Nucleotide Composition and Amino Acid Usage in AT-Rich Hyperthermophilic Species

Subhash Mohan Agarwal[1,*] and Atul Grover[2]

[1]*Bioinformatics Center, School of Information Technology, Jawaharlal Nehru University, New Delhi 110067, India and* [2]*Department of Bioscience and Biotechnology, Banasthali University, Banasthali 304022, India*

Abstract: Nucleotide composition, codon usage and amino acid content are important molecular signatures that vary in different groups of organisms. AT-rich (or GC poor) hyperthermophiles have relatively been unexplored in these aspects. In this study, we have examined the compositional characteristics of AT rich genomes viz. *Methanococcus jannaschii*, *Sulfolobus solfataricus*, *Sulfolobus tokodaii* and *Nanoarcheum equitans* by their comparison with four mesophiles having similar genomic GC content. The analysis revealed a significant increase in purine content of ORFs due to increase in guanine content. Moreover, the influence of dinucleotide composition on protein thermostability was found even larger. Accordingly, increased usage of codons that are constituted of dinucleotides RR was observed. Arginine, proline, valine and tyrosine were most abundant amino acids in hyperthermophilic proteomes, and similar bias was seen when dipeptidic composition of proteins was compared. Further amino acid composition analysis of alpha helices indicates an increased usage of E, K, R and decreased usage of N and Q. Summing up, the study suggested that elevated growth temperature impose selective constraints at all the three molecular levels- nucleotide composition, codon usage and amino acid content.

Keywords: Hyperthermophiles; nucleotide bias; codon usage; amino acid composition.

INTRODUCTION

Hyperthermophiles constantly face the challenge of maintaining the stability of their genome. Increasing the melting point of their DNA by keeping relatively higher GC [1] is one of the methods they have constituted to address the issue. However, the GC content of the genomes does not correlate with optimal growth temperature (OGT) [2, 3]. Various additional attributes have been suggested that contribute in maintaining the stability of genomic DNA of hyperthermophiles [2, 3]. Infact a number of hyperthermophiles have GC content of their DNA lesser than 40% [1]. On the other hand GC content of rRNA and tRNA show strong correlation with optimal growth temperature [4, 5]. Various studies have established that these living organisms are subject to a variety of selection pressures that act not only at the level of global phenotype but at each level of the cell's organization i.e. DNA, RNA and proteins [6]. For example, there is evidence that the proteins of thermophiles are characterized by a distinct pattern of amino acids [7-10]. Moreover a difference in the pattern of synonymous codon usage between thermophiles and mesophiles has been observed [7, 10].

Although considerable studies have focused on understanding the mechanisms that makes life possible under these conditions, it still remains unclear that whether it is due to external conditions or natural selection [4, 7, 11-14]. In order to infer the molecular mechanistic adjustments to the thermal stress, it is desired to compare the genomic characteristics of hyperthermophiles with mesophilic genera. Singer and Hickey [14] made such an attempt considering the genera that show optimal growth temperature (OGT) near or above 50°C, while the AT-rich hyperthermophilic genomes were ignored in their analysis. Das *et al.* [5] looked into some of the hyperthermophilic genomes that had their GC content lower than 50%. A shortcoming of this study was the broad range of OGT (>13°C) over which the genera under study varied. Thus, in order to minimize the ascertainment bias in terms of codon usage and nucleotide composition between different species we have picked up various mesophiles and hyperthermophiles in an even narrower OGT range of 7.8°C for comparative analysis among mesophiles and hyperthermophiles.

The hyperthermophilic archaebacteria, *Nanoarchaeum equitans* is one of the interesting examples qualifying for this kind of analysis. The archaebacteria is known to host smallest non viral genome to date, which spans 490 Kb and is constituted of 537 protein coding genes [15]. The genome displays short intergenic regions, large number of split genes, few pseudogenes, and lacks many of the vital metabolic genes [15]. Further phylogenetic analysis suggested that it diverged early in archaeal lineage even before the emergence of Euryarchaeota and Crenarchaeota, representing basal archaeal lineage [16]. Considering *Nanoarcheum equitans* to be one of the simplest genome of cellular organisms and of course simplest among the genera under study, genomic features of *N. equitans* have been dealt as a special case within the hyperthermophilic group.

Thus, the present paper outlines comparisons of nucleotide bias, codon usage patterns and amino acid bias drawn between mesophiles and hyperthermophiles having average GC content close to 31%.

MATERIALS AND METHODS

The coding sequences (CDS) and the corresponding amino acid sequences for all of the eight genomes (Table **1**) were downloaded from ftp site of GenBank. Following CDS

*Address correspondence to this author at the Bioinformatics Center, School of Information Technology, Jawaharlal Nehru University, New Delhi 110067, India; E-mail: smagarwal@yahoo.co.in

1875-0362/08

Table 1. List of Organisms Studied in the Analysis

Species Name	Abbreviation	GC Content	OGT (°C)
Mesophile			
Campylobacter jejuni	*Cjej*	31	43
Borrelia burgdorferi	*Bbur*	28	37
Lactococcus lactis	*Llac*	35.3	30
Rickettsia prowazekii	*Rpro*	29	35
Hyperthermophile			
Methanococcus jannaschii	*Mjan*	31.3	85
Sulfolobus solfataricus	*Ssol*	35.8	80
Sulfolobus tokodaii	*Stok*	32.8	80
Nanoarchaeum equitans	*Nequ*	31.6	90

integrity check a number of genes were excluded from analysis. For a CDS to be selected presence of a start and stop codon at the beginning and end of each CDS respectively, along with no detectable frameshift was required. Moreover CDS that were smaller than 300 nucleotides were removed. The genes thus shortlisted were analyzed for base compositional bias by studying the prevalence of mononucleotide bases and combinations of dinucleotide bases. CodonW (available from http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html) was used for calculating number of each codon and relative synonymous codon usage (RSCU) for each of the gene within a genome. Similarly, amino acid and dipeptide compositional bias was calculated in the predicted peptides. Secondary structures were predicted using GOR(IV) [11] to study the bias in frequencies of amino acids in three dimensional helical structures. Subsequently to identify patterns showing significant differences between the two groups (mesophilic and hyperthermophilic) t-test was performed. Initially mean values for 4 mesophilic and 3 hyperthermophilic genomes excluding *N. equitans* was evaluated, to derive a general pattern of similarities and differences between the groups. Later the mean value for thermophilic genomes with *N. equitans* was calculated and compared with mesophilic genomes mean to garner information regarding *N. equitans* adaptation towards environment.

RESULTS AND DISCUSSION

NUCLEOTIDE COMPOSITIONAL BIAS AS THERMOPHILIC ADAPTATION

On comparison of occurrence of different nucleotides in four mesophiles and four thermophiles, a significant decrease in the thymine content in thermophiles was noted (Table **2**). This was coupled with an overall increase in purine content (A+G) in thermophilic genomes. Similarly, a significant increase in the dinucleotide pairs AG, GA and GG was also found in hyperthermophiles (Table **3**) coupled with a fall in the frequency of the pair TT. The long standing hypothesis is that GC-richness of protein coding genes can not be considered as thermophilic signatures [17]. The above results confirm the findings of Paz *et al.* [13] who reported abundance of polypurine tracts in thermophilic mRNA sequences, as purine loading of mRNAs is expected to reduce RNA-RNA interactions and thus prevent formation of double stranded RNA molecules [18]. Subsequently, frequency of each nucleotide for each of the three codon positions is analyzed. Although no significant difference is observed for any of the nucleotide at the first and second codon position, however, a marked decrease in the thymine content and a corresponding significant increase in guanine content was seen at the third codon position (Table **2**). This may be considered as a means to increase the GC content at codon third sites (GC3). Hurst and Merchant [17] suggested maintenance of higher GC3 by thermophiles. However, Singer and Hickey [14] reported a very significant increase in adenine at all codon position and decrease in cytosine at first and second codon position. Singer and Hickey [14] hypothesized the significant increase in purine content due to the increase in frequency of adenine. The same hypothesis does not hold true when analyzed for GC poor genomes. On the other hand, the increase in purine amount in these genomes was found due to the increase in overall guanine content. Further, purine richness of codons in terms of AG, GA and GG as two of the three bases in codons is also likely to determine the supercoiling of double stranded DNA which affects the thermostability in the absence of nucleosome structures.

CODON USAGE AND RCSU BIAS IN HYPERTHERMOPHILES

There is a recent interest of scientific community to relate the codon usage with OGT of an organism [19]. Obviously, elevated growth temperatures impose selective constraints on codon-anticodon interactions as well [20]. Our comparisons on codon usage revealed changes in the absolute frequency of 11 codons- significant increase in the frequency of five codons (GAG, AAG, CCA, AGG and AGA) coding for glutamic acid, lysine, proline and arginine respectively and a significant decrease in the three codons (AAU, CAA and CUU) coding for asparagine, glutamine and leucine in the hyperthermophiles (Table **4**). Including *N. equitans* in statistical analysis led to fall in the frequencies of three additional codons i.e., ATT, CGT and CGC encoding isoleucine and

Table 2. Distribution of Nucleotide Frequencies

Nucleotide	*Cjej*	*Bbur*	*Rpro*	*Llac*	Avg-Meso	*Mjan*	*Ssol*	*Stok*	Avg-Thermo		*Nequ*	Avg-thermo+*Nequ*	
A	36.2	37.7	36.7	33.0	35.9	38.2	34.4	35.7	36.1	ns	39.7	37.0	ns
G	18.1	16.9	17.2	19.9	18.0	20.7	21.7	19.9	20.8	*	18.3	20.2	ns
C	12.8	12.0	13.3	16.4	13.6	11.3	14.7	13.7	13.2	ns	12.9	13.2	ns
T	32.9	33.4	32.8	30.7	32.5	29.8	29.2	30.7	29.9	*	29.1	29.7	**
G+C	30.9	28.9	30.5	36.3	31.7	32.0	36.4	33.6	34.0	ns	31.2	33.3	ns
G+A	54.3	54.6	53.9	52.9	53.9	58.9	56.1	55.6	56.9	*	58.0	57.2	**
A(1)	35.9	39.2	37.6	31.4	36.0	37.6	35.5	35.5	36.2	ns	37.5	36.5	ns
G(1)	29.9	27.1	27.7	32.8	29.4	32.9	31.0	30.7	31.5	ns	29.1	30.9	ns
C(1)	12.8	10.9	13.5	15.9	13.3	8.6	12.3	12.0	11.0	ns	10.6	10.9	ns
T(1)	21.4	22.8	21.3	19.9	21.4	20.9	21.3	21.7	21.3	ns	22.8	21.7	ns
A(2)	36.7	37.4	34.7	34.0	35.7	37.1	32.5	32.9	34.2	ns	38.2	35.2	ns
G(2)	14.0	12.2	13.2	13.7	13.3	13.9	15.2	14.4	14.5	ns	12.7	14.0	ns
C(2)	16.5	15.9	18.5	20.9	18.0	15.8	18.4	18.7	17.6	ns	16.2	17.3	ns
T(2)	32.9	34.5	33.5	31.3	33.1	33.2	33.8	34.0	33.7	ns	32.9	33.5	ns
A(3)	36.1	36.3	38.1	33.8	36.1	40.0	35.1	38.7	37.9	ns	43.4	39.3	ns
G(3)	10.2	11.5	10.7	13.0	11.4	15.3	18.8	14.3	16.1	*	12.9	15.3	*
C(3)	16.9	9.2	11.5	12.3	12.5	13.2	13.7	10.5	12.5	ns	12.0	12.4	ns
T(3)	36.7	43.0	39.7	41.0	40.1	31.5	32.4	36.5	33.5	*	31.7	33.0	**

The values shown are the percentage of nucleotides in the complete coding sequences of each genome. Mean values for the mesophilic (Avg-meso) and hyperthermophilic (Avg-thermo; Avg-thermo+nequ) are shown. In addition, values of G+C and purines (G+A) are shown. Also, individual nucleotide values at each of the three codons positions are shown. The codon positions are shown in parentheses. Moreover significance based on a t-test are shown. ns ($p>0.05$); * ($p<0.05$); ** ($p<0.01$).

arginine (Table **4**). The contrast of codon usage in *N. equitans* was thus found more drastic from mesophiles, as compared to other hyperthermophiles. Such drastic deviations might be because of the additional adaptations in *N. equitans* due to parasitic mode of habits [5].

Relative synonymous codon usage (RSCU) was analyzed to measure the codon behavior in a group rather than detecting the overall increase or decrease in codons. The comparison of RSCU mean for each codon in mesophilic and hyperthermophilic genomes, without *N. equitans* identified 12 codons that show change in the absolute frequencies. Significant fall in RCSU mean was noted in hyperthermophiles for five codons when data for *N. equitans* was included (Table **5**). Overall there were significant increases in the frequency of eight codons (GAG, AGG, AAC, ATA, TCC, TAC, TTC and CTA) and significant decrease in nine codons (AAT, ATT, CTT, GAA, TAT, TTT, CCT, CGT and CGC). The results do not stand in agreement with Lobry and Necsulea [19], who suggested synonymous usage of codon for arginine as the most discrimintating factor between hyperthermophiles and non thermophiles. de Farias and Bonato [21] also reported a strong codon bias for arginine.

Nevertheless, predominance of purine-rich codons in hyperthermophiles can clearly be noticed (Tables **4** & **5**). Understandably, for eight amino acids (glutamic acid, arginine, asparagine, isoleucine, tyrosine, phenylalanine and leucine), the increase in frequency of one codon was at the cost of frequency of its alternative codon. The richness of transcribed strand in RY (or YR) nucleotides is thus a characteristic feature of hyperthermophilic genomes, and the degree of this richness might be correlated with the OGT of the organism [5].

Even though, patterns of codon usage are distinct between the two groups of organisms and could be under the influence of natural selection, yet these may simply be phylogenetic trends and may not reflect the thermophilic adaptations. Notably, on clustering the various species based on codon usage, Lobry and Necsulea [19] found archael hyperthermophiles and archael psychrophiles placed close to each other. While a common selection pressure may be determining codon usage in archael thermal extremophiles, the selection pressure linked to OGT is overruled to determine the codon usage and RCSU [19]. Thus, we suggest a thorough investigation on this aspect by considering more species in the analysis representing different archael families and occupying different habitats.

Table 3. Distribution of Dinucleotide Frequencies

Nucleotide	*Cjej*	*Bbur*	*Rpro*	*Llac*	Avg-Meso	*Mjan*	*Ssol*	*Stok*	Avg-Thermo		*Nequ*	Avg-thermo+*Nequ*	
AT	10.0	11.1	11.8	9.1	10.5	10.9	9.7	10.3	10.3	ns	11.3	10.6	ns
AG	6.7	6.4	6.5	5.7	6.3	8.5	8.5	8.2	8.4	***	7.7	8.2	***
AC	3.4	3.3	4.2	4.7	3.9	3.5	4.6	4.3	4.1	ns	4.0	4.1	ns
AA	16.1	16.7	14.2	13.5	15.1	15.3	11.6	12.8	13.2	ns	16.7	14.1	ns
TA	9.1	9.5	11.7	6.5	9.2	9.4	10.2	10.8	10.1	ns	11.5	10.5	ns
TG	6.5	6.0	5.8	7.5	6.4	6.8	5.6	5.7	6.0	ns	4.9	5.7	ns
TC	3.5	3.8	3.8	5.2	4.1	2.9	4.3	4.1	3.7	ns	2.6	3.5	ns
TT	13.9	14.1	11.5	11.6	12.8	10.8	9.1	10.2	10.0	*	10.1	10.1	*
CA	4.7	4.7	5.0	6.0	5.1	4.7	4.7	4.6	4.7	ns	4.6	4.6	ns
CT	4.9	4.6	4.8	5.3	4.9	3.8	5.2	5.3	4.8	ns	3.9	4.6	ns
CG	1.4	1.0	1.8	2.5	1.7	0.7	2.2	1.5	1.4	ns	1.4	1.4	ns
CC	1.7	1.8	1.7	2.6	2.0	2.0	2.8	2.4	2.4	ns	3.0	2.6	ns
GC	4.2	3.1	3.5	3.9	3.7	2.9	3.1	2.9	3.0	ns	3.3	3.0	ns
GT	4.1	3.6	4.7	4.7	4.3	4.3	5.2	4.9	4.8	ns	3.8	4.5	ns
GA	6.3	6.7	5.9	7.0	6.5	8.8	7.9	7.5	8.0	*	6.9	7.8	*
GG	3.4	3.6	3.1	4.2	3.6	4.8	5.4	4.5	4.9	*	4.3	4.7	*

The values shown are the percentage of dinucleotides in the complete coding sequences of each genome. Mean values for the mesophilic (Avg-meso) and hyperthermophilic (Avg-thermo; Avg-thermo+nequ) are shown. Also significance based on a t-test are shown. ns ($p>0.05$); * ($p<0.05$); *** ($p<0.001$).

AMINO ACID COMPOSITION IS RELATED WITH OGT

The average proportion of each amino acid in the mesophiles under study on one hand and in hyperthermophiles under study on the other hand was analyzed to complement codon usage data (Table **6**). As expected, changes were observed in the frequency of seven amino acids. The proteome analysis indicated that the frequency of four amino acids (Arginine, Proline, Valine, and Tyrosine) was markedly higher while that of three amino acids (Asparagine, Phenylalanine and Glutamine) was substantially lower. Earlier Klipcan *et al.* [22] have associated seven amino acids with thermophiles, the so-called class I amino acids. Among these seven amino acids, valine did not find mention, but Suhre and Claverie [23] recognized preference of valine in thermophilic proteomes. Similarly de Farias and Bonato [21] found an increase in Glutamate and Lysine corresponded with an equivalent fall in frequencies of glutamine and histidine and thus maintaining (E+K)/(Q+H) ratio. On the other hand, predominance of glutamate and valine in thermophilic proteomes was reported by Pasamontes and Garcia-Vallve [24].

The abundance of purine-rich codons is the possible reason for high frequency of arginine in thermophilic proteomes (Table **6**). The skewness in the frequencies of the amino acids in hyperthermophilic proteomes has been suggested to be related with the stability of the proteins under extremes of temperature [14, 23, 25, 26]. Increased occurrence of proline residues in loops are thought to enhance the thermostability of proteins [27, 28]. Similarly, valine is known to provide rigidity to the three dimensional structures of proteins causing smaller conformational entropy increase upon unfolding [29]. Higher frequencies of tyrosine despite being encoded by purine-poor codons (TAT and TAC) in hyperthermophilic proteomes, however is explained due to its property of providing thermostability to protein structures [30]. On the other hand, decrease in the asparagine and glutamine frequencies reduces the potential deamination of proteins and thus confers stability to thermophilic proteins [25].

Further the amino acid compositions of the helices in mesophilic and hyperthermophilic genomes were also found varied. It was observed that the amount of oppositely charged residues glutamate, lysine and arginine were higher while asparagine and glutamine were found under-represented in hyperthermophilic genomes (Fig. **1**). It has been suggested that increase in charged residues is responsible for increased number of salt bridges in hyperthermophiles and thereby provides the thermostability to protein [25]. It is well recognized that minute changes in local weak interactions can bring about thermostability in proteins [31], while the overall protein conformations may not see any changes. For example, Goldstein [32] found measures of thermostability

Table 4. Number of Codons Per Thousand

Codon	*Cjej*	*Bbur*	*Rpro*	*Llac*	Avg-Meso	*Mjan*	*Ssol*	*Stok*	Avg-Thermo		*Nequ*	Avg-thermo+*Nequ*	
GGG	5.8	7.8	5.6	7.8	6.8	10.4	9.7	7.3	9.1	ns (0.0790)	10.4	9.4	*
GAG	12.8	17.6	13.3	11.7	13.8	34.8	29.4	23.0	29.1	**	18.9	26.5	*
AGG	2.7	6.4	3.4	1.4	3.5	9.8	17.5	11.9	13.1	**	11.8	12.7	**
AGA	16.0	20.9	15.0	8.1	15.0	27.4	25.1	26.1	26.2	*	24.2	25.7	**
AAG	12.9	21.4	15.5	11.9	15.4	30.7	37.3	27.9	32.0	**	18.5	28.6	*
AAU	54.0	59.2	56.5	41.5	52.8	15.5	32.9	34.9	27.8	*	35.1	29.6	**
AUU	43.7	59.6	51.9	53.6	52.2	48.6	33.7	40.3	40.8	ns (0.0850)	30.4	38.2	*
CGU	6.4	1.8	9.5	15.0	8.2	0.3	1.7	1.4	1.1	ns (0.0860)	0.7	1.0	*
CGC	3.8	0.9	1.9	3.9	2.6	0.1	0.6	0.4	0.4	ns (0.0520)	0.6	0.4	*
CAA	28.4	18.7	24.6	31.1	25.7	9.0	15.5	15.6	13.4	*	20.5	15.2	*
CUA	6.8	8.7	11.6	7.4	8.6	8.5	19.2	16.4	14.7	ns (0.0940)	18.4	15.7	*
CUU	32.1	30.5	20.4	25.5	27.1	9.1	15.2	18.6	14.3	*	6.4	12.3	**
CCA	8.7	9.0	10.8	15.6	11.0	22.5	16.1	17.3	18.6	*	18.8	18.7	*

The values shown are number of codons within each genome. The numbers are scaled to a total of 1000 for each genome. Only those codons that show significant differences are listed. Also significance based on a t-test are shown. ns ($p>0.05$); * ($p<0.05$); ** ($p<0.01$).

Table 5. Relative Synonymous Codon Usage

Codon	*Cjej*	*Bbur*	*Rpro*	*Llac*	Avg-Meso	*Mjan*	*Ssol*	*Stok*	Avg-Thermo		*Nequ*	Avg-Thermo+*Nequ*	
GAG	0.36	0.52	0.46	0.34	0.42	0.81	0.87	0.66	0.78	**	0.49	0.52	*
GAA	1.64	1.48	1.54	1.66	1.58	1.19	1.13	1.34	1.22	**	1.51	1.49	*
AGG	0.53	1.2	0.61	0.23	0.64	1.55	2.25	1.74	1.85	**	1.84	1.07	**
AAU	1.72	1.63	1.7	1.6	1.66	1.41	1.33	1.43	1.39	**	1.33	1.50	***
AAC	0.28	0.37	0.3	0.4	0.34	0.59	0.67	0.57	0.61	**	0.67	0.50	***
AUA	0.91	1.12	1.26	0.33	0.91	1.30	1.58	1.51	1.46	ns (0.078)	1.91	1.11	*
AUU	1.52	1.67	1.43	2.1	1.68	1.39	1.07	1.22	1.23	ns (0.065)	0.87	1.51	*
UAU	1.73	1.59	1.73	1.58	1.66	1.55	1.30	1.48	1.44	*	1.56	1.59	*
UAC	0.27	0.41	0.27	0.42	0.34	0.45	0.70	0.52	0.56	*	0.44	0.41	*
UUU	1.86	1.81	1.71	1.58	1.74	1.59	1.19	1.39	1.39	*	1.38	1.57	*
UUC	0.14	0.19	0.29	0.42	0.26	0.41	0.81	0.61	0.61	*	0.62	0.43	*
UCC	0.18	0.27	0.23	0.26	0.24	0.37	0.68	0.43	0.49	*	0.67	0.38	*
CGU	1.29	0.33	1.69	2.52	1.46	0.04	0.21	0.20	0.15	ns (0.06)	0.11	1.03	*
CGC	0.76	0.16	0.34	0.66	0.48	0.01	0.07	0.06	0.05	*	0.09	0.31	*
CUA	0.37	0.5	0.69	0.45	0.50	0.54	1.11	0.96	0.87	ns (0.075)	1.06	0.64	*
CUU	1.78	1.76	1.21	1.55	1.58	0.58	0.88	1.09	0.85	*	0.37	1.02	**
CCU	2.34	1.77	2.01	1.45	1.89	1.02	1.30	1.46	1.26	ns (0.051)	1.25	1.40	*

The values shown are the relative frequencies of synonymous codon usage within each codon group. Only those codons that show significant differences are listed. Also significance based on a t-test are shown. ns ($p>0.05$); * ($p<0.05$); ** ($p<0.01$); *** ($p<0.001$).

Table 6. Amino Acids Per Thousand

Amino Acids	*Cjej*	*Bbur*	*Rpro*	*Llac*	Avg-Meso	*Mjan*	*Ssol*	*Stok*	Avg-Thermo		*Nequ*	Avg-thermo+*Nequ*	
Methionine	22.2	19.0	21.6	25.0	21.9	23.0	21.9	21.2	22.0	ns	16.9	21.7	ns
Alanine	68.1	45.4	60.6	74.0	62.0	55.2	56.3	55.9	55.8	ns	51.8	60.8	ns
Cysteine	12.2	6.6	10.9	4.4	8.5	12.7	6.0	6.3	8.3	ns	8.0	8.4	ns
Aspartic acid	52.9	52.0	48.4	53.0	51.6	55.2	47.0	46.4	49.5	ns	49.9	52.4	ns
Glutamic acid	70.4	68.2	57.8	69.7	66.5	86.2	68.0	70.1	74.7	ns	78.1	75.1	ns
Phenylalanine	60.3	62.4	48.8	47.5	54.7	42.6	44.6	45.5	44.2	ns	44.6	47.4	*
Glycine	56.1	52.6	54.3	66.3	57.3	64.2	64.5	63.2	64.0	ns	53.2	60.3	ns
Histidine	16.5	12.4	19.1	18.0	16.5	14.4	12.9	13.0	13.4	ns	13.5	15.6	ns
Isoleucine	86.7	107.6	108.8	77.0	95.0	105.1	94.5	99.4	99.7	ns	105.0	95.5	ns
Lysine	94.6	102.3	83.2	73.3	88.4	103.0	77.0	79.6	86.5	ns	107.5	93.0	ns
Leucine	108.2	103.5	101.1	98.8	102.9	94.2	103.7	103.0	100.3	ns	104.3	100.0	ns
Asparagine	63.0	72.9	66.5	52.1	63.6	52.7	49.7	48.9	50.5	ns	53.1	55.4	*
Proline	26.8	25.3	31.6	32.6	29.1	33.8	38.2	39.4	37.1	*	40.3	33.9	**
Glutamine	31.3	22.8	31.5	37.0	30.7	14.3	20.9	20.8	18.7	*	22.0	26.0	*
Arginine	29.9	32.1	33.7	36.0	32.9	38.2	46.8	41.1	42.0	*	38.5	36.4	*
Serine	64.5	74.5	67.5	66.5	68.3	45.2	67.1	66.9	59.7	ns	46.7	56.6	ns
Threonine	40.4	39.5	52.2	57.5	47.4	40.8	47.3	47.9	45.3	ns	41.2	46.7	ns
Valine	52.4	53.6	55.9	66.0	57.0	68.4	74.7	72.1	71.7	*	59.0	62.6	*
Tryptophan	6.5	5.0	7.2	10.0	7.2	7.2	10.6	10.2	9.3	ns	9.7	8.5	ns
Tyrosine	36.9	42.3	39.0	35.5	38.4	43.7	48.4	49.1	47.1	*	56.8	43.6	*

The values shown are the numbers of amino acids within each proteome, scaled to a total of 1000, in order to offset the effects of variations in proteome size. The significance based on a t-test are shown. ns ($p>0.05$); * ($p<0.05$); ** ($p<0.01$).

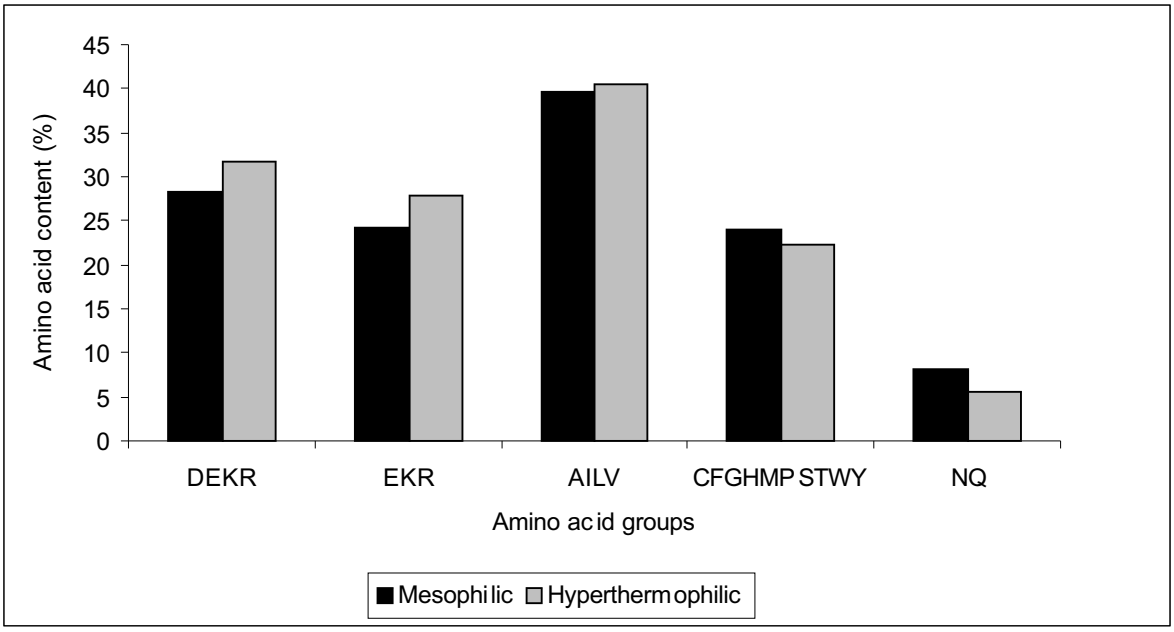

Fig. (1). Amino acid content in helices of mesophilic and hyperthermophilic genomes.

being increased number of charged residues. Das *et al*. [5] reported these residues being positively charged and found a positive correlation between the OGT and P/N ratio of amino acids in proteome. Further, salt bridges were reported to be a characteristic feature of mesophilic and psychrophilic protein folds [32]. An observation that falls consistent with insignificant presence of cysteine residues in thermophiles. Thus, simple amino acid substitutions can shift the balance towards thermophilic adaptations. Klipcan *et al.* [22] suggested the thermophilic adaptation of proteins is a sequence based phenomenon in place of structure based phenomenon. The suggested (E+K)/(Q+H) ratio [26] was calculated, which can be used as an indicator for discriminating organisms according to their OGT. The average ratio for hyperthermophilic genomes, ought to be higher than 4.5 [26], was found 4.7 and 5.1 respectively when calculated without and with *N. equitans*. The reason for exhibiting higher ratio is the higher abundance of purine tracts in hyperthermophiles compared with mesophiles because the glutamic acid and lysine are encoded only by pure-purinic codons.

Another discriminating factor between mesophilic and hyperthermophilic genomes is the absolute difference between the frequency of charged and polar amino acid residues, CvP-bias [23, 26]. The variations in the use of charged and polar residues have been related to large differences in surface accessibilities of the proteins [23, 26, 33] and therefore the CvP bias is further analyzed (Fig. **2**). It was observed that *N. equitans* exhibited a strong bias for the use of charged residues (Asp, Glu, Lys, Arg) at the expense of polar residues (Asn, Gln, Ser, Thr).

The genomic and proteomic composition of *N. equitans* is thus biased, like other hyperthermophilic organisms and causes an increase in charged residues on the molecular surface of the proteins that allows more ion pairs to be formed and thus enhancing protein stability at temperature extremes [25].

DIPEPTIDE COMPOSITION

The trends of occurrence of single amino acid are followed at dipeptide level also (Table **7**). For example, marked increase in tyrosine (Y) content demonstrated its effect as all the 14 dipeptides that exhibited significant difference had an increased frequency of tyrosine, even when it occurred with amino acid that show significant decrease (YN and NY). Similarly, increase in arginine, valine and glutamic acid produced the same effect. On the other hand, the significant decrease in glutamine leads to decrease in content of 14 dipeptides in hyperthermophiles even when it occurred with amino acids that show increase (K, E, I). Amino acids that show increased occurrence in hyperthermophiles frequently occurred in tandem with lysine, which in itself did not exhibit significant bias between mesophiles and hyperthermophiles. Thus there are certain dipeptides which significantly differ in their frequency between mesophiles and hyperthermophiles including *N. equitans* and thus influencing the thermostability of the protein. These trends support the hypothesis put forward by Klipcan *et al*. [22] that thermophilicity has been achieved at the level of sequence without bringing about any significant changes in the conformation of proteins. Conformational change in protein structures is obviously undesired as this would affect the nature of vital metabolic reactions a great deal. While preferential occurrence of amino acid residues adjacent to each other obviously affect intramolecular interactions and thus are instrumental in adjustment of proteins to the growth temperature of the organism [31].

CONCLUSION

The present study examined and analyzed the contributions of nucleotide, amino acid and synonymous codon usage pattern on the genomes of four GC-poor hyperthermophilic archaeal species. Nucleotide composition indicated that the influence of dinucleotide composition on protein

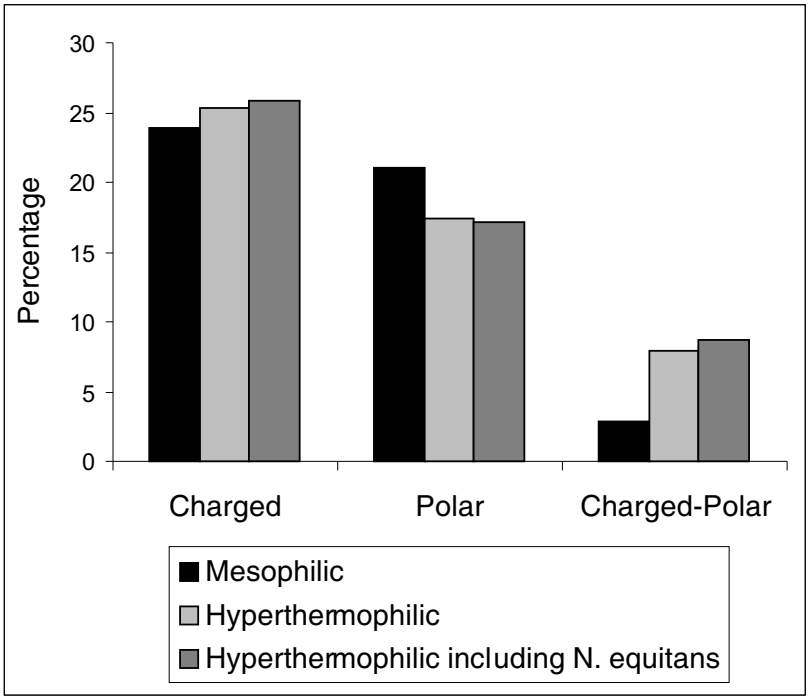

Fig. (2). Plot of the sum of percentages of charged, polar amino acids and the difference between the two categories in mesophilic and hyperthermophilic genomes.

Table 7. Dipeptides that Exhibit Significant Differences Between Mesophilic and Thermophilic Genomes

	M	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	W	Y
M					* +							* -		* -	* +					
A												* -								
C																				
D														** -			* -	* +		
E					* +		** +						* +	* -	** +			* +		* +
F		* -		* -										** -						
G					* +				* +	*** +					* +					*** +
H						** -										** -				
I													** +	** -	** +			** +		
K							** +						* +		*** +			*** +	* +	
L																				* +
N				*** -		* -		* -						** -			* -			
P							* +						** +		* +					*** +
Q	* -			** -	** -	* -			** -	** -		** -				** -				
R				** +	*** +		** +			*** +								*** +		
S				** -											* +					** +
T																				
V					* +				* +	**** +					** +					** +
W					** +				* +	** +										
Y				** +			*** +					*** +						*** +	* +	* +

+ indicates an increase in content of particular dipeptide in the direction mesophilic to hyperthermophilic genomes; - indicates an decrease in content of particular dipeptide in the direction mesophilic to hyperthermophilic genomes; The significance based on a t-test are shown. * ($p<0.05$); ** ($p<0.01$), *** ($p<0.001$) and ** **($p<0.0001$).

thermostability is larger than influence of mononucleotide composition. Codon usage analysis pointed towards the compositional constraint acting on the genome. Further, minor amino acid substitutions seemingly are sufficient for thermo-adaptability in place of drastic structural or conformational changes, and thus also maintain the intrinsic nature of various metabolic reactions. Together, these minor adjustments in genomic and proteomic contents might be considered as the means that have guided the survival of hyperthermophiles under drastic environments.

REFERENCES

[1] R. M. Atlas, and R. Bartha, "Microbial ecology-fundamentals and applications", Pearson Education (Singapore) Pte. Ltd, pp. 305-311, 2005.

[2] D. W. Grogan, "Hyperthermophiles and the problem of DNA instability", *Mol. Microbiol.*, vol. 28, pp. 1043-1049, 1998.

[3] R. J. Klein, Z. Misulovin, and S. R. Eddy, "Noncoding RNA genes identified in AT-rich hyperthermophiles", *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 7542-7547, 2002.

[4] N. Galtier, and J. R. Lobry, "Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperatures in prokaryotes", *J. Mol. Evol.*, vol. 44, pp. 632-636, 1997.

[5] S. Das, S. Paul, S. K. Bag, and C. Dutta, "Analysis of *Nanoarcheum equitans* genome and proteome composition: indications for hyperthermophilic and parasitic adaptations", *BMC Genomics*, vol. 7, pp. 186, 2006.

[6] D. P. Kreil, and C. A. Ouzounis, "Identification of thermophilic species by the amino acid compositions deduced from their genomes", *Nucleic Acids Res*, vol. 29, pp. 1608-1615, 2001.

[7] R. Schwartz, C. S. Ting, and J. King, "Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life", *Genome Res*, vol. 11, pp. 703-709, 2001.

[8] J. R. Lobry, and D. Chessel, "Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria", *J. Appl. Genet.*, vol. 44, pp. 235-261, 2003.

[9] R. Friedman, J. W. Drake, and A. L. Hughes, "Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles", *Genetics*, vol. 167, pp. 1507-1512, 2004.

[10] K. U. Foerstner, C. von Mering, S. D. Hooper, and P. Bork, "Environments shape the nucleotide composition of genomes", *EMBO Rep.*, vol. 6, pp. 1208-1213, 2005.

[11] J. Garnier, J. F. Gibrat, and B. Robson, "GOR method for predicting protein secondary structure from amino acid sequence", *Methods Enzymol.*, vol. 266, pp. 540-553, 1996.

[12] T. Kawashima, N. Amano, H. Koike, *et al*, "Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermo-*

plasma volcanium", *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 14257-14262, 2000.

[13] A. Paz, D. Mester, I. Baca, E. Nevo, and A. Korol "Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes" *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 2951-2956, 2004.

[14] G. A. Singer, and D. A. Hickey, "Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content", *Gene*, vol. 317, pp. 39-47, 2003.

[15] E. Waters, M. J. Hohn, I. Ahel, *et al*, "The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism", *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 12984-12988, 2003.

[16] H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter, "A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont", *Nature*, vol. 417, pp. 63-67, 2002.

[17] L. D. Hurst, and A. R. Merchant, "High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes", *Proc. Biol. Sci.*, vol. 268, pp. 493-497, 2001.

[18] R. J. Lambros, J. R. Mortimer, and D. R. Forsdyke, "Optimum growth temperature and the base composition of open reading frames in prokaryotes", *Extremophiles*, vol. 7, pp. 443-450, 2003.

[19] J. R. Lobry, and A. Necsulea, "Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes", *Gene*, vol. 385, pp. 128-136, 2006.

[20] S. Basak, and T. C. Ghosh, "On the origin of genomic adaptation at high temperature for prokaryotic organisms", *Biochem. Biophys. Res. Commun*., vol. 330, pp. 629-632, 2005.

[21] S. T. de Farias, and M. C. Bonato, "Prefered codons and amino acid couples in hyperthermophiles", *Genome Biol.*, vol. 3, preprint0006, 2000.

[22] L. Klipcan, I. Safro, B. Temkin, and M. Safro, "Optimal growth temperature of prokaryotes correlates with class II amino acid composition", *FEBS Lett.*, Vol. 580, pp. 1672-1676, 2006.

[23] K. Suhre, and J. M. Claverie, "Genomic correlates of hyperthermophilicity: an update", *J. Biol. Chem.*, vol. 278, pp. 17198-17202, 2003.

[24] A. Pasamontes, and S. Garcia-Vallve, "Use of a multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes", *BMC Bioinformatics*, vol. 7, pp. 257, 2006.

[25] R. Das, and M. Gerstein, "The stability of thermophilic proteins: a study based on comprehensive genome comparison", *Funct. Integr. Genomics*, vol. 1, pp. 76-88, 2000.

[26] C. Cambillau, and J. M. Claverie, "Structural and genomic correlates of hyperthermostability", *J. Biol. Chem.*, vol. 275, pp. 32383-32386, 2000.

[27] K. Watanabe, Y. Hata, H. Kizaki, Y. Katsube, and Y. Suzuki, "The refined crystal structure of *Bacillus cereus* oligo-1,6-glucosidase at 2.0 A resolution: structural characterization of proline-substitution sites for protein thermostabilization", *J. Mol. Biol.*, vol. 269, pp. 142-153, 1997.

[28] O. Bogin, M. Peretz, Y. Hacham, *et al*, "Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase", *Protein Sci.*, vol. 7, pp. 1156-1163, 1998.

[29] K. H. Lee, D. Xie, E. Freire, and L. M. Amzel, "Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation", *Proteins*, vol. 20, pp. 68-84, 1994.

[30] M. M. Gromiha, "Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins", *Biophys. Chem.*, vol. 91, pp. 71-77, 2001.

[31] I. Berezovsky, and E. Schakhnovich, "Physics and evolution of thermophilic adaptation" *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 12742-12747, 2005.

[32] R. A. Goldstein, "Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: Insights from the quasi-chemical approximation", *Protein Sci.*, vol. 16, pp. 1887-1895, 2007.

[33] S. Chakravarty, and R. Varadarajan, "Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study", *Biochemistry*, vol. 41, pp. 8152-8161, 2002.

Received: February 04, 2008 Revised: March 04, 2008 Accepted: March 04, 2008