NBU Journal of
Plant Sciences
© University of North Bengal. All Rights Reserved

Vol. 3, March 2009
ISSN No. 0974 6927

Investigation of codon and amino-acid usages in a *Rhizobium* phage

S Sur, B Bajwa, M Bajwa, B Basistha, AK Bothra[1] and A Sen*
Molecular Genetics Laboratory, Department of Botany, University of North Bengal, Siliguri 734013, India
[1]Department of Chemistry, Raiganj College, Raiganj

Abstract
In this study codon and amino-acid usage of the phage genes were analyzed to understand the functionality of the protein coding genes. Low bias was noticed. The genes were influenced by mutational bias. Reasonable number of optimal codons was noticed. GC3 content does not play any role in synonymous codon usage; however gene expression levels and hydropathicity determine codon and amino acid usage variation. Highly expressed genes play an important role in infectivity.

Keywords: *Rhizobium* phage 16-3, codon usage, amino acid usage, CAI, hydropathicity

In order to develop any pathogenic or symbiotic interactions and sites for bacteriophage attachment the bacterial surface is of utmost significance (Putnoky *et al.*, 2004). The bacteriophage 16-3 is a temperate phage of *Rhizobium meliloti* (Ordogh and Szende, 1961) as well some other rhizobacteria. During the process of infection symbiotic nitrogen-fixing partners of various leguminous plants as well as exopolysaccharides, capsular polysaccharides, and lipopolysaccharides play a vital role in rhizobia ((Putnoky *et al.*, 2004). The bacteriophage 16-3 is known to introduce its genome into the chromosome of *Rhizobium meliloti* by the mechanism of site-specific recombination (Papp *et al.*, 1993). It is a double stranded DNA phage with an overall GC content of 58.95%. The genome size of this virus belonging to the *Caudovirales* family is 60195 bp (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) and harbors 110 protein coding genes. Very little is known about the structure and function of the protein coding genes. In this regard, the study of synonymous codon usage is essential in understanding how the protein-coding genes in this organism have been shaped.

It is well known that synonymous codon usage is non-random, species specific (Sur *et al.*, 2007), and is influenced by directional mutational bias and translational selection (Sau *et al.*, 2007). Codon bias varies considerably between and within organisms across the genome (Sen *et al.*, 2007). Codon and amino usage bias has already been studied in T_4, T_7, mycobacteriophages, staphylococcal phages etc. enriching the science of phage biology (Sau, 2006). Amino acid usage is known to be predisposed by hydrophobicity, aromaticity, cysteine content etc. (Lobry and Gautier, 1994; Garat and Musto, 2000; Zavala *et al.*, 2002; Banerjee *et al.*, 2004; Basak *et al.*, 2004; Naya *et al.*, 2004). In this communication we have attempted to look into the synonymous codon usage bias and amino acid usage of *Rhizobium* phage 16-3 to infer upon its way of life and identify predicted highly expressed genes.

Materials and Methods

Genome sequence of the *Rhizobium* phage 16-3 (Accession No. NC_011103) was obtained from the IMG database (Markowitz *et al.*, 2006). All of the protein coding genes were explored using Codon W software (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=codonw) (Sur *et al.*, 2008) and ACUA (Umashankar *et al.*, 2007). ACUA was used to calculate the GC content, GC3 (amount of G or C codons in the third position), Nc (Effective number of codons), GC skew and AT skew (Sur *et al.*, 2008). Nc is a simple measure of codon bias whose value ranges from 20 to 61 (Ghosh, 2004). The codon adaptation index (CAI) was calculated with ACUA. CAI is a measure of the relative adaptiveness of the codon usage of genes in the direction of codon usage of highly expressed genes (Wu *et al.*, 2005). CAI values fluctuate from 0 to 1 with higher values signifying that gene of concern has a codon usage pattern similar to reference genes. Hydropathicity (GRAVY score) and aromaticity (Peden, 1999) of the protein coding genes were determined using Codon W. The former is calculated as arithmetic mean of the sum of the hydropathic indices of each amino acid, while the latter is an index for amino acid usage given that disparity in amino acid composition can have application for analyzing codon usage (Lobry and Gautier, 1994). Correspondence analysis is a type of multivariate statistical analysis. It was carried out on codon usage of codon count and amino acid usage using Codon W (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=codonw). The technique explore key trends codon and amino-acid usage disparity among the genes within the genome.

Results and Discussion

The initial purpose was to settle upon the degree of codon usage variation among the viral genes. Codon heterogeneity is commonly coupled with the expression level of the genes (Sen *et al.*, 2008). The GC3 and Nc values of the genes were analyzed to find out the degree of heterogeneity. Effective number of codons and GC

*Corresponding author:
E-mail: senarnab_nbu@hotmail.com

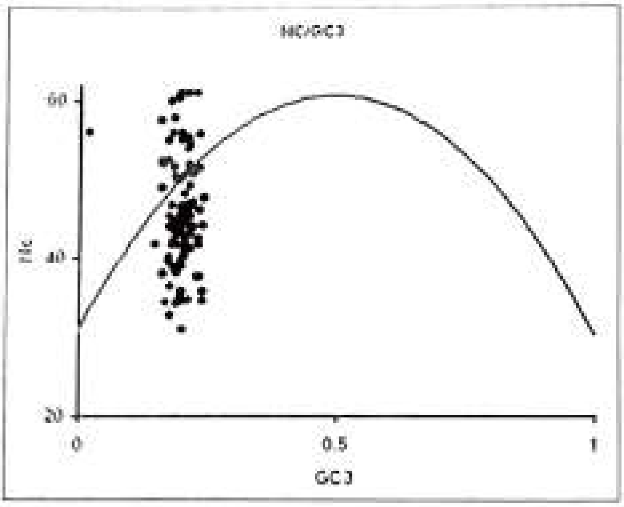

Fig. 1: Effective number of codons used (Nc) in each gene (Y axis) plotted against the G+C content at synonymous third position of codons (GC3) (X axis).The continuous curve denotes the null hypothesis that GC bias at the synonymous site is exclusively due to mutation and not selection.

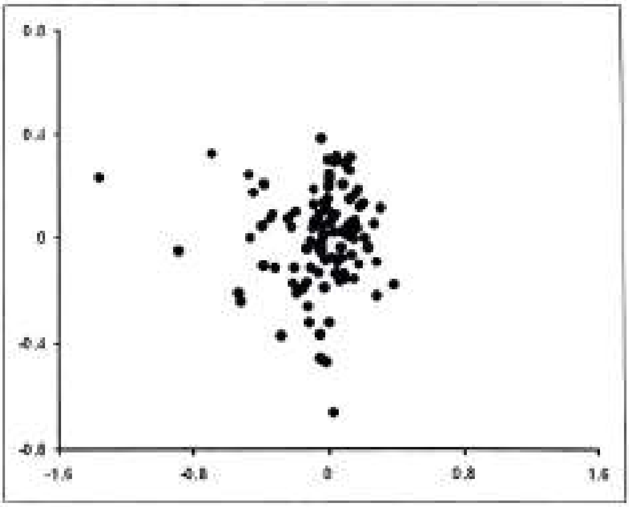

Fig. 2: Correspondence analysis of codon usage patterns on codon count. X and Y axes correspond to axes 1 and 2 of the analysis

content at the third position detects codon usage variation among the genes with the Nc/GC3 plots highlighting the variations. Nc values ranged from 30±1 to 61±0, while GC3 values ranged from 15±1 to 24±1. Although the Nc values varied, GC3 values did not show significant variation implying that there is marked heterogeneity among the genes. Low mean Nc values indicate that there is not much bias among the genes. The GC content is moderate, while the GC3 percentage is extremely low.

Figure 1 shows the Nc/GC3 plot for the studied genes. It is seen from the figure that a small number of genes are lying on the expected curve which is undoubtedly derived from extreme mutational bias and a few number of genes lying above the expected curve. If GC3 content exclusively prejudiced codon bias Nc values would have had fallen on the expected curve in the GC3 and Nc plot suggesting that *Rhizobium* phage genes have additional codon bias independent of GC3s. The mean Fop value for the protein coding genes in *Rhizobium* phage 16-3 is moderate. If mutational bias exclusively controlled codon usages bias the protein coding genes would have had low Fop values. Moderate Fop value indicates the presence of reasonable number of optimal codons.

Table 1: Potentially highly expressed (PHX) genes of *Rhizobium* phage 16-3

Genes	Locus tag	CAI value
P-049	RM163_049	0.815
P-079	RM163_079	0.813
P-073	RM163_073	0.794
P-020	RM163_020	0.792
P-077	RM163_077	0.791
P-078	RM163_078	0.788
P-031	RM163_031	0.784
P-106	RM163_106	0.783
P-094	RM163_094	0.781
P-082	RM163_082	0.776
Putative tail fiber protein H	RM163_022	0.769

Correspondence analysis on codon count of the 110 protein coding genes of *Rhizobium* phage 16-3 was carried out to inspect whether amino-acid compositions play a role in codon usage. It is quite apparent from the plot (Figure 2) that there is not much variation in codon usage among the genes in *Rhizobium* phage. The position of the first major axis of variation was negatively correlated with A3 (r=-0.256), positively correlated with G3 (r=0.235), T3 (r=0.190) and C3 (r=0.194). CAI values were positively correlated (r=0.390) with Axis 1 and strong negative correlation with Nc (r=-0.7535). No significant correlation was obtained for GC3. These results indicate that expression levels of the genes manifested by the CAI values play an important role in codon usage variation among the genes. Strong negative correlation of the principal axis of variation with Nc values may be due to decrease in codon bias among genes lying towards the left of Axis 1. No correlation with GC3 may be due to absence of any function of GC3 compositions in effecting synonymous codon usage variation.

Correspondence analysis of amino acid usage (Figure 3) for the 110 protein coding genes revealed two major axis of variation. It was carried out to identify the forces in defining functionality of the encoded proteins. The first major axis of variation showed strong negative correlation (r=-0.814) with grand average hydropathicity (GRAVY) score. GRAVY score depicting the measure of hydrophobic character present in the amino acid is used during analysis of the structure of a protein sequence. However, almost no correlation was obtained for the principal axis of variation and aromaticity. Few numbers of genes present on the negative side of the axis are anticipated to have GC rich amino acids.

CAI is a measure of gene expression level. As per Wu *et al.*, 2005, top 10% of the genes in terms of CAI values are potentially highly expressed. The corresponding CAI cut off for the protein coding genes of *Rhizobium* phage was 0.769. The potentially highly expressed genes are

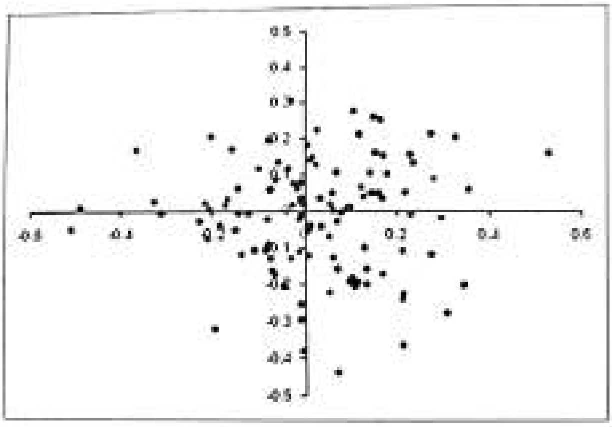

Fig. 3: Correspondence analysis of codon usage patterns on amino acid usage. X and Y axes correspond to axes 1 and 2 of the analysis

shown in Table 1. To get insights into the functional distribution of the potentially highly expressed genes the Cluster of Orthologous genes (COGs) were studied. COGs consist of paralogs from three lineages matching up with ancient phylogenetic lines. Out of the 6 categories of COG genes in *Rhizobium* phage none of them were found to fall in the highly expressed category. Most of the potentially highly expressed genes are associated with phage infectivity and integration. The presence of highly expressed genes associated with infectivity indeed helps the phage in survival.

Conclusion

Rhizobium phage genes show low codon bias. The genes may be influenced by mutational bias and possess reasonable number of optimal codons. GC3 content does not play any role in synonymous codon usage, however gene expression levels determine codon usage variation. Hydrophobic nature of the amino-acids is revealed by amino acid usage. Potentially highly expressed genes associated with infectivity divulge its role as a phage

Acknowledgement

The authors are grateful to the Department of Biotechnology (DBT), Govt. of India for providing financial help in setting up Bioinformatics Facility in the University of North Bengal.

References

Banerjee T, Basak S, Gupta SK, Ghosh TC. 2004. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. *J Biomol Struct Dyn* 22: 13-23

Garat B, Musto H. 2000. Trends of Amino acids usage in the proteins from the unicellular Parasite *Giardia Lamblia*. *Biochem Biophys Res Commun* 279: 996-1000

Ghosh TC, Gupta SK, Majumdar S. 2000. Studies on codon usage in *Entamoeba histolytica*. Int J Parasitol 30:715-722

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res* 22: 3174-3180

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, *et al.* 2006. An experimental metagenome data management and analysis system. *Bioinformatics* 22:e359-67

Naya H, Zavala A, Romero H, Rodriguez-Maseda H, Musto H. 2004. Correspondence analysis of amino acid usage within the family *Bacillaceae*. *Biochem Biophys Res Commun* 325: 1252-1257

Ordogh F, Szende K. 1961. Temperate bacteriophages isolated from *Rhizobium meliloti*. *Acta Microbiol Hung* 8: 65-71

Papp I, Dorgai L, Papp P, Jonas E, Olasz F, Orosz L. 1993. The bacterial attachment site of the temperate *Rhizobium* phage 16-3 overlaps the 3' end of a putative proline tRNA gene. *Mol Gen Genet* 240: 258-264

Peden J. 1999. Analysis of codon usage. PhD thesis, The University of Nottingham, UK

Putnoky P, Deak V, Bekasi K, Palvolgyi A, Maasz A, Palagyi Z, Hoffmann G, Kerepesi I. 2004. H protein of bacteriophage16-3 and RkpM protein of *Sinorhizobium meliloti* 41 are involved in phage adsorption. *J Bacteriol* 186: 1591-1597

Sau K, Gupta SK, Sau S, Mandal SC, Ghosh TC. 2007. Studies of synonymous codon and amino acid usage biases in broad host range bacteriophage KVP40. *The J Microbiol* 45: 58-63

Sau K. 2007. Studies on synonymous codon and amino acid usage in *Aeromonas hydrophila* phage Aeh1: architecture of protein-coding genes and therapeutic applications. *J Microbiol Immunol Infect* 40:24-33

Sen A, Sur S, Bothra AK, Benson DR, Normand P, Tisa LS. 2008. The implication of life style of codon usage patterns and predicted highly expressed genes for three *Frankia* genomes. *Anton van Leeuwen* 93: 335-346

Sur S, Sen A Bothra A. 2007. Codon usage analysis of some archaeal (methanogenic) nitrogen fixing genes: relationship to gene expression and biasness. *Bioinform Trends* 2: 47-60

Sur S, Bhattacharya M, Bothra AK, Tisa LS, Sen A. 2008. Bioinformatic analysis of codon usage patterns in a free living diazotroph *Azotobacter vinelandii*. *Biotechnology* 7:242-249

Umashankar V, Arun Kumar V, Sudarsanam D. 2007. ACUA: A software tool for automated codon usage analysis. *Bioinformation* 2: 62-63

Wu G, Culley DE, Zhang W. 2005. Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiol 151: 2175-2187

Zavala A, Naya H, Romero H, Musto H. 2002. Trends in codon and amino acid usage in *Thermotoga maritima*. *J Mol Evol* 54: 563-568.