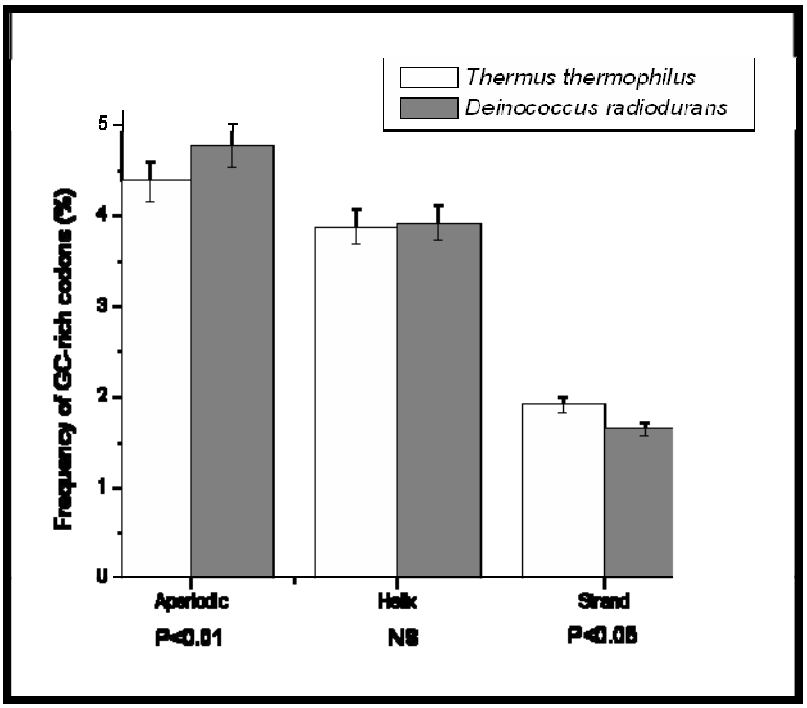

Figure 1: Frequency of SSS codons in three protein secondary structures of *T. thermophilus* and *D. radiodurans*.

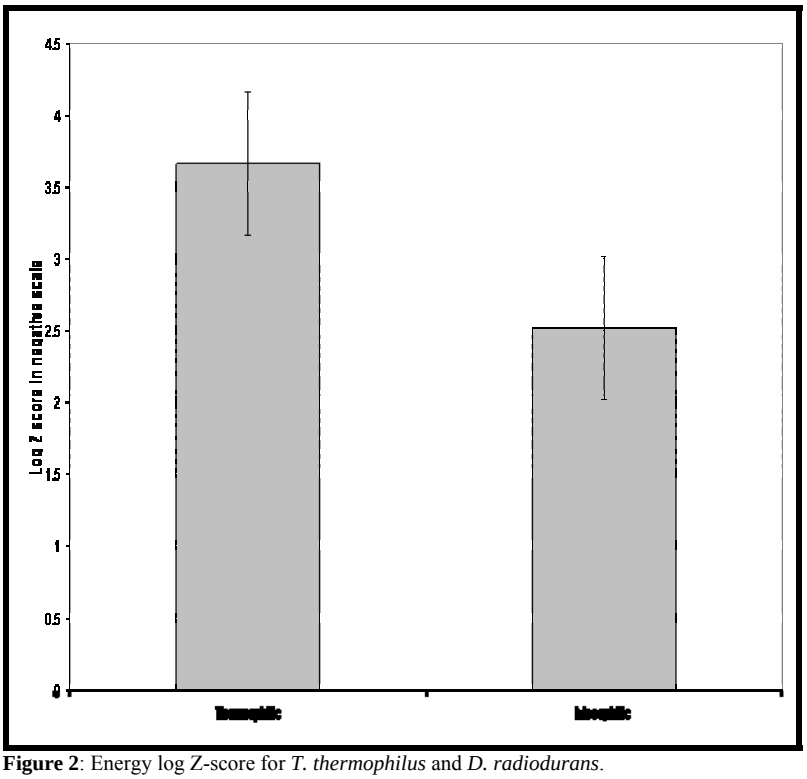

Figure 2: Energy log Z-score for *T. thermophilus* and *D. radiodurans*.

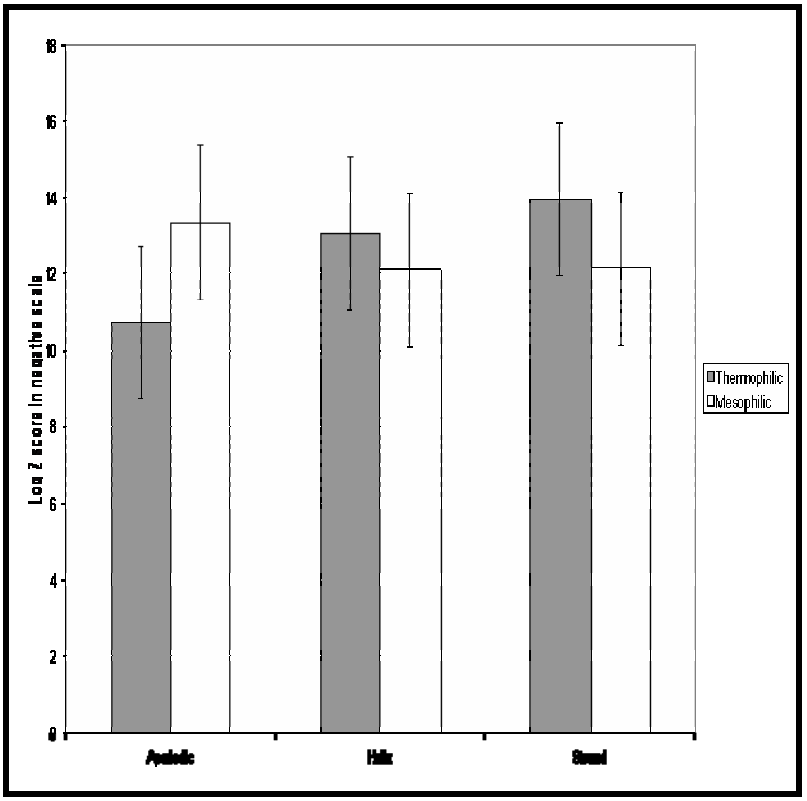

Figure 3: Contribution of mRNA regions corresponding to three protein secondary structures on energy log Z-score for *T. thermophilus* and *D. radiodurans.*

Results and Discussion

Implications of higher GC-level towards genomic thermal adaptation

To investigate whether a higher GC-level of thermophilic prokaryotes increases the thermodynamic stability of mRNA secondary structure, we analyzed the *Bacillaceae* family where highest correlation coefficient in the regression analysis between genomic G+C content and optimal growth temperature has been observed by Musto et al. **[4]**. We observed significant positive correlation ($r=0.682$, $p<0.05$), similar to the observation made by Musto et al [4], between average G+C content of genes available for each species and optimal growth temperature. Higher G+C content at higher growth temperature implies that mRNA molecules of thermophilic genomes should have higher stability than the mesophilic genomes. In order to examine the above hypothesis, we have measured the correlation coefficient between average G+C content against average free folding energy (ΔG) (used to measure mRNA stability for a predicted RNA secondary structure **[13]** (http://dinamelt.bioinfo.rpi.edu/zipfold.php)) of mRNA molecules. A positive correlation was observed ($r=0.42$, $p<0.05$) between average G+C content and average free folding energy (ΔG) of mRNA molecules. Here, increase of ΔG with the increase in G+C levels indicates that higher G+C levels at higher growth temperature does not have any role to stabilize mRNA secondary structure. Similar trend has been observed when we performed the analysis by taking homologous set of genes from the species belong to *Bacillaceae* family (data not shown). Although, in the thermodynamic stability hypothesis it was presumed that a higher G+C level should represent more stable mRNA, but the present analysis suggest some other functional relevance of higher G+C level at higher temperature. A negative correlation between mRNA stability and GC_3 content for *Drosophila melanogaster* gene sequences has been observed by Stenoien and Stephan **[14].** These authors proposed an alternative explanation that guanine and cytosine might be confined to the unpaired regions of secondary structural elements and thus not a contributor to the stability of mRNA molecule. However, for rRNA molecule, Wang and Hickey **[9]** observed that increased G+C content of thermophiles is almost entirely concentrated within the double-stranded stem regions. But the biochemical basis for the increase of G+C content with optimal growth temperature for mRNA molecule in the *Bacillaceae* family remains to be elucidated. It is well known that stable secondary structures play a role in transcriptional regulation **[15]** and translation **[16]** and the stability of mRNA secondary structure would have the greatest inhibitory effect on translation **[16, 17].** The decrease in stability of mRNA secondary structure at high growth temperature might therefore be related to an increase in translation rates.

Does variation in GC content always influence mRNA secondary structural stability at high growth temperature? A case study between T. thermophilus and D. radiodurans homologous genes.

Mesophile-thermophile pair may differ in other environmental variables, apart from the optimal growth temperature. We have taken two closely related organisms namely, *Thermus thermophilus* and *Deinococcus radiodurans* where phylogenetic relationships are expected to be fairly accurate **[18, 19, 20].** Optimal growth temperature of *Deinococcus radiodurans* and *Thermus thermophilus* are 30°C and 73°C respectively (www.dsmz.de). We have retrieved 382 homologous gene pairs **(Table 1 in supplementary material)** from these two related eubacteria. It is known that translational speed is not uniform along the coding sequence and varies according to the secondary structure of the encoded proteins; hence analyses of the coding sequences corresponding to the secondary structures of the encoded proteins are very much effective in extracting biological information of a gene.

The genomic GC levels of thermophilic bacteria, *Thermus thermophilus* is 69.4% and that of its mesophilic counterpart *Deinococcus radiodurans* is 66.6%. The same trend has been observed

in the average levels of GC and GC_3 of all the homologous gene pairs as well as in coding regions corresponding to different protein secondary structures of *T. thermophilus* and *D. radiodurans* genes **(Table 1 in supplementary material).**

We then calculated the cumulative frequency of all GC-rich codons, i.e. all SSS codons (S= G or C) in the coding sequences corresponding to the three protein secondary structures of *T. thermophilus* and *D. radiodurans* genes. It is interesting to note that the cumulative frequency of GC-rich codons in the coding sequences of three proteins secondary structures are not always higher in *T. thermophilus* than *D. radiodurans* **(Figure 1)**. In the aperiodic structures, GC-rich codons are under-represented in *T. thermophilus* as compared to *D. radiodurans*, whereas it does not vary significantly in the helix structure. On the other hand, GC-rich codons are preferred in *T. thermophilus* compared to *D. radiodurans* in the strand structure. It is well-established that the reduction in the frequency of GC-rich codons is the consequence to avoid the formation of side-by-side GC base pairs in the codon–anticodon interaction for better translational efficiency **[7, 8].** Thus, our results suggest that the genomic segments coding for aperiodic structure of the encoded protein are responsible for better translational efficiency at high growth temperature.

It was suggested that potential secondary structures of single-stranded mRNA may play important functional roles in translational regulation and gene expression by modulating synonymous codon usage **[17, 21, 22, 23].** Here, we have determined the stability of mRNA secondary structure by the minimization of the free energy of the molecule.
Z-score for *T. thermophilus* genes has been observed to be significantly lower than *D. radiodurans* **(Figure 2)**. It indicates that of mRNA secondary structure of *T. thermophilus* is more stable than that of *D. radiodurans*. We then compared the contributions of mRNA regions corresponding to three protein secondary structures ($Z_{aperiodic,}$ Z_{helix} and Z_{strand}) between *T. thermophilus* and *D. radiodurans* homologous gene pairs (Figure 3). $Z_{aperiodic}$ is significantly higher in *T. thermophilus* genes than that of *D. radiodurans*. No significant difference in Z_{helix} values has been observed between *T. thermophilus* and *D. radiodurans* homologous gene pairs. However, significantly lower Z_{strand} has been observed for *T. thermophilus* compared to *D. radiodurans*. These results further suggest that the higher folding stability of mRNA secondary structure of *T. thermophilus* compared to *D. radiodurans* is contributed from the coding sequences related to the strand structure of protein. Higher Z-score of *T. thermophilus* contributed from the aperiodic structure indicates that, these segments of coding sequence are actually responsible for better translational efficiency **[17].**

Conclusions:
Prokaryotic genomes have wide compositional range where GC levels vary approximately from 25% to 75%. It is a long standing debate whether the wide variation in base composition is under selection in prokaryotes. Based on the analysis of genomic GC levels and optimal growth temperature in prokaryotic families, Musto et al. **[4]** claimed that optimal growth temperature influences the variation of GC levels in prokaryotes. However, this aspect remains very controversial **[5, 6, 24].** The selective advantage of higher GC level is the thermodynamic stability required by DNA, RNA and proteins at high growth temperature **[1].** However it is also important to analyze if higher GC levels are always associated with higher thermodynamic stability of mRNA molecules at high growth temperature.

Here, we observed that GC content is significantly higher in *T. thermophilus* as compared to *D. radiodurans* in all the protein secondary structures. However nucleotide sequences corresponding to all three protein secondary structures do not have the same influence on mRNA secondary structural stability. Nucleotide sequences corresponding only to the strand structure have the stabilizing influence on mRNA secondary structure. On the other hand, there is a strong tendency to avoid the formation of side-by-side GC-base pair in the aperiodic structure which suggests its important contribution to maintain translational efficiency. In our previous article **[7],** we argued that an organism's strategy to survive at high temperature is linked to its ability to produce sufficient amounts of functional gene products under the extreme thermal condition. This might be achieved through a combination of adaptive changes at several molecular levels in a proper synchronization, i.e., stability at one molecular level cannot be achieved at the cost of other biological processes(s). Thus all these analyses lead us to conclude that the positive outcome has always been to maximize biological function at high growth temperature.

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

References
[1] G. Bernardi *et al., J Mol Evol*, **24:** 1 (1986) [PMID: 3104608]
[2] N. Galtier *et al., J Mol Evol*, **44:** 632 (1997) [PMID: 9169555]
[3] L.D. Hurst *et al., Proc Biol Sci,* **268:** 493 (2001) [PMID: 11296861]
[4] H. Musto *et al., FEBS Lett*, **573**: 73 (2004) [PMID: 15327978]
[5] S.A. Marashi *et al., Biochem Biophys Res Commun*, **325**: 381 (2004) [PMID: 15530402]
[6] S. Basak *et al., Biochem Biophys Res Commun*, **327:** 969 (2005) [PMID: 15652489]
[7] S. Basak *et al., Biochem Biophys Res Commun,* **330:** 629 (2005) [PMID 15809043]
[8] J.L. Bennetzen *et al., J Biol Chem*, **257:** 3026 (1982) [PMID: 7037777]
[9] H.C. Wang *et al., Nucleic Acids Res,* **30:** 2501 (2002) [PMID: 12034839]
[10] S.F. Altschul *et al., Nucleic Acids Res*, **25:** 3389 (1997) [PMID: 9254694]
[11] G. Deleage *et al., Biochimie*, **79:** 681 (1997) [PMID: 9479451]
[12] L. Katz *et al., Genome Res*, **13:** 2042 (2003) [PMID: 12952875]
[13] N.R. Markham *et al., Nucleic Acids Res*, **33:** W577-W581 (2005) [PMID: 15980540]
[14] H.K. Stenoien *et al., J Mol Evol*, **61:** 306 (2005) [PMID: 16044249]
[15] M. Weik *et al., J Virol,* **76:** 8532 (2002) [PMID: 12163572]
[16] T.D. Schmittgen *et al., J. Biol. Chem*, **269:** 16269 (1994) [PMID: 8206932]
[17] M. Jia *et al., FEBS Lett*, **579:** 5333 (2005) [PMID: 16194536]
[18] J.H. McDonald, *Mol Biol Evol*, **18:** 741 (2001) [PMID: 11319258]
[19] R. Hensel *et al., Int J Syst Bacteriol*, **36:** 444 (1986)
[20] W.G. Weisburg *et al., Syst Appl Microbiol*, **11:** 128 (1989) [PMID: 11542160]
[21] S. Basak *et al., FEBS Lett*, **581:** 5825 (2007) [PMID: 18047839]
[22] S. Basak *et al., FEBS Lett*, **580:** 3895 (2006) [PMID: 16797014]
[23] D.B. Carlini *et al., Genetics*, **159:** 623 (2001) [PMID: 11606539]
[24] H. Musto et al., *Biochem Biophys Res Commun*, **330:** 357 (2005) [PMID: 15796890]

Edited by P. Kangueane
Citation: Basak ***et al,*** Bioinformation 4(8): 352-356 (2010)

Supplementary material:

Table 1: GC and GC_3 compositions (%) of *T. thermophilus* and *D. radiodurans*

	T. thermophilus				*D. radiodurans*			
	Overall	Aperiodic	Helix	Strand	Overall	Aperiodic	Helix	Strand
GC	69.10	73.00	68.08	61.46	66.85	70.57	66.14	57.96
GC_3	93.36	92.68	93.88	93.83	86.29	86.81	85.76	87.49