ISSN 1682-296X (Print)
ISSN 1682-2978 (Online)

Bio Technology

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Biotechnology 7 (2): 242-249, 2008
ISSN 1682-296X

Bioinformatic Analysis of Codon Usage Patterns in a Free Living Diazotroph, *Azotobacter vinelandii*

[1]Saubashya Sur, [1]Malay Bhattacharya, [2]Asim K. Bothra, [3]Louis S. Tisa and [1,3]Arnab Sen
[1]Department of Botany, Faculty of DBT Bioinformatics, University of North Bengal, Siliguri, WB 734013, India
[2]Laboratory of Chemiinformatics Bioinformatics, Raiganj University College, Raiganj, WB 733134, India
[3]Department of Microbiology, University of New Hampshire, Durham, NH 03824, USA

Abstract: Synonymous codon usage analysis of the protein coding genes, nitrogen fixation related genes and ribosomal protein genes of *Azotobacter vinelandii* were performed and potentially highly expressed genes were detected. Codon usage was highly biased. Genes in the genome exhibited considerable amount of heterogeneity. However, unlike ribosomal protein genes, which were governed by translational selection, those genes associated with nitrogen fixation were affected by mutational pressure. Using the Codon Adaptation Index (CAI) as a numerical estimator of gene expression level, highly expressed genes in *Azotobacter* were predicted with ribosomal protein genes as a reference. Highly expressed genes are affluent in of GC rich codons. We have identified 503 potentially highly expressed genes having diverse functions. PHX genes present in the COG groups were identified. Most of them are associated with major metabolic functions. Ten PHX genes linked to the nitrogen fixing mechanism has also been identified. These results specify the capability of the bacterium to survive in a free-living state, compete with other soil bacteria and fix nitrogen in a manner somewhat different from the conventional.

Key words: *Azotobacter*, nitrogen fixation, codon usage, COGs, PHX

INTRODUCTION

Nitrogen is a plant nutrient, which is commonly deficient in most soil environment contributing to reduced agricultural yields throughout the world (Dixon and Kahn, 2004). Although molecular nitrogen or dinitrogen (N_2) makes up four-fifths of the atmosphere, it is metabolically unavailable for direct use by higher plants or animals. Several microbial species are able to convert atmospheric nitrogen to ammonia by the enzyme nitrogenase. Nitrogenases are composite metalloenzymes with conserved structural and mechanistic characteristics (Rees and Howard, 2000; Lawson and Smith, 2002).

The free living diazotroph *Azotobacter vinelandii* is a gram negative, strictly aerobic bacterium, with broad ranging metabolic capabilities (http: //genome.jgi-psf.org/draft_microbes/azovi/azovi.info.html). This bacterium can grow on a wide variety of carbohydrates, alcohols and organic acids (Rediers *et al*., 2005). *A. vinelandii* is of unusual interest to the scientists engaged in nitrogen fixation studies due to two important features: (i) besides molybdenum-containing nitrogenase enzyme, they synthesize two other nitrogenases; one in which molybdenum is replaced by vanadium and a second which contains only iron (Eady, 1991, 1996; Bishop and Premakumar, 1992). (ii) *A. vinelandii* has developed a number of physiological mechanisms to permit it to fix nitrogen aerobically (Dixon and Kahn, 2004), one of which is high respiration rate, which prevents oxygen reaching the site of nitrogenase reaction (Rediers *et al*., 2005).

Recently, the *A. vinelandii* genome was sequenced and the availability of these sequence data opened up the possibility to explore molecular nature of the activities and potential gene expression by this organism. An attempt was made in this work using bioinformatics tools to study the synonymous codon usage patterns. Synonymous codon usage is species specific and differs significantly among the genes within the same organism (Peden, 1999). Diverse patterns of codon usage may arise from dissimilar factors. It has been reported that directional mutational pressure and natural selection working at the level of translation are the main reasons of codon usage variation among the genes in different organisms (Grantham *et al*., 1981; Peden, 1999). In extremely AT or GC rich unicellular

Corresponding Author: Dr. Arnab Sen, Department of Botany, Faculty of DBT Bioinformatics, University of North Bengal, Siliguri, WB 734013, India Tel: 91-353-2699106/91-353-6528172 Fax: 91-353-2699001

organisms, compositional bias shapes codon usage variation among the genes (Gupta *et al*., 2004). Besides other mechanisms, codon usage bias influence gene expression by favoring the translation rate (Ikemura, 1981; Bernardi, 1995). It is based on the selection of the third codon position to acclimatize coding sequences to the most abundant tRNAs in the cell or to those with more efficient codon-anticodon interaction (Martin-Galiano *et al*., 2004). In lowly expressed genes, codon usage may be governed by mutational bias since they are less controlled by translational selection (Banerjee *et al*., 2004). To analyze the patterns of codon usage and assess the degree and direction of codon bias, many indices have been proposed. Among these indices, the Codon Adaptation Index (CAI) was anticipated as a measure of codon usage within a gene relative to a reference set of genes (Sharp and Li, 1987; Sen *et al*., 2007). This index has been shown to associate with mRNA expression levels (Peden, 1999) and has been used to predict highly expressed genes in various organisms (Dos Reis *et al*., 2003; Martin-Galiano *et al*., 2004; Wu *et al*., 2005a,b). In addition to CAI, the effective number of codons (Nc), which is defined as the number of equal codons that would generate the same codon usage bias as observed and the frequency of optimal codons (Fop), which is defined as the proportion of synonymous codons that are optimal (Peden, 1999), are also used for the same purpose. An optimal codon is one codon whose incidence of usage is appreciably higher in putatively highly expressed genes (Stenico *et al*., 1994).

The aim of the present study was to analyze the synonymous codon usage patterns and envisage expression level of the protein coding genes of *A. vinelandii* with special reference to those genes involved in nitrogen fixation. A Cluster of Orthologous Groups (COGs) consist of individual proteins and paralogs from at least three lineages corresponding to an ancient conserved domain. (Tatusov *et al*., 2003). We have explored the correlation between the predicted expression level of the genes present in various COG groups and the life style of *Azotobacter vinelandii*. We believe that the outcome of this study will be helpful to the community of scientists who work on this important and unusual microbe.

MATERIALS AND METHODS

The genome sequence of *Azotobacter vinelandii* AvOP (henceforth will be referred to as AvOP) was obtained from Integrated Microbial Genomes (http: //www.img.jgi.doe.gov) website. All of the protein coding genes, nitrogen fixation associated genes, TTA codon containing genes and those genes allied with the ribosomal proteins were taken for the study.

The software CodonW (http: //codonw.sourceforge.net) (Peden, 1999) was used to determine genomic G+C composition in the third position of codon (GC3s), effective number of codons (Nc) (Peden, 1999) and frequency of optimal codon (Fop) values (Sur *et al*., 2007). The Codon Adaptation Index (CAI) (Wu *et al*., 2005a) was calculated taking the codon usage of the ribosomal protein genes, which are certainly highly expressed as reference. CAI value was calculated by using a web-based application: the CAI Calculator 2 (http: //www.evolvingcode.net/codon/cai/cais.php) (Wu *et al*., 2005a).

GC3 signifies the frequency of guanine and cytosine at the synonymous third positions of codons. The effective number of codons (Nc) serves as a measure of general codon bias (Sur *et al*., 2006). The Nc value depicts the number of equal codons that would create the same codon usage bias as was observed. Nc values range from 20 (when only one codon is per amino acid) to 61 (when all codons are used in equal probability).

Codon Adaptation Index (CAI) is an extensively used measure of codon bias in prokaryotes and eukaryotes (Peden, 1999). It is a measurement of relative adapted-ness of a gene's codon usage towards the codon usage of highly expressed genes. The relative adapted-ness of each codon is the relationship of the usage of each codon, to that of the most abundant codon within the same synonymous family (Peden, 1999).

F_{op} is the fraction of synonymous codons that are optimal codons (Peden, 1999). Its value ranges form 0 (denoting a gene has no optimal codons) and 1.0 (when a gene is wholly composed of optimal codons). It is just a ratio between the incidence of optimal codons and the total number of synonymous codons. For the equations of calculating Nc, CAI and Fop values please refer to Sur *et al*. (2006, 2007).

In order to test whether the values of the aforesaid indices in nitrogen fixing genes, ribosomal protein genes and TTA codon containing genes significantly differ from that of the protein coding genes, Z score (Walpole *et al*., 2004) was determined. It gives the standard normal cumulative distribution function. Random samples of the genes were taken. The average of each of the indices from the sample was calculated. This furnishes the random value. The process was repeated 100 times and the average computed. This is the mean. The Standard Deviation (SD) was also calculated. With these data the Z score was calculated using the formula:

$$Z = \frac{(\text{Mean} - \text{Observed})}{\text{SD}}$$

Where, the observed value is the one observed for the indices for the types of genes undertaken in the study.

Correspondence analysis (COA) was performed using CodonW (http: //codonw.sourceforge.net) (Peden, 1999). Correspondence analysis creates a series of orthogonal axes to identify trends to explain the data variation, with each subsequent axis explaining a decreasing amount of the variation (Benzecri, 1992). It was carried out on simple codon count and amino acid frequencies. The file containing the gene sequences were loaded in Codon W (Peden, 1999). For calculating the former the correspondence analysis menu (Menu 5) was selected. It had four options. Option 1 was used for correspondence analysis on codon count. In this option advanced correspondence analysis sub option was preferred so as to have greater control during correspondence analysis. The toggle level was changed to exhaustive; the numbers of axis altered and the program was run. Correspondence analysis on amino acid usage was performed with the help of option 3 in the correspondence analysis menu (Menu 5).

RESULTS AND DISCUSSION

Overall synonymous codon usage: The primary aim in this study was to detect the degree of codon usage heterogeneity which is generally linked with gene expression level. Highly expressed genes have higher frequencies of codons considered optimal for translation (Lafay *et al*., 2000). Most bacteria having a balanced AT/GC genome content, show considerable amount of codon heterogeneity (Sen *et al*., 2007). Since *A. vinelandii* AvOP has a high G+C content (65.71%), the GC3s and Nc values for all genes in this genome were calculated to determine if heterogeneity exists among genes in this genome. Figure 1 shows the Nc /GC3s plot, which have been suggested to be an effective means to explore the codon usage variations among genes in the same genome (Peden, 1999). The Nc values of AvOP genes range from 24 to 61 suggesting that this GC-rich genome exhibited substantial heterogeneity in codon usage. The genes encoding ribosomal proteins, which are anticipated to be expressed at high levels during rapid cell growth, were recognized and are highlighted in the Nc plots. Most of the Ribosomal Protein Genes (RPGs) of the AvOP genome cluster at the low ends of the plot, which is quite analogous to the results observed in genomes of *Xanthomonas* (Sen *et al*., 2007), *Escherichia coli* and *Streptomyces* (Wu *et al*., 2005a) and designate a significant strong codon bias in these genes resulting out of selection for translational efficiency (Cutter *et al*., 2003). The position of the genes related with nitrogen

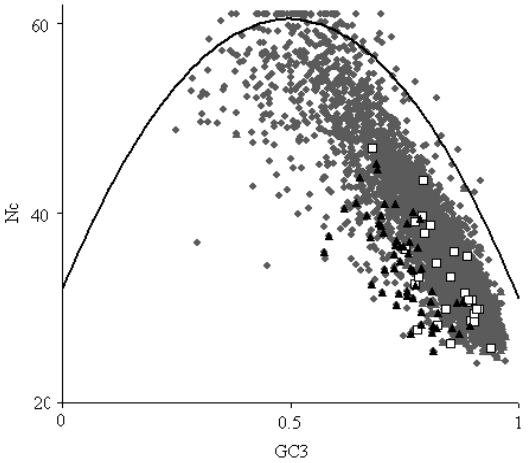

Fig. 1: Effective number of codons (Nc) plotted against the GC content at the synonymous third position in *Azotobacter vinelandii* AvOP. The continuous curve represents the null hypothesis that the GC bias at the synonymous site is exclusively due to mutation, but not selection (Sen *et al*., 2007). The protein coding genes are represented by gray circles, nitrogen fixation related genes in white squares and ribosomal protein genes in black triangles

fixation (NFGs), are also shown in the Nc plots (Fig. 1). NFGs are more or less clustered along with the ribosomal protein genes. The continuous curve in the Fig. 1 indicated the factor influencing codon usage bias. If synonymous codon bias was completely dictated by GC3s, Nc values should fall below the expected curve of Nc/GC3 (Sur *et al*., 2006, 2007). Nevertheless, we found that apart from a very few genes, the values obtained for the bulk of the genes were well below the expected values (Fig. 1).

Table 1 shows the mean values of different indices used to study codon usage patterns in AvOP. We can see from Table 1 that the effective number of codons (Nc) decreased with the corresponding increase of GC3. The low Nc values specify a high degree of codon bias. The nitrogen fixation related genes and ribosomal protein genes are more biased than the protein coding genes as evidenced by their lower Nc values. On the other hand, ribosomal protein genes and nitrogen fixation related genes had elevated Fop values compared to the PCG values. The higher Fop value indicates the presence of higher proportion of optimal codons in these genes. If mutational bias barely influenced codon bias, these genes would have had a low Fop value. The values of NFG, RPG and TTA shown in Table 1 were meticulously tested with

Table 1: Mean values of, GC%, GC3%, Nc, CAI and Fop of the studied genes in the *Azotobacter vinelandii* AvOP

Genes group	GC%	GC3%	Nc	CAI	Fop
Protein coding genes (PCG)	65.5±5.3	81.4±11.5	36.9±7.8	0.62±0.13	0.58±0.07
N_2 fixation related genes (NFG)	63.9±4.1	84.0±6.2	33.1±5.2	0.72±0.09	0.63±0.05
Ribosomal protein genes (RPG)	60.5±3.5	74.7±6.8	34.6±4.9	0.71±0.08	0.63±0.05
TTA codon containing genes (TTA)	59.9±7.8	68.3±15.2	45.5±9.1	0.47±0.15	0.51±0.08

Mean±Standard deviation

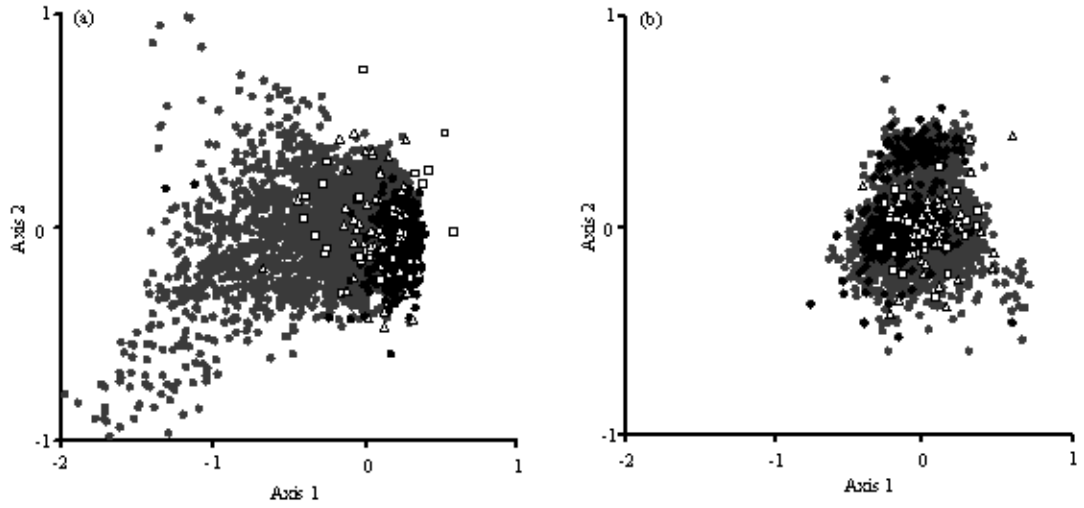

Fig. 2: Correspondence analysis of simple codon count (a) and amino acid usage (b) for the *Azotobacter vinelandii* AvOP genome. For each plot, the X and Y axis correspond to axis 1 and axis 2 of the analysis. Protein coding genes are represented by gray circles, nitrogen fixation related genes by white squares, ribosomal protein genes by white triangles and highly expressed genes by black circles

that of PCG for any significance difference. The Z values for NFGs, RPGs and TTA codon containing genes of different indices revealed differences from PCGs. Although Z values of CAI and Fop showed moderate disparity, significant differences were observed for the same in GC3 and GC. On the other hand the Z-values of Nc for RPGs and NFGs varied quite significantly from that of PCG. These observations imply that there is discrepancy in the characteristics of the studied genes even though they belong to the same genome.

Correspondence analysis: The multivariate statistical analysis is a commonly used method to study the variations in codon usage among the genes in different organisms (Ghosh *et al*., 2000). Correspondence analysis is one of the most vital multivariate statistical technique in which the data is plotted in a multidimensional space of 59 axes (excluding Met, Trp and stop codons) and 20 axis in case of codon usage and amino acid usage, respectively and then the most prominent axes contributing to variation of codon usage or amino acid usage among the genes is determined (Banerjee *et al*., 2004).

Figure 2a shows the positions of the genes along the first and second major axes of the correspondence analysis on simple codon count. Correspondence analysis on simple codon count scrutinizes whether amino acid compositions put forth any control on synonymous codon usage. It was seen that correspondence analysis on simple codon count accounted for 47.5 and 5.94% of the total variation of the first and second major axes, respectively. Thus, in AvOP genome, there is a single major explanatory axis on the codon usage variation among the genes. The position of the genes on the first major axis showed significant negative correlation ($r = -0.825$, $p<0.001$) with Nc and a strong positive correlation with gene expression level ($r = 0.908$, $p<0.001$). Negative correlation of the principal axis with Nc is caused by decrease in codon bias among the genes lying towards the left of Axis 1. We have not found any correlation with GC3. Thus local variations in GC3 content do not play any role in synonymous codon selection. Interestingly it is seen from Fig. 2a that majority of the highly expressed genes are clustered together at one end of the major axis produced by correspondence analysis on codon count. The highly expressed genes are also lying on the positive side of the first major axis. The nitrogen fixation related genes are dotted in the centre of the axis. The position of the genes on the first major axis shows

strong positive correlations with C3 and G3 and significant negative correlations with T3 and A3.

The correspondence analysis on amino-acid usage was done to identify the probable forces in defining the functional adaptations of encoded proteins. Correspondence analysis on amino acid usage accounted for 14.85 and 3.25% of the total variation in protein amino acid content, respectively. This observation also suggests that there is a single major explanatory axis on the amino acid variation among these genes. The positions of the first two major axes are plotted in Fig. 2b. It is seen that highly expressed genes and other genes form groups along the horizontal axis. The position of the genes on the first major axis was correlated with CAI. A weak negative correlation was observed with CAI ($r = -0.115$, $p<0.001$). A number of highly expressed genes are located on the negative side of the first major axis and it may be assumed that these genes may be affluent in GC rich amino acids compared to the lowly expressed genes. The nitrogen fixation related genes remain scattered and those which are present in the negative side of the major axis are expected to be having greater numbers of GC rich amino acids.

Identification of predicted highly expressed (PHX) genes: In the past, CAI values have been extensively used to calculate the expressivity of genes (Gupta *et al*., 2004; Wu *et al*., 2005a, b). This index evaluates the degree to which selection has been successful in molding the pattern of codon usage. In that respect, this index is helpful for predicting the level of expression of a gene (Peden, 1999). The CAI values were calculated for all the genes with highly expressed ribosomal protein genes used as references. The distributions of the CAI values are shown in Fig. 3. The CAI values ranged from 0.17-0.90, with the majority of the genes having CAI values between 0.55 and 0.85. The median CAI value for the genes was 0.648. Mean CAI values of the nitrogen fixation related genes and ribosomal protein genes are higher in comparison to the protein coding genes (Table 1). Further analysis showed no significant correlation between CAI values and gene length suggesting that codon bias was not the key mechanism shaping the efficient translation of long genes. A significant positive correlations of CAI values was observed with GC3 ($r = 0.866$, $p<0.001$). These observations suggest that gene expression levels elevated with the increase in GC3 composition. The plot of the frequency distribution of CAI values (Fig. 3) has a distinct distribution patterns that picked in the 0.65-0.70 CAI range, which was followed by steady decline. This result suggests that majority of the genes were moderately expressed in AvOP.

As defined by Wu *et al.* (2005a), the top 10% of the genes, in terms of CAI values, were classified as the Predicted Highly Expressed (PHX) genes. This corresponded to a cut-off value of 0.764 and included 503 genes with a median CAI of 0.648. These 503 genes included 14 ribosomal protein genes and 10 nitrogen fixation related genes. Table 2 shows the top 20 PHX genes in the genome.

Functional analysis of PHX genes: Clusters of Orthologous Groups of proteins (COGs) were used to recognize the functional distribution of the PHX genes in the *Azotobacter* genome. Each COG type consists of individual proteins or groups of paralogs from at least 3 lineages and thus, corresponds to a primeval conserved domain. To help the investigation, each of the COG categories were clustered in the subsequent 4 COG

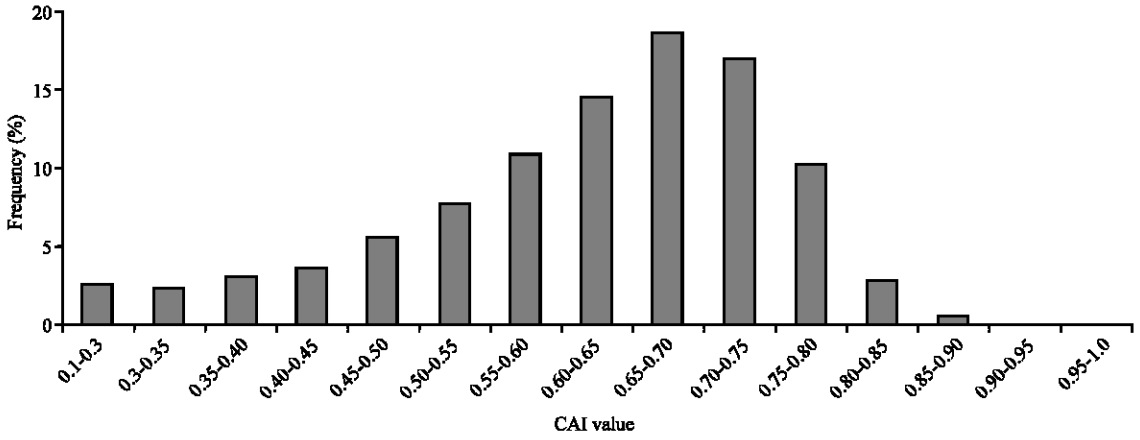

Fig. 3: Frequency distribution of CAI values for all coding genes in *Azotobacter vinelandii* AvOP genome

Table 2: Top 20 PHX genes for the *Azotobacter* AvOP genome

GenBank accession	Description	Gene length (bp)	CAI value
Metabolism			
ZP_00418729	Methionine adenosyltransferase	1191	0.896
ZP_00415164	Aconitase	2604	0.882
ZP_00419981	L-glutamine synthetase	1404	0.877
ZP_00418722	Fructose bisphosphate aldolase	1065	0.873
ZP_00415187	Aconitase	2610	0.870
ZP_00419003	AP endonuclease	858	0.870
ZP_00419361	Molybdenum pterin binding domain	429	0.870
ZP_00418942	2-hydroxy-3-oxopropionate reductase	891	0.869
ZP_00416063	Serine hydroxymethyltransferase	1254	0.862
ZP_00418125	Nucleoside diphosphate kinase	432	0.857
ZP_00416685	Nitrogenase	1428	0.856
ZP_00418692	Malate synthetase	2175	0.855
ZP_00419027	Nickel dependent hydrogenase, large subunit	1809	0.855
ZP_00415335	ATP synthase F1, beta subunit	1377	0.854
Information storage and processing			
ZP_00417428	Translation elongation factor	538	0.878
ZP_00416115	LSU ribosomal protein L25P	597	0.867
ZP_00418502	SSU ribosomal protein S18P	231	0.859
Cellular processes			
ZP_00419975	Alkyl hydroperoxidase reductase/Thiol specific antioxidant/ Mal allergen	564	0.863
ZP_00418004	Translation elongation factor Ts	870	0.857
ZP_00419983	Small GTP binding protein domain: GTP binding protein TypA	1821	0.868

Table 3: List of PHX genes associated with nitrogen fixation in *Azotobacter* AvOP

GenBank accession	Description	Gene length (bp)	CAI value
ZP_00416685	Nitrogenase	1428	0.856
ZP_00416680	Nitrogenase iron protein	873	0.848
ZP_00417578	Nitrogenase molybdenum-iron protein alpha chain	1479	0.838
ZP_00416684	Nitrogenase	342	0.831
ZP_00417577	Nitrogenase iron protein	873	0.829
ZP_00416683	Nitrogenase vanadium-iron protein, alpha chain	1425	0.824
ZP_00417579	Nitrogenase molybdenum-iron protein beta chain	1572	0.820
ZP_00417584	Nitrogenase molybdenum iron cofactor biosynthesis protein	1428	0.790
ZP_00416672	Nitrogenase MoFe biosynthesis protein nif E	1410	0.765
ZP_00419328	Nitrogenase cofactor biosynthesis protein nif B	1512	0.764

functional groups: Information storage and processing consisting of COGs linked to transcription, translation, RNA processing, DNA replication, replication recombination and repair, chromatin structure (group 1); cellular processes including, cell division and cell cycle control, nuclear structure, defense mechanisms, signal transduction, cell wall/envelope biogenesis, cell motility, cytoskeleton, extra cellular structures, intercellular trafficking and posttranslational modification (group 2); metabolism consisting of energy production and conversion, carbohydrate transport, amino acid transport, nucleotide transport, coenzyme metabolism, inorganic ion transport and secondary metabolites biosynthesis (group 3); genes with general function predictions and unknown functions (group 4). CAI values of all the genes present in different COG groups were calculated and the PHX genes were identified as per the cut off values mentioned above. The AvOP genome had 11.47, 14.54, 62.70 and 11.27% PHX genes in the group 1, 2, 3 and 4, respectively. As expected COG functional group 3 (Metabolism) had the sizeable portion of PHX genes for the *Azotobacter* genome.

The top 5 COG categories for *Azotobacter* were: energy production and conversion, amino acid transport and metabolism, carbohydrate transport and metabolism, translation and general function prediction. This presents some insight into the genes required for the lifestyles of the bacterium. The high number of PHX genes in the above mentioned COG specifies their capacity to stay alive in a free-living condition and compete with other microorganisms present in soil. Interestingly about 10 genes linked to nitrogen fixation in this bacterium have been found to fit into the highly expressed category. This is particularly fascinating as they can be good candidates for expression in other organisms. Table 3 shows the PHX genes involved in nitrogen fixation. This incorporates the genes encoding nitrogenase vanadium-iron protein, nitrogenase molybdenum-iron protein alpha chain, nitrogenase molybdenum-iron protein beta chain which play vital roles in the nitrogen fixing machinery of the bacterium. As mentioned earlier the former is synthesized under Mo-deficient conditions in the presence of vanadium and the latter are synthesized in the absence of a fixed nitrogen source when molybdenum is available.

TTA codons in AvOP genes: Like many other G+C rich microorganisms, TTA codon is the rarest one in AvOP genome. TTA codon, corresponding in mRNA to the UUA codon, one of the six alternative leucine codons, has been found to play important role in antibiotic production and aerial mycelia formation in *Streptomyces coelicolor* A3 (2) (Leskiw *et al*., 1991a, b; Li *et al*., 2007). However, no work has been done on TTA containing genes of AvOP. We have identified 686 TTA containing genes which are 13.75% of total protein coding AvOP genes. The expected frequency of TTA codons in the AvOP genome was estimated as 0.35% by multiplying together the overall frequencies of T1 (12.46%), T2 (28.27%) and A3 (9.81%) codons. The observed frequency of TTA codons was 0.076% which is only 21.92% of the expected frequency. However, this is not a very unusual situation as TTA codons in other organisms were also found to be less than expected (6 to 52%, Li *et al*., 2007). The mean GC3%, GC%, CAI and Fop of TTA genes are all less than the average values of protein coding genes where as mean Nc value is more (Table 1). In a nutshell, all these indices indicate that these TTA containing genes are under mutational pressure and less biased in their codon usage. Their expression level is also predicted to be less than the average protein coding genes. Only 6 TTA genes are featured in the PHX gene category which is less than 1% of all TTA containing genes and none of the TTA genes are present in the Top 20 PHX gene list (Table 2). The TTA containing PHX genes are IMP cyclohydrolase gene which is the last enzyme required in IMP biosynthesis pathway and take active part in purine metabolism, one Aldehyde dehydrogenase, one extracellular solute-binding protein, Malate: quinone-oxidoreductase, one methionyl-tRNA synthetase and a Beta-ketoacyl synthase gene.

CONCLUSION

The *Azotobacter* genes show codon bias. The nitrogen fixation related genes and ribosomal protein genes are more biased compared to the protein coding genes. There is considerable amount of heterogeneity among the genes in this bacterium. GC3 composition does not play any role in affecting codon usage variation among the genes in this organism. The gene expression levels are more or less high. The highly expressed genes are affluent in GC rich amino acids. Scattering of the nitrogen fixation related genes along the centre of the axis of correspondence analysis of codon count indicated their conserved nature. Codon usage based strategy was used to approximate the gene expressions in *Azotobacter vinelandii* and identify a set of Potentially Highly Expressed (PHX) genes. We have identified 503 potentially highly expressed genes having diverse functions. Majority of the PHX genes present in the COG categories are associated with metabolic functions. About 10 genes linked to nitrogen fixation are also PHX. These results indicate the ability of the bacterium to persist in a free-living state, compete with other soil bacteria and fix nitrogen in a manner somewhat dissimilar from the conventional method.

ACKNOWLEDGMENTS

Arnab Sen acknowledges DBT for providing Overseas Associateship. The authors are grateful to the Department of Biotechnology (DBT), Government of India, for providing financial help in setting up Bioinformatics Infrastructural Facility at University of North Bengal. This investigation was supported in part by NSF EF-0333177, DBT Overseas Associateship from Govt. of India and by the College of Life Sciences and Agriculture, University of New Hampshire-Durham.

REFERENCES

Banerjee, T., S. Basak, S.K. Gupta and. T.C. Ghosh, 2004. Evolutionary forces in shaping codon and amino acid usages in *Blochmannia floridanus*. J. Biomol. Str. Dyn., 22 (1): 13-23.

Benzecri, J.P., 1992. Correspondence Analysis Handbook. Marcel Dekker, New York.

Bernardi, G., 1995. The human genome: Organization and evolutionary history. Ann. Rev. Genet., 29: 445-476.

Bishop, P.E. and R. Premakumar, 1992. Alternative Nitrogen Fixation Systems. In: Biological Nitrogen Fixation, Stacey, G., R.H. Burris and H.J. Evans (Eds.). Chapman and Hall, New York, pp: 737-762.

Cutter, A.D., B.A. Payseur, T. Salcedo, A.M. Estes, J.M. Good, E. Wood, T. Hartl, H. Maughan, J. Strempel, B. Wang, A.C. Bryan and M. Dellos, 2003. Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. Genome Res., 13 (12): 2651-2657.

Dixon, R. and D. Kahn, 2004. Genetic regulation of biological nitrogen fixation. Nat. Rev. Microbiol., 2 (8): 621-631.

Dos, Reis, M., L. Wernisch and R. Savva, 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res., 31 (23): 6976-6985.

Eady, R.R., 1991. The Mo-, V- and Fe-based nitrogenase systems of *Azotobacter*. Adv. Inorg. Chem., 36: 77-102.

Eady, R.R., 1996. Structure minus sign function relationships of alternative nitrogenases. Chem. Rev., 96 (7): 3013-3030.

Ghosh, T.C., S.K. Gupta and S. Majumdar, 2000. Studies on codon usage in *Entamoeba histolytica*. Int. J. Parasitol., 30 (1): 715-722.

Grantham, R., C. Gautier, M. Gouy, M. Jacobzone and R. Mercier, 1981. Codon catalogue usage is a genome strategy for genome expressivity. Nucleic Acids Res., 9 (1): 213.

Gupta, S.K., T.K. Bhattacharya and T.C. Ghosh, 2004. Synonymous codon usage in *Lactococcus lactis*: Mutational bias versus translational selection. J. Biomol. Str. Dyn., 21 (4): 1-9.

Ikemura, T., 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol., 146 (1): 1-21.

Lafay, B., J.C. Atherton and P.M. Sharp, 2000. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology, 146 (4): 851-860.

Lawson, D.M. and B.E. Smith, 2002. Metal Ions in Biological Systems. Sigel, A. and H. Sigel (Eds.). Vol. 39, Marcel Dekker, New York, pp: 75-119.

Leskiw, B.K., E.J. Lawlor, J.M. Fernandez-Abalos and K.F. Chater, 1991a. TTA codons in some genes prevent their expression in a class of developmental, antibiotic-negative, *Streptomyces* mutants. Proc. Natl. Acad. Sci., 88 (6): 2461-2465.

Leskiw, B.K., M.J. Bibb and K.F. Chater, 1991b. The use of a rare codon specifically during development? Mol. Microbiol., 5 (12): 2861-2867.

Li, W., W. Wu, W. Tao, C. Zhao, Y. Wang, X. He, G. Chandra, X. Zhou, Z. Deng, K. Chater and M. Tao, 2007. A genetic and bioinformatics analysis of *Streptomyces coelicolor* genes containing TTA codons, possible targets for regulation by a developmentally significant tRNA. FEMS Microbiol. Lett., 266 (1): 20-28.

Martin-Galiano, A.J., J.M. Wells and A.G. de la Campa, 2004. Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. Microbiology, 150 (7): 2313-2325.

Peden, J., 1999. Analysis of codon usage. Ph.D Thesis, The University of Nottingham, United Kingdom.

Rediers, H., P.B. Rainey, J. Vanderleyden and R. De Mot, 2005. Unraveling the secret lives of bacteria: Use of *in vivo* expression technology and differential fluorescence induction promoter traps as tools for exploring niche-specific gene expression. Microbiol. Mol. Rev., 69 (2): 217-261.

Rees, D.C. and J.B. Howard, 2000. Nitrogenase: Standing at the crossroads. Curr. Opin. Chem. Biol., 4 (5): 559-566.

Sen, G., S. Sur, D. Bose, U. Mondal, T. Furnholm, A.K. Bothra, L. Tisa and A. Sen, 2007. Analysis of codon usage patterns and predicted highly expressed genes for six phytopathogenic *Xanthomonas* genomes shows a high degree of conservation. In: Silico Biology 7: 0039 http: //www.bioinfo.de/isb/ 2007/07/0039/

Sharp, P.M. and W.H. Li, 1987. The Codon Adaptation Index-a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res., 15 (3): 1281-1295.

Stenico, M., A.T. Lloyd and P.M. Sharp, 1994. Codon usage in *Caenorhabditis elegans*- delineation of translational selection and mutational biases. Nucleic Acids. Res., 22 (13): 2437-2446.

Sur, S., A. Sen and A.K. Bothra, 2006. Codon usage profiling and analysis of intergenic association of *Frankia* EuIK1 nif genes. Ind. J. Microbiol., 46 (4): 363-369.

Sur, S., A. Sen and A.K. Bothra, 2007. Mutational drift prevails over translational efficiency in *Frankia* nif operons. Ind. J. Biotechnol., 6 (3): 321-328.

Tatusov, R.L., N.D. Federova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumdar, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin and D.A. Natale, 2003. The COG database: An updated version includes eukaryotes. BMC Bioinformatics, 4 (1): 41.

Walpole, R.E., R.H. Myers, S.L. Meyers and K. Ye, 2004. Probability and Statistics for Engineers and Scientists, Pearson Education, Singapore, Pte. Ltd., Indian Branch, 482 F.I.E. Patparganj, Delhi 110 092, India.

Wu, G., D.E. Culley and W. Zhang, 2005a. Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiology, 151 (7): 2175-2187.

Wu, G., L. Nie and W. Zhang, 2005b. Predicted highly expressed genes in *Nocardia farcinica* and the implication for its primary metabolism and nocardial virulence. Anton. Van. Leewen, 89 (1): 135-146.