BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(7)

Statistical analysis of pentose phosphate pathway genes from eubacteria and eukarya reveals translational selection as a major force in shaping codon usage pattern

Ayon Pal[1], Subhasis Mukhopadhyay[2] & Asim Kumar Bothra[3]*

[1]Department of Botany, Raiganj College (University College) P.O.- Raiganj, Dist.- Uttar Dinajpur, PIN-733134, West Bengal, India; [2]Bioinformatics Centre, Department of Biophysics, Molecular Biology and Bioinformatics University of Calcutta, 92 APC Road, Kolkata-700009, West Bengal, India; [3]Cheminformatics Bioinformatics Lab, Department of Chemistry, Raiganj College (University College) P.O.- Raiganj, Dist.- Uttar Dinajpur, PIN-733134, West Bengal, India; Asim Kumar Bothra – Email: asimbothra@gmail.com; *Corresponding author

Received March 26, 2013; Accepted March 27, 2013; Published April 13, 2013

Abstract:
Comparative analysis of metabolic pathways among widely diverse species provides an excellent opportunity to extract information about the functional relation of organisms and pentose phosphate pathway exemplifies one such pathway. A comparative codon usage analysis of the pentose phosphate pathway genes of a diverse group of organisms representing different niches and the related factors affecting codon usage with special reference to the major forces influencing codon usage patterns was carried out. It was observed that organism specific codon usage bias percolates into vital metabolic pathway genes irrespective of their near universality. A clear distinction in the codon usage pattern of gram positive and gram negative bacteria, which is a major classification criterion for bacteria, in terms of pentose phosphate pathway was an important observation of this study. The codon utilization scheme in all the organisms indicates the presence of translation selection as a major force in shaping codon usage. Another key observation was the segregation of the *H. sapiens* genes as a separate cluster by correspondence analysis, which is primarily attributed to the different codon usage pattern in this genus along with its longer gene lengths. We have also analyzed the amino acid distribution comparison of transketolase protein primary structures among all the organisms and found that there is a certain degree of predictability in the composition profile except in *A. fumigatus* and *H. sapiens*, where few exceptions are prominent. In *A. fumigatus*, a human pathogen responsible for invasive aspergillosis, a significantly different codon usage pattern, which finally translated into its amino acid composition model portraying a unique profile in a key pentose phosphate pathway enzyme transketolase was observed.

Keywords: Metabolic pathway, codon usage, pentose phosphate pathway, Nc, CAI, transketolase.

Background:
An outstanding facet of metabolism is the congruence of the key metabolic pathways amongst diverse species. It might have taken place as a result of their early advent in evolutionary history as well as the process possesses high efficiency. One thus looks for an opportunity for carrying out comparative analysis of metabolic pathways among widely diverse species with an aim to extract information about the functional relation of organisms **[1, 2].** The pentose phosphate pathway or hexose monophosphate shunt exemplifies one such important metabolic pathway which meets the need of all organisms for providing reducing power in the form of NADPH to execute anabolism. Apart from generating NADPH for use in reductive biosynthesis reactions, the pentose phosphate pathway generates pentose sugars which is utilized by actively growing cells to put together nucleic acids, aromatic amino acids, cell

wall constituents, vitamins and coenzymes like NADH, ATP and others **[3-7].** Biochemically, the pentose phosphate pathway has two distinct phases – the irreversible oxidative phase which generates pentose phosphates and NADPH, and the reversible non-oxidative phase that recycles pentose phosphates to glucose 6-phosphate. In enteric bacteria such as *Escherichia coli* pentose phosphate pathway is the sole route for utilizing sugars like D-xylose, D-ribose and L-arabinose **[8, 9]** whereas, in many ascomycetous fungi, mainly yeasts, pentose phosphate pathway defends the cell from oxidative stress. From the medical perspective, pentose phosphate pathway is of immense significance and mutations in different genes connected with this pathway results in metabolic disorders which includes at least three congenital deficiencies **[10, 11].**

There are overall ten enzymatic reactions in the bi-phasic pentose phosphate pathway with most of the consequent enzymes being subdued by the final products of the reaction **[12]**. The reactions of the oxidative phase is concerned with the oxidative NADPH generation and formation of the five carbon sugar ribose 5-phosphate. These reactions are catalyzed by the enzymes glucose 6-phosphate dehydrogenase, lactonase and 6-phosphogluconate dehydrogenase. On the other hand, in the non-oxidative phase, the enzymes phosphopentose isomerase, phosphopentose epimerase, transketolase and transaldolase catalyzes the interconversion of three to seven carbon containing monosaccharides in a chain of non-oxidative reactions. This may result in the synthesis of five carbon containing sugars for nucleotide biosynthesis. Further, conversion of these excess five carbon sugars into intermediates for glycolytic pathway may occur. The enzymes transketolase and transaldolase are vital enzymes in this regard as they create a reversible link between glycolysis and pentose phosphate pathway.

In view of the immense significance of the pentose phosphate pathway and the multitude of functions this pathway performs, we have undertaken an in-depth bioinformatic analysis of the pentose phosphate pathway in 10 different organisms representing diverse metabolic niche as well as habits and include prokaryotes as well as eukaryotes. In this study, we have attempted a comprehensive and comparative codon usage analysis of the pentose phosphate pathway genes of a diverse group of organisms representing different niches and the related factors affecting codon usage with special reference to the major forces influencing codon usage patterns. Correspondence analyses of codon usage and amino acid usage was performed to investigate the major trends in codon and amino acid variations existing in the pentose phosphate pathway genes. We have also tried to correlate the codon usage bias with the tRNA content and analyze codon adaptation index to predict the potential expression level of some of the genes related to the pentose phosphate pathway. Along with this we have also undertaken an elaborate amino acid profiling of a key pentose phosphate pathway enzyme transketolase of the ten different organisms featured in our study to find out if there is any significant difference in the amino acid usage pattern among diverse groups that might have cropped up as a result of adaptation.

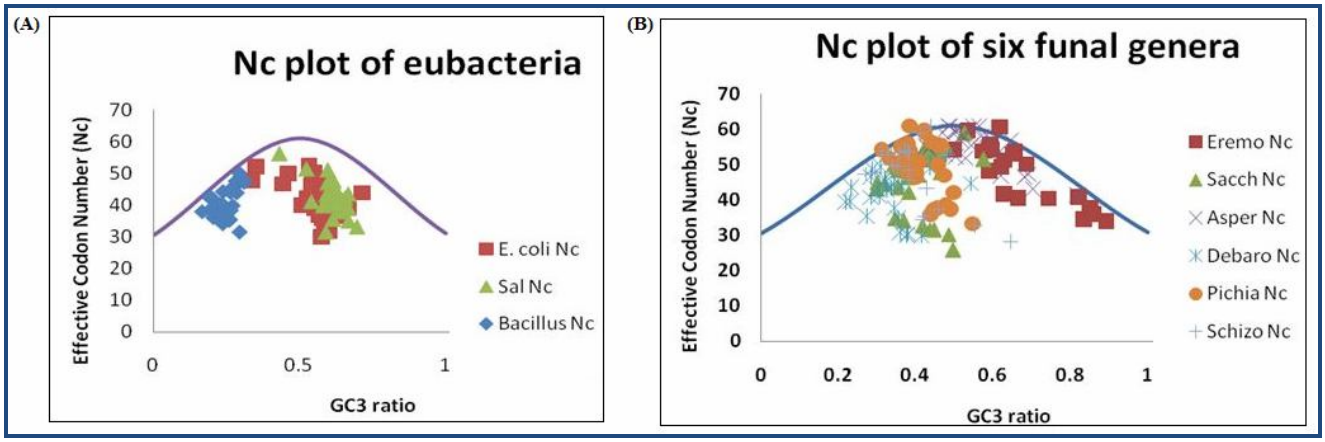

Figure 1: (A) Nc plot of the three eubacterial genera where the markers indicate pentose phosphate pathway gene sequences of E. coli=*Escherichia coli* 55989; Sal=*Salmonella enterica* subsp. *enterica* serovar *Typhimurium*; and Bacillus=*Bacillus cereus* 03BB102. The continuous curve represents the null hypothesis that the GC bias at the synonymous site is solely due to mutation but not selection; **(B)** Nc plot of the six fungal genera where the markers indicate pentose phosphate pathway gene sequences of Eremo=*Eremothecium gossypii* ATCC 10895; Sacch=*Saccharomyces cerevisiae* S288C; Asper=*Aspergillus fumigatus* Af293; Debaro=*Debaryomyces hansenii* var *hansenii* CBS767; Pichia=*Pichia pastoris* GS115; and Schizo=*Schizosaccharomyces pombe* 972h. The continuous curve represents the null hypothesis that the GC bias at the synonymous site is solely due to mutation but not selection.

Methodology:

The complete genome sequences of 10 organisms including both prokaryotes and eukaryotes were downloaded from the Integrated Microbial Genomes website (http:// www.img.jgi.doe.gov) **[13]**. These life forms epitomize diverse metabolic niches ranging from mesophilic plant and animal pathogens, free livings, halotolerants to methylotrophs. They include organisms like *Eremothecium gossypii* ATCC 10895 (= *Ashbya gossypii* ATCC 10895) **[14]**, *Aspergillus fumigatus* Af293 **[15]**, *Bacillus cereus* 03BB102, *Debaryomyces hansenii* var *hansenii* CBS767 **[16]**, *Escherichia coli* 55989 **[17]**, *Homo sapiens* **[18]**, *Pichia pastoris* GS115 **[19]**, *Saccharomyces cerevisiae* S288C **[20]**, *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* SL1344, and *Schizosaccharomyces pombe* 972h **[21]**. Detailed information

regarding the organisms including NCBI sequence id, habitat, relevance etc., is listed in **Table 1 (see supplementary material).**

The nucleotide sequences along with their corresponding amino acid sequences encoding the information for the production of proteins and enzymes of pentose phosphate pathway were sorted out using references from KEGG database **[22-24].** The 10 organisms included in the study, a total of nearly three hundred gene sequences along with their amino acid counterparts was sorted out. The gene sequences of the various enzymes employed in pentose phosphate pathway of *Escherichia coli* 55989 and *Saccharomyces cerevisiae* S288C was taken as the prokaryotic and eukaryotic standard model respectively. The *E. coli* genome has 31 different gene sequences coding for the different enzymes of the pentose phosphate pathway, whereas there are about 28 different gene sequences for the same purpose in *S. cerevisiae*.

The effective number of codons (ENc or Nc), which is a measure of synonymous codon usage bias **[25]**, was calculated for each nucleotide sequence encoding enzymes of the pentose phosphate pathway. Further, the frequency of guanine and cytosine at the synonymous third position of codon, known as GC3 content was calculated. CodonW (http: //codonw. sourceforge.net//) was employed to carry out both these calculations. Nc plots were further constructed by plotting the Nc values against the corresponding GC3 values obtained. Codon Adaptation Index or CAI **[26]**, a commonly used and well-accepted measure for calculating the expression levels of gene sequences was calculated using the CAI Calculator present in the E-CAI server (http: //genomes.urv.es /CAIcal/) **[27].**

A multivariate statistical analysis technique, called correspondence analysis was performed. In this technique, high dimensional data are reduced to a limited number of variables or axes and the most prominent axes contributing to the codon usage variation among the gene sequences is considered **[28]**. Correspondence analysis, based on the pentose phosphate pathway was employed to identify the relation and distinction existing among the different organisms included in our study. Correspondence analysis based on codon usage pattern, relative synonymous codon usage or RSCU and amino acid usage was done to find out the similarities and dissimilarities in terms of synonymous codon usage and amino acid usage pattern.

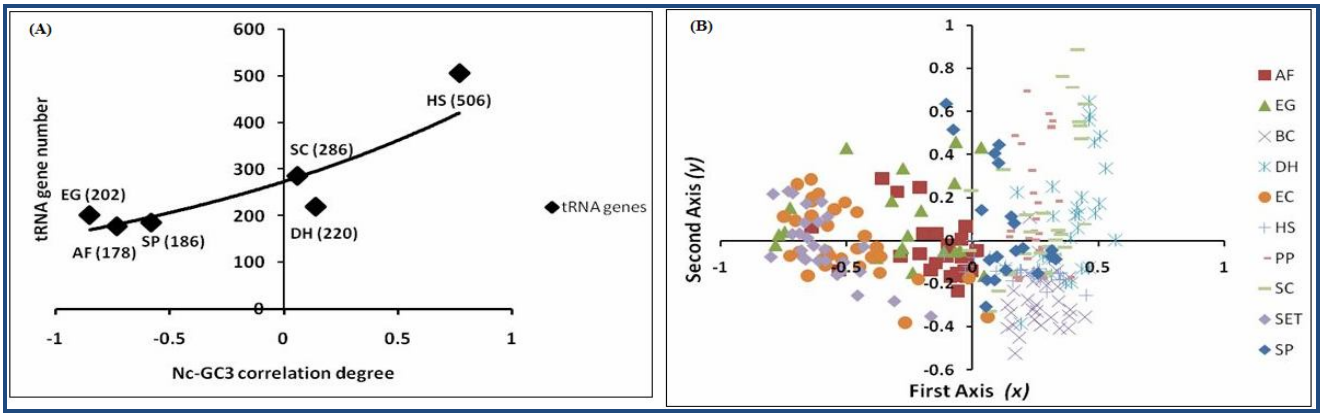

Figure 2: (A) Scatter plot showing exponential increase in the number of tRNA genes (within parentheses) decoding the twenty standard amino acids with respect to increasing degree of positive correlation between Nc and GC3 values in EG= *Eremothecium gossypii* ATCC 10895; AF= *Aspergillus fumigatus* Af293; SP= *Schizosaccharomyces pombe* 972h; DH= *Debaryomyces hansenii* var *hansenii* CBS767; SC= *Saccharomyces cerevisiae* S288C; and HS=*Homo sapiens;* **(B)** Organism wise correspondence analysis on RSCU of pentose phosphate pathway gene sequences in all the test organisms. Organism abbreviations as mentioned in Table1.

Results & Discussion:

Degeneracy due to the presence of synonymous codons is an imperative virtue of the genetic code. The existence of preference or biasness for a certain subset of codons within organisms is governed by different factors which include expression level **[26]**, protein structure and composition **[29],** tRNA abundance **[30]**, GC composition **[31]**, strand specific compositional bias **[32]**, and gene length **[33].** The Nc index is a simple measure of overall codon bias and ranges from twenty to sixty one, where 20 is the value obtained when only one codon is used for each amino acid (i.e., the codon bias is maximum) and 61 is the value obtained when all synonymous codon for each amino acid are equally used (i.e., no codon bias). In our study of the pentose phosphate pathway genes, covering 10 different organisms spanning across diverse metabolic niches we observed that in the prokaryote *E. coli* the effective codon number (Nc) ranges from 29.9 to 52.4 with an average of 42.52, whereas in case of *S. cerevisiae* the effective codon number lies within 25.9 to 58.8 with an average of 45. **Table 2 (see supplementary material)** lists the effective codon number range of the pentose phosphate pathway coding sequences of the different organisms along with the mean Nc scores. From this table it is quite clear that the mean effective codon number of the prokaryotic eubacteria is less than that of the eukaryotes like fungi and *H. sapiens*. The Nc average for the pentose phosphate pathway genes is lowest in terms of *B. cereus* (40.88) whereas it is highest in the case of *A. fumigatus* (54.86) preceded by *H. sapiens* (51.80). The observation that prokaryotic eubacteria returns a lower Nc score is crucial in terms of the fact that it implies they have a higher codon bias compared to the eukaryotes.

Two crucial factors, considered to have strategic effect on codon usage patterns of organisms are natural selection and mutation

pressure **[34].** In order to explore whether the determinative factors for codon usage variation in pentose phosphate pathway genes is mutation pressure or natural selection, we carried out a correlation analysis between Nc score and GC3. In conjunction with this we put together the Nc plot for a better understanding of the forces shaping codon usage distinction.

The Nc plots of the different organisms selected for the study were separately constructed. Looking at the Nc versus GC3 plots **(Figure 1A)**, we observed that in the case of prokaryotic eubacteria, a genus specific clustering of the pentose phosphate pathway genes is prevalent. This is a clear indication of the fact that organism specific codon usage bias percolates into vital metabolic pathway genes irrespective of their near universality. Superimposing the individual Nc plots of the three eubacterial genera, a striking feature emerged showing a clear distinction between the codon usage pattern of gram positive *Bacillus cereus* 03BB102 and gram negative *Escherichia coli* 55989 or *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* SL1344. This discrete clustering of the pentose phosphate pathway genes on the Nc plot based on their gram nature, which is a major classification criterion for bacteria, is a significant upshot of this study. The Nc plots in all these cases indicate the presence of translation selection as a major force in shaping codon usage.

Analyzing the Nc plots of the eukaryotic fungal genera in **Figure 1B,** it is observed that barring *S. cerevisiae* and *D. hansenii*, all the other genera display a certain degree of mutational pressure acting on some of their pentose phosphate pathway genes. Apart from these two genera, many of the gene sequences coding for pentose phosphate pathway enzymes in *Eremothecium gossypii*, *Debaryomyces hansenii* var *hansenii*, *Schizosaccharomyces pombe*, *Pichia pastoris* and *Aspergillus fumigatus* indicate G+C compositional constraints, as they are found to lie on or immediately below the null hypothesis curve. The gene sequences coding for the enzymes of the pentose phosphate pathway genes in *H. sapiens* forms a distinct constellation on the Nc plot where we observe a significant positive correlation (r=0.79; p>>0.01) between the Nc and GC3 values. Moving forward, we worked out the degree of correlation existing between the Nc and GC3 values of the pentose phosphate pathway genes and plotted this against the number of tRNA genes employed by each of this eukaryotic organism for decoding the standard twenty amino acids **[35].** It was observed that there is a marked increase in the number of tRNA genes decoding the twenty standard amino acids with respect to increasing degree of positive correlation between Nc and GC3 values **(Figure 2A).** This points to the fact that as codon bias decreases, large amount of tRNA coding genes are required to carry out translation. In this case, we find that the fungal species *S. cerevisiae* encodes the largest number of amino acid decoding tRNA genes, about 286, whereas in comparison only 178 tRNA genes for the same function is present in *A. fumigatus*. In *H. sapiens* we find the highest number of tRNA genes, about 506 in number employed for decoding the twenty standard amino acids. This observation is thus in line with our general observation regarding the other eukaryotes, where, an increase in positive correlation between Nc and GC3 values point towards reduced codon bias, thus employing higher amount of tRNA genes to encounter base sequence mutation in tRNAs for proper translation.

Correlation analysis of effective codon number and codon adaptation index

A correlation between the Nc score, which is a measure of codon bias and the potential expression level of the gene or CAI was calculated. CAI is a numerical value associated with each gene of a given genome which expresses its synonymous codon bias and helps us to study the effect of translational bias on gene expression. The Codon Adaptation Index ranges from 0 to 1.0, with higher CAI values signifying that the gene of interest has a higher degree of expressivity **[26].** Generally, genes with biased codon usage are potentially highly expressed and in the case of *E. coli*, we observed a significant anti-correlation between Nc and CAI (r= -0.85; p>>0.01) corroborating the fact that biased genes are potentially highly expressed. Similarly, the correlation between Nc and CAI was calculated in case of *S. cerevisiae* and similar degree of anti-correlation was obtained (r = -0.85; p>>0.01). We worked out the correlations between Nc and CAI in the other test organisms and there also we found the existence of significant anti-correlation. This anti-correlation clearly suggest the tendency of nature to conserve gene sequences related to vital functions like pentose phosphate pathway metabolic reactions which is crucial in terms of supplying reducing power to the cell along with pentose phosphates, the building blocks of nucleic acids.

Correspondence analysis on relative synonymous codon usage (RSCU)

To investigate the major trend in relative synonymous codon usage and amino acid usage variation, CodonW was used in carrying out the correspondence analysis. For precisely carrying out the correspondence analysis across the 10 diverse genera selected for this study we meticulously sorted out only those gene sequences which are present in all the organisms or at least in 75% of the organisms subjected to the tests. It was observed that there are nine such sequences present in all the organisms and three sequences present in at least 3/4th of the test subjects, coding for some of the important pentose phosphate pathway enzymes. These include 5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase, 6-phosphogluconate dehydrogenase, fructose 1,6-bisphosphate aldolase, fructose-1,6-bisphosphatase, phosphofructokinase, phosphoglucose isomerase, ribokinase, transaldolase, transketolase, gluconate kinase, phosphoglucomutase and ribose-5-phosphate ketol-isomerase. **Table 3 (see supplementary material)** lists the different functions catalyzed by some of these enzymes.

Correspondence analysis of about three hundred coding sequences encoding pentose phosphate pathway enzymes of the ten organisms were carried out along with correspondence analysis of their amino acid counterparts. Correspondence analysis on RSCU detected a single major trend of codon usage variation and in comparison to all the axes generated, the first axis and second axis accounted for 26.25% and 17.07% of the total variations respectively, which are the highest among all the axes. The position of the genes along the first and second major axes was plotted in **(Figure 2B).** Correspondence analysis on RSCU clearly shows a clear and distinct organism-specific clustering of the pentose phosphate pathway coding sequences. In line with our previous observation, we find that here also there is a clear distinction between the gene clusters of gram negative and gram positive eubacteria. The two gram negative eubacterial genera in this study are found to remain together in

an inseparable cluster on the upper left quadrant of the plot, whereas the gram positive *B. cereus* remain in a tight cluster on the lower right quadrant of the plot **(Figure 2B).**

From this plot we also find that *A. fumigatus* clearly possesses a distinct codon usage pattern among the eukaryotes and forms a separate group in comparison with other eukaryotes, including *H. sapiens* which forms an overlapping core cluster of eukaryotic pentose phosphate pathway coding sequences at the junction of the upper and lower quadrant on the right side of the plot. Correspondence analysis performed on codon usage of the pentose phosphate pathway enzyme coding sequences detected a single major trend of codon usage variation and in comparison to all the axes generated, the first axis and second axis accounted for 28.78% and 19.31% of the total variations respectively. The position of the genes along the first and second major axes was plotted in **(Figure 3A)**. This further substantiate our findings clearly, earmarking the gram positive genus from the gram negative eubacterial genera. The codon usage pattern of *A. fumigatus* was found to be in line with our previous observation. A major feature was the separation of the *H. sapiens* genes in a separate cluster on the right hand side of the plot, a fact that is primarily attributed to the significantly different codon usage pattern in this genus along with its longer gene length. All the remaining fungal yeast genera shared a similar pattern of codon usage as is evident from the correspondence analysis on codon usage.

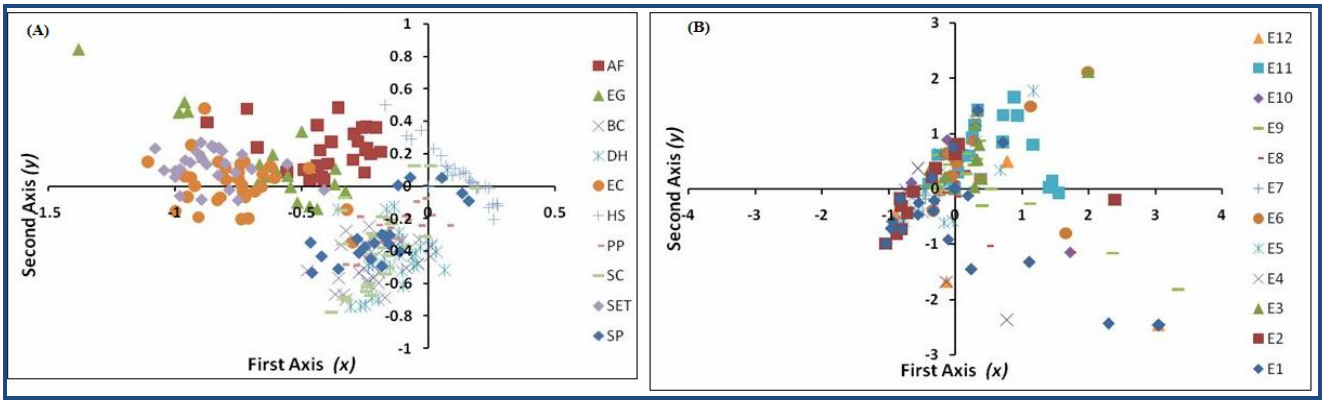

Figure 3: (A) Organism wise correspondence analysis on codon usage of pentose phosphate pathway gene sequences in all the test organisms. Organism abbreviations as mentioned in Table1; **(B)** Enzyme wise correspondence analysis on amino acid usage of twelve different enzymes of pentose phosphate pathway in all the test organisms where, E1=5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase; E2=6-phosphogluconate dehydrogenase; E3=Fructose 1,6-bisphosphate aldolase; E4=Fructose-1,6-bisphosphatase; E5=Phosphofructokinase; E6=Phosphoglucose isomerase; E7=Ribokinase; E8=Transaldolase1; E9=Transketolase; E10=Gluconate kinase; E11=Phosphoglucomutase; and E12=Ribose-5-phosphate ketol-isomerase.

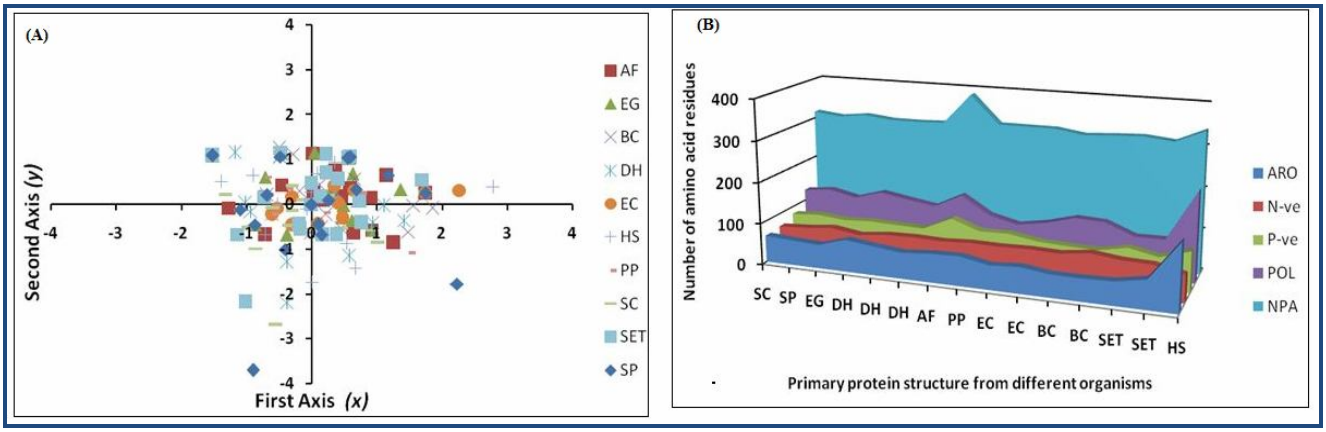

Figure 4: (A) Organism wise correspondence analysis on amino acid usage of pentose phosphate pathway enzyme sequences in all the test organisms. Organism abbreviations as mentioned in Table 1; **(B)** 3-D area plot for comparing the amino acid type distribution in 15 transketolase protein primary structures from ten different organisms (organism abbreviations as mentioned in Table1). ARO=aromatic amino acids; N-ve and P-ve=amino acid with negatively and positively charged side chains respectively; POL=amino acids with polar side chains and NPA=non-polar aliphatic side chain amino acids.

Estimation of relative amino acid usage (RAAU)

Protein function is essentially a virtue of its structure and in order to perform a function with fidelity and efficiency, conservation of structural property is a pre-requisite. The degeneracy prevailing at the order of gene sequences is reduced at the amino acid level significantly to achieve this purpose. In order to find out whether at the level of primary structure there is an organism-specific distinction among the enzymes, we carried out correspondence analysis on relative amino acid usage along with the study of individual amino acid and codon

usage configuration of a vital pentose phosphate pathway enzyme. Correspondence analysis on relative amino acid usage of pentose phosphate pathway proteins and enzymes detected a major trend of amino acid usage variation and in comparison to all the axes generated, the first axis and second axis accounted for 8.77% and 8.52% of the total variations respectively, which are the highest among all the axes. The position of the genes along the first and second major axes was plotted in **(Figure 3B)**. From the correspondence analysis on amino acid usage study we find that though there is a degree of separation in terms of the amino acid usage pattern among the different enzymes of the oxidative and non-oxidative phases of pentose phosphate pathway **(Figure 3B)**, but when comparing among organisms spread out across two different domains of life there is no significant distinction in the amino acid usage pattern of any single enzyme **(Figure 4A).**

Amino acid configuration of transketolase

We want to scrutinize the fact that if there exists any species-specific amino acid configuration in the pentose phosphate pathway enzymes, we made an in-depth study of a key enzyme in the non-oxidative branch of the pentose phosphate pathway that transfers a two-carbon glycoaldehyde unit from ketose-donor to aldose-acceptor sugars, called transketolase. This enzyme is also involved with Calvin cycle in photosynthetic plants as well as bacteria. Transketolase, in mammals, links up the pentose phosphate pathway to glycolysis, feeding surplus sugar phosphates into the core carbohydrate metabolic pathways. Its presence is necessary for the production of NADPH, especially in tissues actively engaged in biosyntheses. Transketolase enzyme activities have also been found to be related to neurodegenerative diseases, diabetes, and cancer **[36].** The codon composition of the enzyme transketolase was analysed in all the ten test organisms and it was observed that the amino acid serine and leucine, which are coded by six triplets have lesser bias in terms of usage of codons. In serine, except AGU, the remaining five triplets AGC, UCA, UCU, UCG and UCC are used randomly, whereas in leucine all the six triplet codons CUG, UUA, CUC, CUU, CUA, and UUG are used without discretion. The scenario is quite different in the case of arginine, the last of the three amino acids coded by six triplet codons where in all the organisms concerned ranging from eubacteria to *H. sapiens* there is a higher degree of bias towards CGU and AGA codons and decreasing order of preference for the other four codons CGC, AGG, CGG and CGA respectively. The inclination among all the test organisms towards the GGU triplet coding for glycine is an interesting finding and these observations justify the fact that in all these organisms, there is a stronger codon-anticodon binding. The binding arising due to formation of five to six hydrogen bonds between codon and anti-codon is preferred **[37].** The role of translational selection in shaping codon composition in vital metabolic pathway gene sequences is thus evident.

Inter-generic amino acid profile comparison of transketolase

Further, analysis of the amino acid profile of about 15 transketolase primary protein structures reveal that the concentration of polar amino acids vary considerably among organisms which might have an effect on the hydrophillic nature and may have a bearing on the habitat or environment of the organism. *A. fumigatus* and *H. sapiens* exhibit a significantly different amino acid usage pattern than *A. fumigatus* having higher levels of non-polar aliphatic amino acids and positively charged amino acids. Supplementary findings from the amino acid composition study of transketolase reveals that—(i) alanine is the most preferred amino acid among all organisms, (ii) tryptophan is the least used aromatic amino acid, (iii) methionine is the least preferred non-polar aliphatic amino acid, (iv) cysteine is the least used amino acid among all organisms, (v) glycine is evenly distributed among all organisms, and (vi) *Homo sapiens* has elevated levels of the amino acids phenylalanine, tyrosine, leucine, isoleucine and serine. Comparing the amino acid distribution in 15 transketolase protein primary structures from ten different organisms we find that **(Figure 4B)** there is a certain degree of predictability in the composition profile of all the organisms barring *A. fumigatus* and *H. sapiens*, where few exceptions are notable, including spiked levels of amino acids with polar side chain and aromatic side chain in *H. sapiens*. In *A. fumigatus*, we observe a significantly different pattern of amino acid family composition with higher concentration of amino acids with non-polar aliphatic side chain, polar side chain and positively charged side chain.

In the case of *A. fumigatus*, a mesophilic human pathogen fungus responsible for invasive aspergillosis, we consistently observed a different codon usage pattern which finally translated in to its amino acid composition model portraying a unique profile for one of its crucial pentose phosphate pathway enzyme, transketolase. Apart from this exception, all our findings substantiate the fact that, in a vital and core energy metabolism pathway genome-specific codon usage bias is overridden or not translated into amino acid degeneracy where purifying selection plays a significant role in safeguarding the form and function of an enzyme.

Conclusion:

The analysis of various codon usage parameters and their inter-relationship, points to the fact that organism-specific codon usage bias is a virtue of an organism. A clear distinction in the codon usage pattern of gram positive and gram negative bacteria in terms of pentose phosphate pathway was an important observation of this study. In the case of eukaryotic fungi, where the genera are quite related some degree of distinction is evident in term of codon usage, but the degree of distinction is less. In the case of *Homo sapiens*, we find a specific and distinct clustering of the pentose phosphate pathway genes in terms of codon usage pattern which may be attributed to its unique codon usage nature and greater gene length. The results of correspondence analysis clearly show that even in the case of a vital life support pathway like pentose phosphate pathway, carrying out important metabolic functions, like generation of reducing power and pentose phosphates for nucleotide synthesis, organism-specific codon usage pattern is clearly evident. Both in terms of codon usage and RSCU pattern, a clear specificity is present that delineates organism. In the human pathogen *A. fumigatus* Af 293, a different codon usage pattern was observed, which finally translated into its amino acid composition model portraying a unique profile in a key pentose phosphate pathway enzyme called transketolase. The functions of a core pathway are universal and functions are a property of enzyme structure, at the level of amino acid usage a distinction among organism specific enzymes is quite blurred, a fact attributable to purifying selection and conservation.

Acknowledgement:
We would like to acknowledge the Department of Biotechnology, Government of India for its two grants BT/BI/04/026/93 and BT/BI/010/019/99.

References:

[1] Ashida Y *et al. IPSJ Digit Cour.* 2008 4: 228
[2] Pal A *et al. Bioinformation.* 2011 **5:** 446 [PMID: 21423891]
[3] Sprenger GA**,** *Arch Microbiol.* 1995 **164**: 324 [PMID: 8572885]
[4] Miosga T & Zimmermann FK, *Curr Genet.* 1996 **30**: 404 [PMID: 8929392]
[5] Lindqvist Y *et al. EMBO J.* 1992 **11:** 2373 [PMID: 1628611]
[6] Blank L M *et al. Genome Biol.* 2005 **6**: R49 [PMID: 15960801]
[7] Kondo H *et al. Biochem J.* 2004 **379**: 65 [PMID: 14690456]
[8] Stephens C *et al. J Bacteriol.* 2007 **189**: 8828 [PMID: 17933895]
[9] Liu S *et al. Microbiology.* 2007 **153**: 3196 [PMID: 17768262]
[10] Collard *F et al. FEBS Lett.* 1999 **2**: 223[PMID: 1051802]
[11] Graille M *et al. Biochimie.* 2005 **8**: 763 [PMID: 16054529]
[12] Ratushny A V *et al. BGRS.* 2006 **2**: 118
[13] Markowitz V M *et al. Nucleic Acid Res.* 2008 **36**: D528 [PMID: 17933782]
[14] Dietrich FS *et al. Science.* 2004 **304**: 304 [PMID: 15001715]
[15] Nierman W C *et al. Nature.* 2005 **438:** 1151 [PMID: 16372009]
[16] Dujon B *et al. Nature.* 2004 **430**: 35 [PMID: 15229592]
[17] Touchon M *et al. PLoS Genet.* 2009 **1**: e1000344 [PMID: 19165319]
[18] Lander E S *et al. Nature.* 2001 **409**: 860 [PMID: 11237011]
[19] De Schutter K *et al. Nat Biotechnol.* 2009 **6**: 561 [PMID: 19465926]
[20] The yeast genome directory, *Nature.* 1997 **387**: 5 [PMID: 9169864]
[21] Wood V *et al. Nature.* 2002 **415**: 871 [PMID: 11859360]
[22] Kanehisa M *et al. Nucleic Acid Res.* 2008 **36**: D480 [PMID: 18077471]
[23] Kanehisa M *et al. Nucleic Acid Res.* 2006 **34**: D354 [PMID: 16381885]
[24] Kanehisa M *et al. Nucleic Acids Res.* 2000 **28**: 27 [PMID: 10592173]
[25] Wright F, *Gene.* 1990 **87:** 23 [PMID: 2110097]
[26] Sharp P & Li W, *Nucleic Acids Res.* 1987 **15**: 1281 [PMID: 3547335]
[27] Puigbo P *et al. Biol Direct.* 2008 **3**: 38 [PMID: 18796141]
[28] Basak S *et al. J Biomol Struct Dyn.* 2004 **22:** 205 [PMID: 15317481]
[29] D'Onofrio G *et al. Gene.* 2002 **300**: 179 [PMID: 12468099]
[30] Kanaya S *et al. Gene.* 1999 **238**: 143 [PMID: 10570992]
[31] Sueoka N & Kawanishi Y, *Gene.* 2000 **261**: 53 [PMID: 11164037]
[32] Lobry JR, *Mol Biol Evol.* 1996 **13**: 660 [PMID: 8676740]
[33] Eyre-Walker A, *Mol Biol Evol.* 1996 **13**: 864 [PMID: 8754221]
[34] Whittle C A *et al. Genome Biol Evol.* 2011 **3**: 332 [PMID: 21402862]
[35] Chan PP & Lowe TM, *Nucleic Acids Res.* 2009 **37** (Database issue): D93 [PMID: 18984615]
[36] Coy JF *et al. Clin Lab.* 2005 **51**: 257 [PMID: 15991799]
[37] Wilhelm T & Nikolajewa S, *J Mol Evol.* 2004 **5:** 598 [PMID: 15693616]

Edited by P Kangueane
Citation: **Pal *et al.*** Bioinformation 9(7): 349-356 (2013)

BIOINFORMATION open access

Supplementary material:

Table 1: Major information regarding the ten organisms featured in the study.

Sl. No.	Organism	Abbreviation	NCBI/RefSeq	Habit/Relevance	Reference
1	*Eremothecium gossypii* ATCC 10895 (=*Ashbya gossypii* ATCC 10895)	EG	NCBI/RefSeq:NC_005782;to NCBI/RefSeq:NC_005788	Mesophile, plant pathogen	(14)
2	*Aspergillus fumigatus* Af293	AF	NCBI/RefSeq:NC_007194;to NCBI/RefSeq:NC_007201	Mesophile, human Pathogen	(15)
3	*Bacillus cereus* 03BB102	BC	NCBI/RefSeq:NC_012472; NCBI/RefSeq:NC_012473	Mesophile, free living	
4	*Debaryomyces hansenii* var *hansenii* CBS767	DH	NCBI/RefSeq:NC_006043;to NCBI/RefSeq:NC_006049;	Mesophile, halotolerant	(16)
5	*Escherichia coli* 55989	EC	NCBI/RefSeq:NC_011748	Mesophile, free living	(17)
6	*Homo sapiens*	HS	NCBI/RefSeq:NC_000001;to NCBI/RefSeq:NC_000024;and NCBI/RefSeq:NC_001807	Mesophile	(18)
7	*Pichia pastoris* GS115	PP	NCBI/RefSeq:NC_012963;to NCBI/RefSeq:NC_012966;	Methylotroph	(19)
8	*Saccharomyces cerevisiae* S288C	SC	NCBI/RefSeq:NC_001133;to NCBI/RefSeq:NC_001148;and NCBI/RefSeq:NC_001224;	Mesophile	(20)
9	*Salmonella enterica* subsp. *enterica* serovar *Typhimurium* SL1344	SET	NCBI/RefSeq:FQ312003	Mesophile, free living, human Pathogen	
10	*Schizosaccharomyces pombe* 972h	SP	NCBI/RefSeq:NC_001326; NCBI/RefSeq:NC_003421;to NCBI/RefSeq:NC_003424;	Mesophile	(21)

Table 2: Effective number of codon (Nc) data for the ten organisms featured in the study.

Organism	Nc scale		Average Nc
	Low	High	
Eremothecium gossypii ATCC 10895	33.80	60.70	47.68
Aspergillus fumigatus Af293	41.19	61.00	54.86
Bacillus cereus 03BB102	31.60	50.00	40.88
Debaryomyces hansenii var *hansenii* CBS767	29.83	52.80	42.45
Escherichia coli 55989	29.90	52.40	42.52
Homo sapiens	46.68	56.01	51.80
Pichia pastoris GS115	33.36	51.00	48.80
Saccharomyces cerevisiae S288C	25.90	58.80	45.00
Salmonella enterica subsp. *enterica* serovar *Typhimurium* SL1344	31.50	56.20	42.23
Schizosaccharomyces pombe 972h	28.10	61.00	47.28

Table 3: Major enzymes of the pentose phosphate pathway along with their functions.

Enzyme	Function(s)	EC id
6-phosphogluconate dehydrogenase	Catalyzes an NADPH regenerating reaction in the pentose phosphate pathway.	EC 1.1.1.44
Phosphoglucose isomerase	Catalyzes the interconversion of glucose-6-phosphate and fructose-6-phosphate; required for cell cycle progression and completion of the gluconeogenic events of sporulation.	EC 5.3.1.9
Ribokinase	Ribokinases phosphorylate ribose to ribose-5-phosphate in the presence of ATP and magnesium.	EC 2.7.1.15
Transaldolase	Enzyme in the non-oxidative pentose phosphate pathway; converts sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate to erythrose 4-phosphate and fructose 6-phosphate.	EC 2.2.1.2
Transketolase	Catalyzes conversion of xylulose-5-phosphate and ribose-5-phosphate to sedoheptulose-7-phosphate and glyceraldehyde-3-phosphate in the pentose phosphate pathway; needed for synthesis of aromatic amino acids.	EC 2.2.1.1
Gluconate kinase	Catalyses the phosphorylation of D-gluconate in the presence of ATP and Mg leading to the formation of 6-P-gluconate.	EC 2.7.1.12
Phosphoglucomutase	Catalyzes the conversion from glucose-1-phosphate to glucose-6-phosphate, which is a key step in hexose metabolism; functions as the acceptor for a Glc-phosphotransferase.	EC 5.4.2.2