Comparative study of synonymous codon usage in bacteria growing at extreme temperatures

Monisha Singhal and Pragya Chaturvedi

Department of Biotechnology,
Birla Institute of Scientific Research,
Jaipur, 302001, India
Email: monishasinghal50@gmail.com
Email: mbi_pragya@yahoo.com

R.K. Gothwal, M.K. Mohan and P.S. Solanki*

Department of Biotechnology,
Birla Institute of Scientific Research,
Jaipur, 302001, India

and

Department of Bioengineering,
Birla Institute of Technology,
Mesra, Jaipur Campus, 302017, Rajasthan, India
Email: rkgothwal@bisr.res.in
Email: kmohan@bisr.res.in
Email: pooransingh@bisr.res.in
*Corresponding author

Abstract: Genome availability has made possible to compare the codon and amino acid usage strategies among different extremophiles. Correspondence analysis was used to characterise various patterns present in 200 genes encoding 10 key enzymes of citric acid cycle from 20 organisms surviving at varying temperature. The study has shown that the different extremophiles follow a specific trend of codon usage and amino acid composition which is affected by temperature variation and base composition which is vital for functional and structural stability of enzymes and hence for their adaptive survival in such harsh environmental conditions. It was found that higher temperature favours high aromaticity score which can be linked to its thermal behaviour. The results and statistical analysis of various parameters of codon usage shows a level of preference in synonymous codons and indicates towards a kind of anonymous selection pressure which help stabilising the genetic material at varying degree of temperature.

Keywords: extremophiles; codon adaptation index; correspondence analysis.

Reference to this paper should be made as follows: Singhal, M., Chaturvedi, P., Gothwal, R.K., Mohan, M.K. and Solanki, P.S. (2021) 'Comparative study of synonymous codon usage in bacteria growing at extreme temperatures', *Int. J. Bioinformatics Research and Applications*, Vol. 17, No. 1, pp.53–68.

Biographical notes: Monisha Singhal is a postgraduate in Biotechnology and pursuing her PhD with keen interest in computational biology and bioinformatics research.

Pragya Chaturvedi is a postgraduate in Bioinformatics and pursuing her doctoral studies with keen interest in drug designing and systems biology.

R.K. Gothwal is working as Scientist at BISR and Faculty at BIT, Mesra, Jaipur Campus. He has obtained his Masters degree in Agricultural Biochemistry from Rajasthan Agricultural University, Bikaner, Rajasthan and PhD degree in Biotechnology from BIT, Mesra. Since 1997, he has worked over various projects including utilisation of lignin from lignocellulosic biomass for making phenol formaldehyde based wood adhesives and on the enumeration of microbial biodiversity of saline lakes and desert regions of Rajasthan. Presently, he is working in rhizosphere microbial community analysis of some economically important plants of desert regions of Rajasthan using 16S-rDNA technology based analysis.

M.K. Mohan has obtained his Masters and PhD from G.B. Pant University, Pantnagar. He started his career from TERI and has been associated with Microbial group in various capacities for over a decade. He has been a member of an International group that worked on innovative bioprocesses for energy generation from solid wastes. Since 1997, he has been with the biotechnology group at BISR. Continuing the environmental mitigation work he handled a project on decolourisation of textile dye effluents. Presently he is involved in assessing Microbial Biodiversity of salt lakes and deserts following the principles of polyphasic taxonomy.

P.S. Solanki is working as Bioinformatics Scientist at Birla Institute of Scientific Research and Faculty at Birla Institute of Technology, Mesra, Jaipur Campus, Jaipur. He obtained his Master degree in Biotechnology from MLS University, Udaipur followed by Advanced Diploma (PG) in Bioinformatics from JNU, New Delhi and PhD in Computational Biology and Bioinformatics from CCT, Rajasthan University, Jaipur. His research interest is in structural bioinformatics and evolutionary genomics. His current working area includes structure based adaptive evolution in extremophilic proteins, understanding interaction specificity in biological complexes, Understanding genomic evolution using base frequencies.

1 Introduction

From various studies it has been identified that codon composition plays very significant role in the structural and functional stability of enzymes. The contribution of synonymous codons were also studied in various organisms and it has been proved that they are not used randomly (Aota et al., 1988). It has been previously shown that in many prokaryotes and some lower eukaryotes, in addition to mutational biases, the natural selection was also influencing non-random codon usage. Against non-random usage of codons, it had been established that highly expressed genes used specific sets of codons (Cancilla et al., 1995; Freirepicos et al., 1994; Gharbia et al., 1995; Gouy and Gautier, 1982; Shields and Sharp, 1987; Stenico et al., 1994) and therefore develop biasness in genome evolution. The atypical codon usage pattern suggests that genes might have been attained by

horizontal transfer (Delorme et al., 1994; Groisman et al., 1992; Medigue et al., 1991). Characterising the pattern of codon usage in specific organism may be of importance due to its various practical applications, as it helps in molecular as well as evolutionary studies of that species (Sharp and Li, 1987a). It had become possible to examine, at what extent and how, the codon usage patterns diverged the related species (Lloyd and Sharp, 1992). Codon usage can also be shaped by natural selection acting through a specific external environmental force (Sharp and Li, 1987b). It has been shown in genomes of thermophilic bacteria, that synonymous codon usage is affected by two major factors, the overall G+C content of genome and growth at high temperature (Lynn et al., 2002). The variance in usage pattern of synonymous codon was different in mesophilic, thermophilic bacteria due to natural selection towards high temperature. In a study of internal correspondence analysis for understanding variability of codon usage, it was found that the trend for the amino-acid composition of thermophilic proteins was under the control of pressure at nucleic acid level and was not present in intergenic spaces (Lobry and Chessel, 2003).

The importance of translational constraint and its effect on protein was shown via multivariate analysis of 999 chromosome-encoded proteins from *Escherichia coli* (Lobry and Gautier, 1994). It was further investigated that there was strong relationships between amino-acid physico-chemical properties and the composition of proteins. In a co-inertia analysis it was found that hydrophobicity was linked to the biological environment of proteins, similarly the expressivity of protein genes were found to linked with the propensity of amino-acids to form alpha helix/beta sheets (Thioulouse and Lobry, 1995). Kimura (1983) reported in neutral theory of molecular evolution that observed neutral amino acid substitutions and polymorphisms do not indicate the fitness of the same (Kimura, 1983).

Using hierarchical clustering and PCA on amino acid composition it was shown that several factors influence the amino acid compositions. It was reported that thermophilic species can be identified by their global amino acid composition alone, although the dominant effect of GC content was not neglected (Kreil and Ouzounis, 2001).

The present study was performed using correspondence analysis approach in order to understand the comparative effect of temperature on codon bias pattern in 200 genes encoding the enzymes of citric acid cycle from 20 selected organisms surviving a varying degree of temperatures and organisms were categorised as psychrophiles, mesophiles, thermophiles and hyperthermophiles.

2 Materials and method

We analysed the patterns of synonymous codon usage in a total of 20 completely sequenced bacterial genomes (listed in Table 1). This set of genomes includes organisms growing at different ranges of optimal growth temperature in which 11 were eubacterial genomes and nine were archaeal genomes. Out of these 11 genomes, four eubacterial genomes were mesophilic, two were thermophilic and five were psychrophilic. The five hyperthermophilic genomes were archae; three were thermophilic archae genomes and one was mesophilic archae genome. Table 1 comprises the detailed listing of selected organisms along with their optimal growth temperature and G+C contents.

Table 1 List of organisms, their optimal growth temperature and average G+C content of genes encoding the 10 key citric acid cycle enzymes

Organism	*Taxonomy*	*Optimal growth temperature* (°C)	*G+C content* (%)
Hyperthermophiles			
Archaeoglobus fulgidus	Archae	83	51.32
Sulfolobus solfataricus	Archae	80	37.67
Aeropyrum pernix	Archae	95	58.43
Sulfolobus tokadaii	Archae	80	34.98
Pyrobaculum aerophilium	Archae	103	51.39
Thermophiles			
Thermoplasma volcanium	Archae	60	43.62
Picrophilus torridus	Archae	60	42.34
Thermoplasma acidophilum	Archae	59	48.98
Thermus thermophilus	Eubacteria	75	66.59
Thermobifida fusca	Eubacteria	50	66.49
Mesophiles			
Halobacterium sp.	Archae	37	67.13
Escherichia coli	Eubacteria	37	63.81
Yersinia pestis	Eubacteria	37	53.96
Deinococcus radiodurans	Eubacteria	30	45.67
Bacillus subtilis	Eubacteria	30	48.40
Psychrophiles			
Desulfotalea psychrophila	Eubacteria	10	51.42
Psychrobacter articum	Eubacteria	9	45.40
Photobacterium profundum	Eubacteria	10	55.56
Leptospira interrogans	Eubacteria	13	42.42
Colwellia psychrerythraea	Eubacteria	–1	40.7

2.1 Dataset

Genes encoding the 10 enzymes of Citric Acid Cycle (CAC) from the selected 20 organisms, were downloaded from the Kyoto Encyclopaedia of Genes and Genomes http://www.genome.jp/kegg in the *fasta* format (Table 2). Partial Coding sequences (CDS) and those having less than 300 bp were discarded. The total number of observed codons was 82,104.

The Mesophilic *Deinococcus radiodurans,* as a relative of thermophilic *Thermus thermophilus* was added to facilitate comparison. The Mesophilic archaeal species *Halobacterium sp*. was added to include an archaeal species to the majority of eubacteria in mesophiles and hence promote uniformity in dataset. This also helps distinguish between the effects of environmental selection and phylogenetic history.

Availability of completely sequenced genomes of psychrophilic bacteria enabled us to study five such species that had optimal growth at very low temperature.

Table 2 Represents dataset containing list of enzymes and genes from citric acid cycle

Name of enzyme of citric acid cycle	*No. of genes*
Citrate synthase	20
Isocitrate dehydrogenase	20
Aconitase	20
Succinyl CoA synthetase alpha subunit	20
Succinyl CoA synthetase beta subunit	20
Succinate dehydrogenase flavoprotein subunit	20
Succinate dehydrogenase iron sulphur subunit	20
Succinate dehydrogenase cytochrome subunit	20
Fumarate hydratase class II	20
Malate dehydrogenase type B	20
Total No. of genes	200

2.2 Parameters

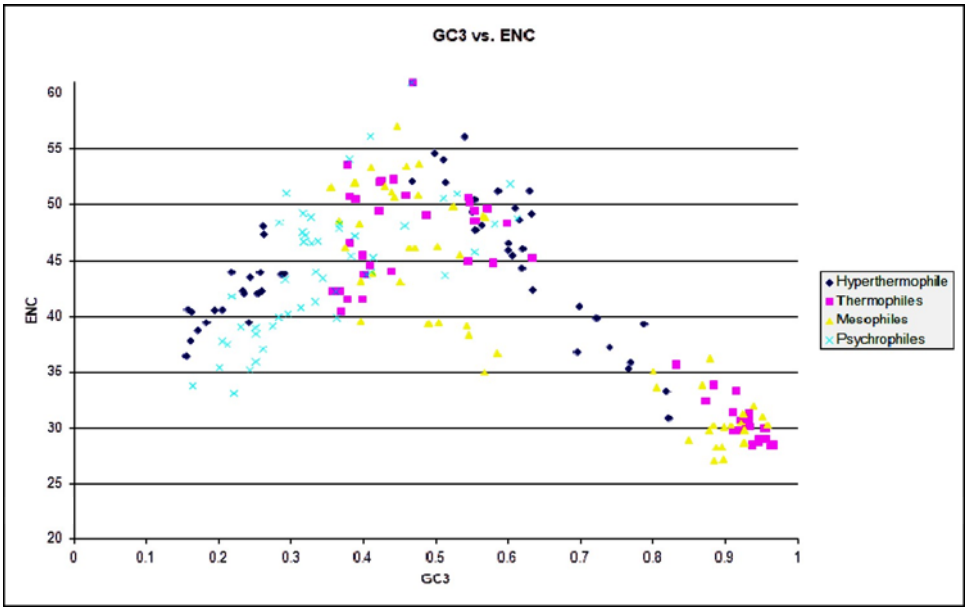
Different researchers have used different parameters for evaluating and analysing the codon usage pattern at genes and genome level. For the current study we have used the parameters and codon usage indices as suggested by Peden (1999). We have calculated the codon usage indices for each gene using the CodonW1.4.2 program by J.F. Peden (Benzecri, 1992). The indices calculated were:

- *CAI*: The Codon Adaptation Index is a most widespread technique to analyse codon usage bias. It is based on a reference set derived from highly expressed genes. It predicts relative importance of that codon and provided with a score. It helps to predict the selection of codon usage in evolution. It is useful for predicting the level of expression of a gene and for making comparisons of codon usage in different organisms (Aota et al., 1988).
- *ENC*: It is the Effective Number of Codons used in a gene which can be calculated from the codon usage data which does not rely upon of gene length and amino acid composition. Nc ranges from 20 to 61. Codon usage patterns can be investigated by the plot of Nc vs. G+C content at synonymous sites (Greenacre, 1984).
- *GC3*: This is the fraction of codons that are synonymous at the third codon position, which have either a guanine or cytosine at that third codon position.
- *Aromaticity score*: The frequency of aromatic amino acids (PHE, TYR, TRP) in the hypothetical translated gene product (Gene, 1990).
- *Hydropathicity*: It is used to calculate the general average hydropathicity scores of hypothetical gene product. It is calculated as the arithmetic mean of the sum of the hydropathic indices of each amino acid (Gene, 1990).

3 Results and discussion

A total of 200 genes encoding the 10 key enzymes of citric acid cycle from all 20 organisms were combined for the analysis. Graphs were plotted between the GC composition at the 3rd base and ENC (Figure 1) for all the four datasets of Hyperthermophiles, Thermophiles, Mesophiles and Psychrophiles.

Figure 1 Graph between ENC and GC3. Hyperthermophiles are shown in dark blue diamonds, thermophiles are shown in pink squares, mesophiles are shown in yellow triangles and psychrophiles are shown in cyan crosses (see online version for colours)



The codon adaptation index values for all the 10 enzymes of 20 organisms are plotted. The reference genes used in the calculation of CAI values were set of highly expressed genes of *E. coli*. Hence codon usage similar to highly expressed genes of *E. coli* would give CAI values closer to 1 whereas those very different from those of *E. coli* would give CAI values closer to 0. Hence from the graph (Figure 2) we infer that Mesophiles were closest related to *E. coli* and have the most similar pattern of codon usage, since *E. coli* is also a mesophile. Next, psychrophiles were more related to mesophiles in codon usage pattern than the thermophiles or the Hyperthermophiles which were far distant.

The yellow curve depicting the mesophiles was most closely adapted to the codon usage pattern of *E. Coli* which was the reference for the calculation of Codon Adaptive Index. Next the psychrophiles shown by the cyan coloured curve, then the thermophiles (pink) and lastly the hyperthermophiles (blue) which were far distant from the codon usage pattern of *E. Coli.*

Figure 2 Graph showing comparison of CAI values of hyperthermophiles, thermophiles, mesophiles and psychrophiles (see online version for colours)

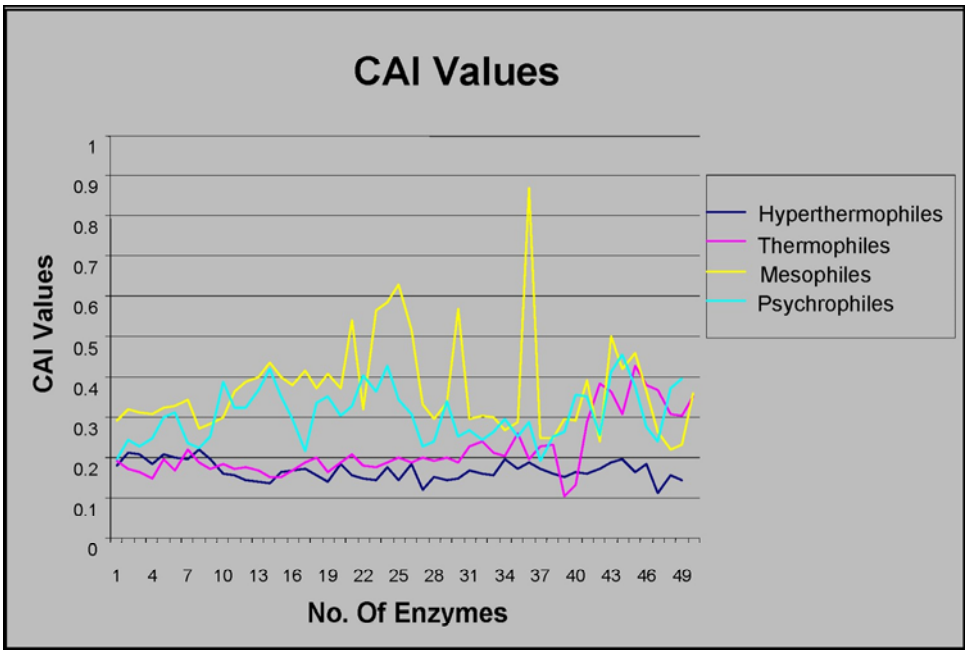
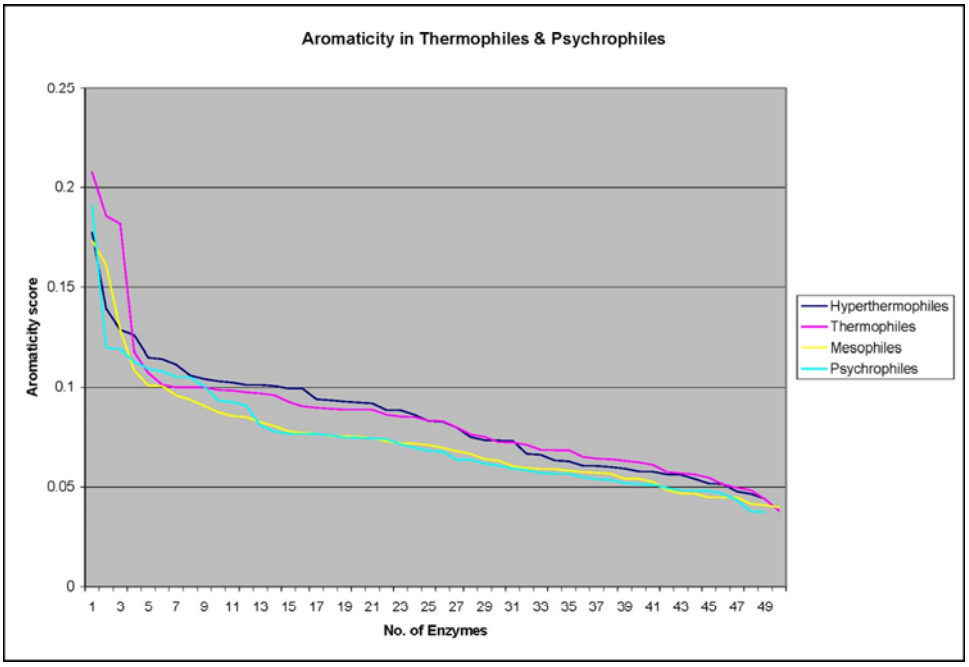


Figure 3 The Aromaticity score of hyperthermophiles, thermophiles, mesophiles and psychrophiles (see online version for colours)

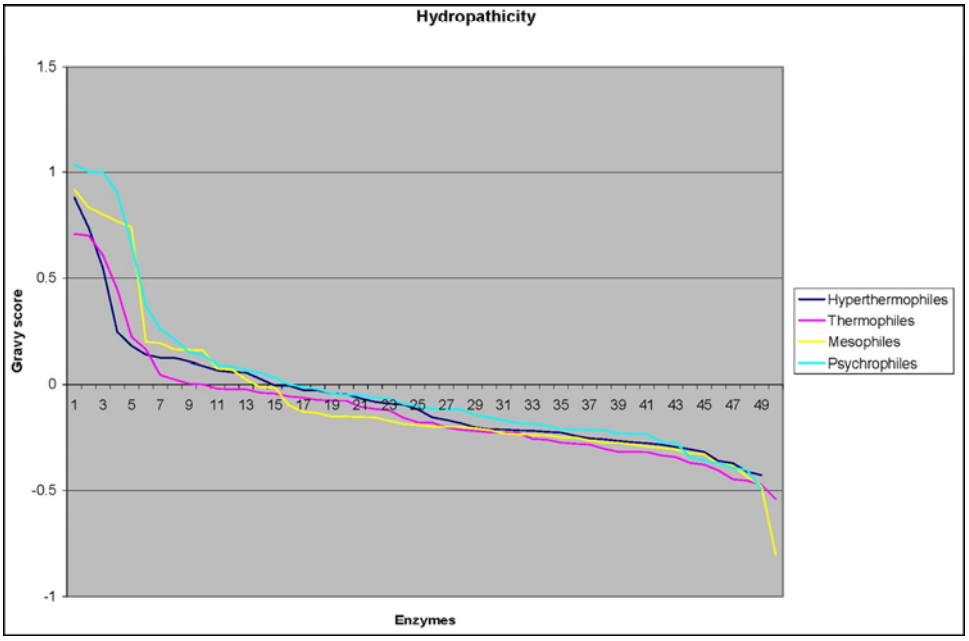


For all enzymes aromaticity purview from 0.05 to 0.20. The scores which were found between mesophiles and pyschrophiles were in overlapped conditions and this could be also observed between thermophiles and hyperthermophiles. The highest aromaticity level was found in thermophiles in collate with others. In this plot we observed that the higher temperature based organisms are lying above the lower temperature based organisms except some interactions in the beginning.

The Aromaticity score of various enzymes in all 20 organisms were calculated using the CodonW program. Figure 3 shows the Aromaticity score of the enzymes in all 20 bacteria.

It was important to study hydrophobic amino acids coded by synonymous codons, as we know it plays crucial role in protein function. In beginning graph shows hydrophobicity profile inverse pattern in comparison with aromaticity and later on shows an overlapping pattern throughout the score in the plot.

Figure 4 Graphical display of hydrophobicity of hyperthermophiles, thermophiles, mesophiles and psychrophiles (see online version for colours)

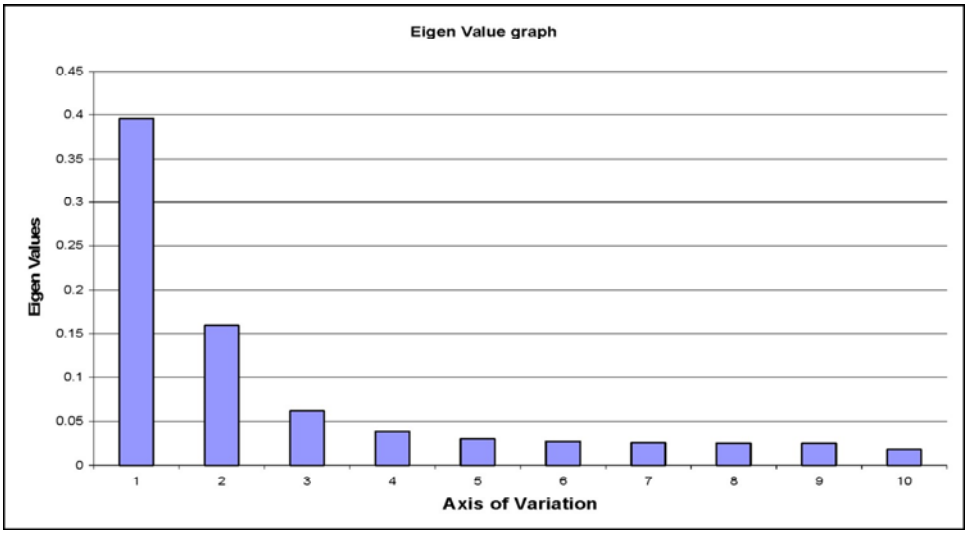


The Hydrophobicity of various enzymes in the 20 organisms selected was again calculated by the CodonW program and the graphical analysis depicted below. Among all the bacteria the psychrophilic bacteria tend to be the most hydrophobic in nature (Figure 4).

Correspondence analysis; the most commonly used multivariate analysis method for the analysis of codon and amino acid usage was performed using CodonW1.4.2. Correspondence Analysis is a multivariate method that projects data in high dimensional space onto low- dimensional sub spaces. The principal factors are therefore along the directions of maximum variability in the dataset. Hence we can find out the trend of maximum variability in the data.

All the 200 genes were used in the correspondence analysis of codon usage. The Eigen values graph obtained is shown in Figure 5.

Figure 5 The Eigen Value graph. There is 39.57 % variation of dataset associated with axis 1 and 15.94% variation associated with axis 2 (see online version for colours)



During the correspondence analysis, each gene was represented as a 59 dimensional vector, and each dimension corresponds to the codon usage value of one sense codon (excluding AUG, UGG and three stop codons).

The Eigen values graph (Figure 5) showed that there was a strong trend in the data with the two major factors accounting for 55.51% of total variability; hence the first factorial map is a good representation of the data. All the other factors were small enough to be ignored. The axis 1 (39.57 % variation) values were plotted on the horizontal axis and the second major trend of variation (15.94%) (Figure 6) was plotted on the vertical axis. All the genes show a broad distribution along the horizontal axis; however these groups are significantly different with respect to their positions along the vertical axis.

The first axis (39.575) was linked to G+C content of the genes, with poor genomic GC content organisms on the left and rich genomic GC content organisms on the right. The Psychrophiles have AT-rich genomes and hence clustered on the left of the horizontal axis (Figure 7). Similarly, GC rich organisms were clustered to the right (Figure 7). The second axis (15.94%) was linked to temperature, with all the thermophilic and hyperthermophilic organisms clustered on the top and the mesophilic and psychrophilic bacteria clustered at the bottom (Figure 8). Due to poor GC content and low optimal growth temperature the psychrophilic genes were clustered in the third quadrant of the first factorial map.

The graph was plotted between first axis (39.54) and second axis (15.94) of inertia for analysing the distribution pattern of GC and AT ending codons in all selected organisms (Figure 9). This plot clearly specifies the distribution of GC ending codon (in dark blue colour) cluster to the right and the AT ending codons (in pink colour) to the left. Thus from the plot it can be visualised that the first axis of inertia is correlated to the overall GC content of the gene.

Figure 6 Correspondence Analysis of codon usage using CodonW. The two major trends of variation were plotted, with the horizontal axis having a 39.54% variation and the vertical axis showing a 15.94% variation. All other axis did not show any major trend of variation (see online version for colours)

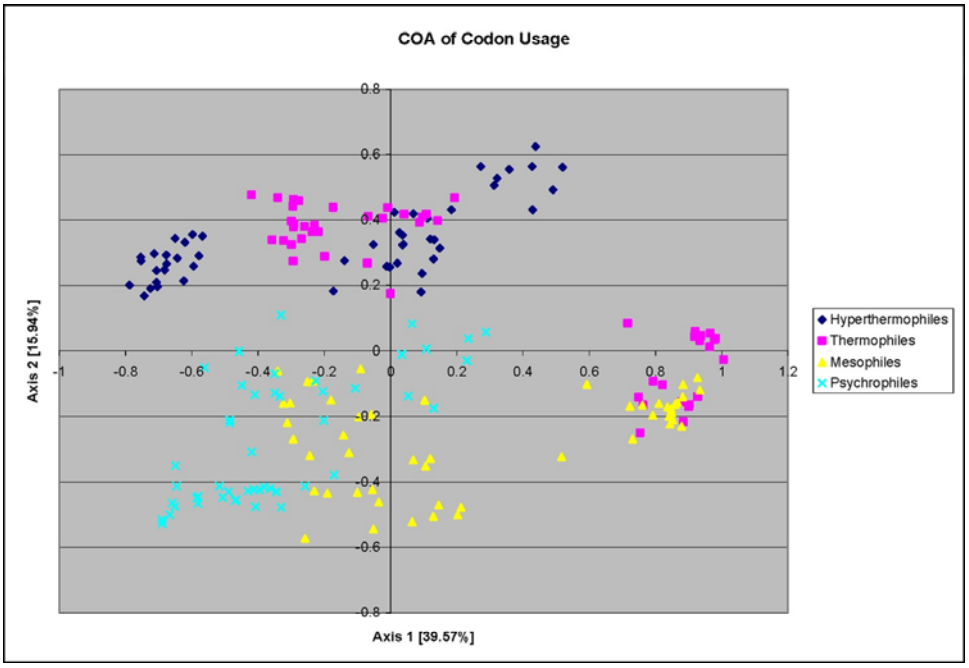


Figure 7 Graphs depicting correlation of axis 1 of correspondence analysis with GC content (see online version for colours)

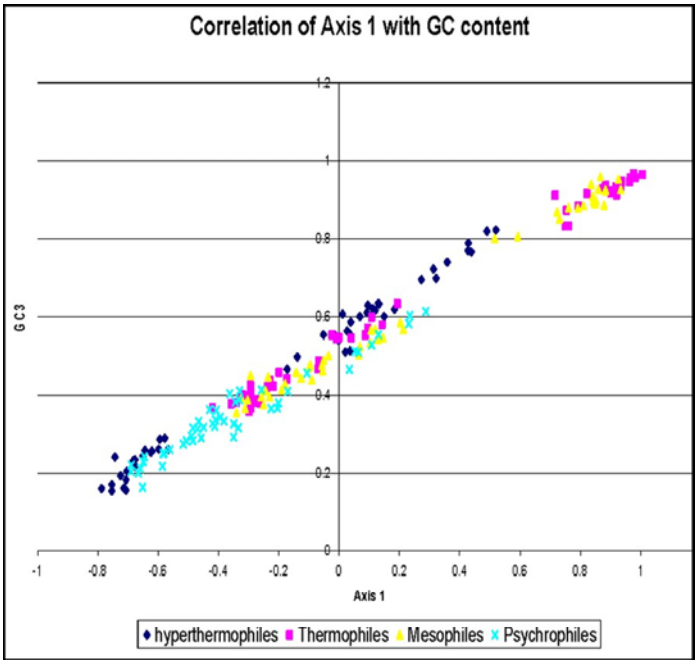


Figure 8 Graphs depicting correlation of axis 2 with temperature (see online version for colours)

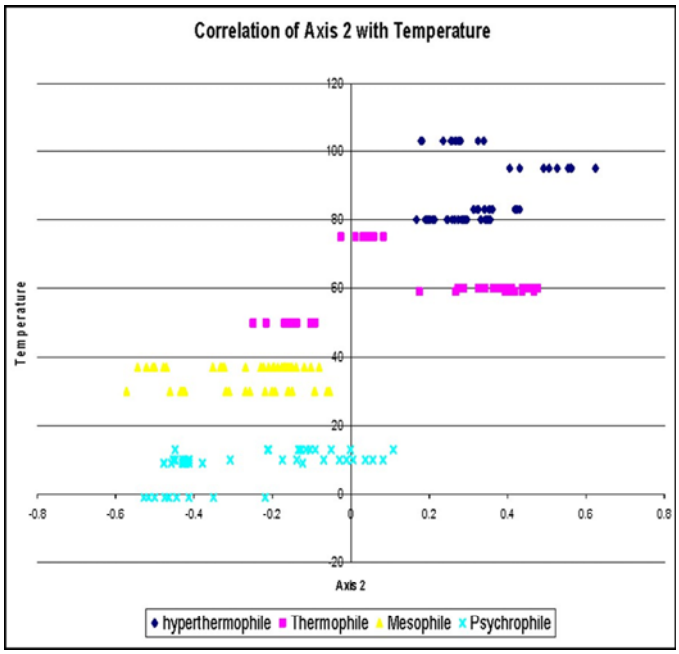
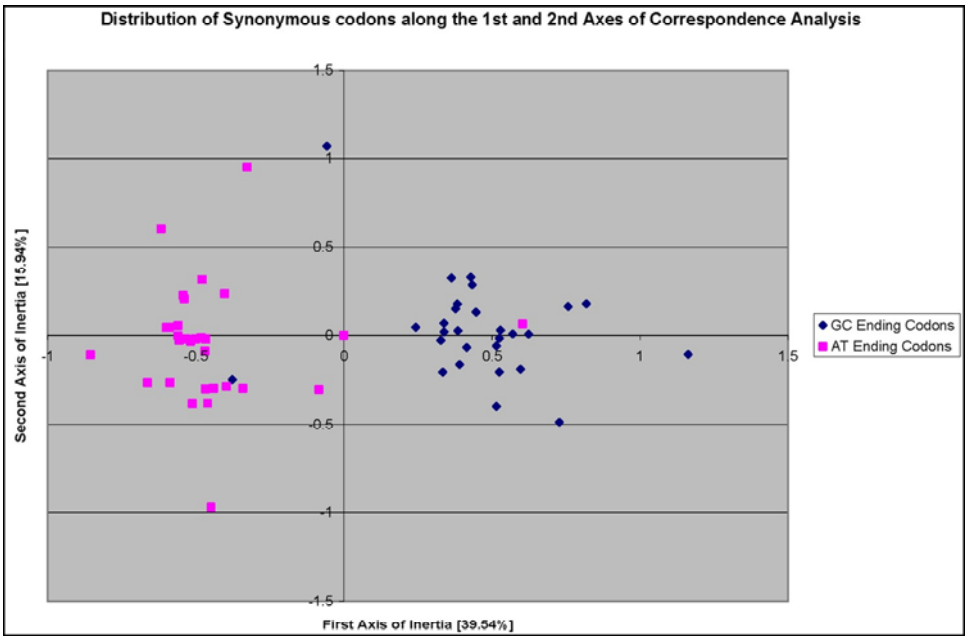


Figure 9 The distribution of synonymous codons along the first and second axes of the correspondence analysis. Clearly, the GC Ending codons (in dark blue) cluster to the right and the AT Ending codons (in pink) cluster to the left. Hence the first axis of inertia is correlated to the overall GC content of the gene (see online version for colours)



Correspondence analysis of amino acid usage was performed using CodonW1.4.2. Table 3 summarises the results. Codon coordinates on the first factorial map reveal that the AT-rich codons were on the left of the first axis and the GC-rich codons were on the right of the first axis, illustrating the correlation between the G+C content and codon usage strategy.

The second axis showed that hyperthermophiles and thermophiles tended to favour codons at the top of the first factorial map e.g., AGG, ATA, AGA and AAG (shown in red in Table 3) and to avoid using codons at the bottom (shown in dark blue in Table 3). The psychrophilic organisms were biased towards using codons shown by dark blue colour in Table 3 (CGT, CGC, ATC and CAA).

Table 3 Codon coordinates on the first factorial map for synonymous codon usage in the given dataset. The first axis is he genomic G+C content with GC-rich codons having positive values. The second factor is linked to thermophily, with positive values corresponding to codons favoured in thermophilic species (see online version for colours)

Amino acid	*Synonymous codons*	*Axis 1*	*Axis 2*
Ala	GCT	–0.51751	–0.03225
	GCC	0.60343	0.06522
	GCA	–0.48760	0.13902
	GCG	0.3336	–0.20574
Arg	CGT	–0.44827	**–0.97127**
	CGC	0.72666	**–0.48684**
	CGA	–0.08431	–0.30350
	CGG	1.161855	–0.10633
	AGA	–0.61693	**0.60044**
	AGG	–0.05820	**1.06617**
Asn	AAT	–0.55982	0.05564
	AAC	0.327397	–0.02748
Asp	GAT	–0.50027	–0.02105
	GAC	0.52923	0.028908
Cys	TGT	–0.51071	–0.38236
	TGC	0.446288	0.130502
Gln	CAA	**–0.66276**	–0.26503
	CAG	0.377479	0.149476
Glu	GAA	–0.34101	–0.29672
	GAG	0.362741	0.322636

Table 3 Codon coordinates on the first factorial map for synonymous codon usage in the given dataset. The first axis is he genomic G+C content with GC-rich codons having positive values. The second factor is linked to thermophily, with positive values corresponding to codons favoured in thermophilic species (see online version for colours) (continued)

Amino acid	*Synonymous codons*	*Axis 1*	*Axis 2*
Gly	GGT	–0.45995	–0.37921
	GGC	0.414528	–0.06734
	GGA	–0.44042	–0.29672
	GGG	0.428026	0.327027
His	CAT	–0.53205	–0.02051
	CAC	0.38493	0.025508
Ile	ATT	–0.39653	–0.28492
	ATC	0.515103	**–0.39733**
	ATA	–0.32618	**0.952772**
Leu	TTA	–0.85543	–0.1086
	TTG	–0.37594	–0.24777
	CTC	0.757301	0.161615
	CTA	–0.54289	0.223983
	CTG	0.596241	–0.18990
Lys	AAA	–0.46730	–0.29933
	AAG	0.433275	**0.283988**
Met	ATG	0	0
Phe	TTT	–0.46623	–0.01998
	TTC	0.338204	0.020056
Pro	CCT	–0.50202	–0.02051
	CCC	0.818075	0.17767
	CCA	–0.59918	0.044702
	CCG	0.390766	–0.16377
Ser	TCT	–0.58664	–0.26415
	TCC	0.62355	0.007345
	TCA	–0.55577	–0.02792
	TCG	0.515045	–0.05885
	AGT	–0.5388	0.205875
	AGC	0.38225	0.175702
Ter	TAA	0	0
	TAG	0	0
	TGA	0	0

Table 3 Codon coordinates on the first factorial map for synonymous codon usage in the given dataset. The first axis is he genomic G+C content with GC-rich codons having positive values. The second factor is linked to thermophily, with positive values corresponding to codons favoured in thermophilic species (see online version for colours) (continued)

Amino acid	*Synonymous codons*	*Axis 1*	*Axis 2*
Thr	ACT	–0.5593	–0.00171
	ACC	0.52465	–0.20511
	ACA	–0.47887	0.31434
	ACG	0.242038	0.04555
Trp	TGG	0	0
Tyr	TAT	–0.46921	–0.08747
	TAC	0.337757	0.068563
Val	GTT	–0.48197	–0.01226
	GTC	0.56950	0.008176
	GTA	–0.58888	0.046366
	GTG	0.524656	–0.01554

4 Conclusion

This study was directed towards understanding the strategies of codon usage patterns among different temperature based groups of organism. The genes encoding the key enzymes of citric acid cycle were selected as dataset. The correspondence analysis, a multivariate analysis method, was used to study the variation among the codon usage in different group of organism. In this study we have analysed the correlation between different factors on inter genomic variations in selected organisms. The patterns and biological basis of codon usage in thermophiles have previously been investigated in detail; however codon bias in psychrophilic bacteria has not been studied due to lack of completely sequenced genomes. It was found that higher temperature favours high aromaticity score which can be linked to its thermal behaviour.

From the reported results it can be inferred that organisms living at different degree of temperatures follow a specific trend of codon usage and amino acid composition which is affected by temperature variation and base composition, which is vital for functional and structural stability of enzymes and hence for their adaptive survival in such harsh environmental conditions.

It can be concluded from this study that nature favours some kind of selective pressure in preferring the codon based upon the survival conditions of temperature. It may be any environmental condition which is vital for survival of an organism. Hence the parameters of living conditions such as temperature, pressure, pH are the key factors in deciding the codon composition of genetic element of an organism.

Acknowledgement

I would like to thank BTIS Sub-DIC Center at Birla Institute of Scientific Research (BISR), Jaipur for providing the resources required for this research work.

References

Aota, S., Gojobori, T., Ishibashi, F., Maruyama, T. and Ikmura, T. (1988) 'Codon usage tabulated from the GenBank genetic sequence data', *Nucleic Acids Res.*, Vol. 16, pp.r315–r402.

Benzecri, J.P. (1992) *The Correspondence Analysis Hankbook; Statistics: Textbooks and Monographs*, Publish. Marcel Dekker.

Cancilla, M.R., Hillier, A.J. and Davidson, B.E. (1995) 'Lactococcus lactis glyceraldehyde-3-phosphate dehydrogenase gene, gap - further evidence for strongly biased codon usage in glycolytic pathway genes', *Microbiology-UK*, Vol. 141, pp.1027–1036.

Delorme, C., Godon, J.J., Ehrlich, S.D. and Renault, P. (1994) 'Mosaic structure of large regions of the Lactococcus lactis subsp cremoris chromosome', *Microbiology-UK*, Vol. 140, pp.3053–3060.

Freirepicos, M.A., Gonzalezsiso, M.I., Rodriguezbelmonte, E., Rodrigueztorres, A.M., Ramil, E. and Cerdan, M.E. (1994) 'Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes', *Gene*, Vol. 139, pp.43–49.

Gene, W.F. (1990) 'The 'effective number of codons' used in a gene', *Gene*, Vol. 87, No. 1, pp.23–29.

Gharbia, S.E., Williams, J.C., Andrews, D.M.A. and Shah, H.N. (1995) 'Genomic clusters and codon usage in relation to gene-expression in oral gram-negative anaerobes', *Anaerobe*, Vol. 1, pp.239–262.

Gouy, M. and Gautier, C. (1982) 'Codon usage in bacteria: correlation with gene expressivity', *Nucleic Acids Res.*, Vol. 10, pp.7055–7074.

Greenacre, M.J. (1984) *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Groisman, E.A., Saier, M.H.J. and Ouchman, H. (1992) 'Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the Salmonella', *EMBO Journal*, Vol. 11, pp.1309–1316.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press.

Kreil, D.P. and Ouzounis, C.A. (2001) 'Identification of thermophilic species by amino acid compositions deduced from their genomes', *Nucleic Acids Res.*, Vol. 29, pp.1608–1615.

Lloyd, A.T. and Sharp, P.M. (1992) 'Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae', *Nucleic Acids Research*, Vol. 20, No. 20, pp.5289–5295.

Lobry, J.R. and Chessel, D. (2003) 'Internal Correspondence analysis of codon and amino acid usage in thermophilic bacteria', *J. Appl. Genetics*, Vol. 44, No. 2, pp.235–261.

Lobry, J.R. and Gautier, C. (1994) 'Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 E. coli chromosome-encoded genes', *Nuc. Acids Res.*, Vol. 22, No. 15, 3174–3180.

Lynn, D.J., Singer, G.A.C. and Hickey, D.A. (2002) 'Synonymous codon usage is subject to selection in thermophilic bacteria', *Nucleic Acids Research*, Vol. 30, No. 19, pp.4272–4277.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) 'Evidence for horizontal gene transfer in Escherichia coli speciation', *Journal of Molecular Biology*, Vol. 222, pp.851–856.

Peden, J.F. (1999) *Analysis of Codon Usage*, PhD thesis, University of Nottingham, Nottingham, UK.

Sharp, P.M. and Li, W.H (1987b) 'The selection-mutation-drift theory of synonymous codon usage', *Nucleic Acids Research*, Vol. 15, pp.1281–1295.

Sharp, P.M. and Li, W.H. (1987a) 'The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Research*, Vol. 15, No. 3, pp.1281–1294.

Shields, D.C. and Sharp, P.M. (1987) 'Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases', *Nuc. Acids Res.*, Vol. 15, pp.8023–8040.

Stenico, M., Lloyd, A.T. and Sharp, P.M. (1994) 'Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases', *Nucleic Acids Res.*, Vol. 22, 2437–2446.

Thioulouse, J. and Lobry, J.R. (1995) 'Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package', *Comput. Appl. Biosci.*, Vol. 11, No. 3, June, pp.321–329.