Research Journal of Microbiology

ISSN 1816-4935

Academic Journals Inc.

www.academicjournals.com

Research Journal of Microbiology 3 (10): 600-613, 2008
ISSN 1816-4935

In silico Analysis of *Chlorobium* Genomes Divulge Insights into the Lifestyle of the Bacteria

[1]Saubashya Sur, [2]Asim K Bothra, [1]Manprit Bajwa, [3]Louis S. Tisa and [1]Arnab Sen
[1]NBU Bioinformatics Facility, Department of Botany,
University of North Bengal, Siliguri-734013, West Bengal, India
[2]Chemiinformatics Bioinformatics Laboratory, Department of Chemistry,
Raiganj University College, Raiganj-733134, West Bengal, India
[3]Department of Microbiology, University of New Hampshire,
Rudman Hall, 46 College Road, Durham NH 03824, USA

Abstract: The finished sequences of three *Chlorobium* genomes were examined and compared to each other for their synonymous codon usage. Codon usage by *Chlorobium* was moderately biased but a considerable amount of variation was observed. GC3 composition plays an important role in the codon usage variation among the genes in the studied genomes. Similar homologs of horizontally transferred nitrogen fixing and photosynthesis related genes having high identity levels indicated their co-evolution within the genus. Correlation of codon usage bias with tRNA content in *Chlorobium* genomes revealed the inability of the translation machinery in these bacteria to co-evolve with higher codon usage resulting in moderate bias. Arrangement of the genes in leading strand and lagging strand of replication had virtually no role in influencing synonymous codon usage variation in these bacteria. Whole genome alignment revealed the conserved nature of the genomes. Using codon adaptation index, a set of potentially highly expressed genes in *Chlorobium* was determined taking ribosomal protein genes as a reference. A sizeable fraction of the potentially highly expressed (PHX) genes in the COG categories were related to metabolism. Quite fascinatingly, some of the genes associated with nitrogen fixation and photosynthesis like hydrogenases, nitrogenase iron protein complexes, bacteriochlorophylls, chlorosomes etc. were also PHX. These results offer insights into the survival patterns of these bacteria thriving under stressed conditions and efficiently carrying out two important metabolic processes especially under reduced light and anoxic environments.

Key words: *Chlorobium*, codon bias, potentially highly expressed (PHX) genes, nitrogen fixation and photosynthesis

INTRODUCTION

The phylum *Chlorobi* contains a unique group of photosynthetic bacteria having narrow diversity commonly referred to as green sulfur bacteria (Overmann, 2000; Garrity and Holt, 2001). The green sulfur bacteria frequently inhabit illuminated, stratified anoxic aquatic habitats, sediments, sulfide-rich environments, hot springs etc. (Van Gemerden and Mas, 1995; Imhoff, 2003). Majority of species under the genus *Chlorobium* are rod shaped anaerobic photolithoautotrophs. They are photosynthetic and can fix nitrogen. They use hydrogen sulfide, elemental sulfur and hydrogen as electron sources (Imhoff, 2003). The genome sequences for some *Chlorobium* strains are available in the public domain. Amongst them only three (*Chlorobium tepidium* TLS, *Chlorobium phaeobacteroides* DSM 266 and

Corresponding Author: Arnab Sen, NBU Bioinformatics Facility, Department of Botany,
University of North Bengal, Siliguri-734013, West Bengal, India Tel: +91-0353-6528172

Chlorobium chlorochromatii CaD3) are complete and the rest draft. The availability of the complete genomes unlocks the opportunity to use bioinformatics approaches to investigate synonymous codon usage patterns, gene expression levels and throw light on their lifestyles. Since, draft genome sequences contain annotation errors and are subject to change we did not considered them for analysis. These *Chlorobium* strains are gram negative nitrogen fixing anaerobes, photooxidising sulfur compounds and possessing special light-harvesting complexes called chlorosomes that contain bacteriochlorophylls and carotenoids (Eisen *et al*., 2002). *Chlorobium tepidium* is thermophilic, gram-negative bacterium originally isolated from sulfide rich hotsprings in New Zealand (Eisen *et al*., 2002). *Chlorobium phaeobacteroides* is a non-motile, rod-shaped bacterium first isolated from the anoxic sulfide containing waters, 19.5 m below the surface of meromictic Lake Blankvann in Norway. It is a representative of the brown-colored green sulfur bacteria species (Imhoff, 2003). It lacks vesicles and grows in fresh water medium. *Chlorobium chlorochromatii* is phototrophic green sulfur bacteria found in freshwater lakes worldwide (Overmann and Schubert, 2002). It is a symbiotic green sulfur bacterium existing as phototrophic consortia. Phototrophic consortia are examples of one of the best known types of bacterial associations (Vogl *et al*., 2006).

Synonymous codon usage is non-random and species specific (Banerjee *et al*., 2004) and vary noticeably among different genes within the same genus (Basak *et al*., 2004). Across the genome, the G+C composition ensuing from mutational bias and/or translational selection has been speculated to decide the major trends in codon usage by high or low G+C organisms (Knight *et al*., 2001). In a given genome, codon bias is greater in highly expressed genes compared to lowly expressed genes (Lafay *et al*., 2000; Dos Reis *et al*., 2003). Codon bias of highly expressed genes is more affected by translational selection compared to lowly expressed genes, which are directed by mutational bias (Banerjee *et al*., 2004). To analyze the patterns of codon usage, various indices have been proposed to gauge the degree and direction of codon bias (Sur *et al*., 2008). Amongst them, the Codon Adaptation Index (CAI) was proposed as a quantum of codon usage within a gene relative to a reference set of genes (usually ribosomal protein genes) (Sur *et al*., 2008). This index has been shown to be associated with mRNA expression levels (Sen *et al*., 2007). Besides CAI (Wu *et al*., 2005a), the effective number of codons (Nc), described as the number of equal codons creating the same codon usage bias as observed and the frequency of optimal codons (Fop) (Sur *et al*., 2008), defined as the portion of synonymous codons that are optimal, are also used for the same purpose.

The microbiological, physiological and biochemical characteristics of *Chlorobium* have already generated lot of interest among the microbiologists. The present study aims to perform a comparative analysis of the codon usage patterns and predicted expression levels for the protein coding genes in these bacteria with special reference to genes associated with nitrogen fixation and photosynthesis. The outcome of this study will sufficiently throw light on their lifestyle and subsistence in their habitat.

MATERIALS AND METHODS

The finished genome sequences for three *Chlorobium* strains (*Chlorobium tepidium* TLS, *Chlorobium phaeobacteroides* DSM 266 and *Chlorobium chlorochromatii* CaD3), (GenBank accession numbers NC 002932, NC 008639 and NC 007514; hence forth to be referred as CTE, CPB and CCM respectively) were obtained from the IMG website (img.jgi.doe.gov) (Markowitz *et al*., 2006). The data of horizontally transferred genes for CCM and CTE were obtained from the website (http://cbcsrv.watson.ibm.com/HGT_SVM/) (Tsirigos and Rigoutsos, 2005). CTE, CCM and CPB have 2308, 2050 and 2799 genes in total. All of the protein coding genes and those associated with the ribosomal proteins, photosynthesis and nitrogen fixation were explored using Codon W software (http://bioweb2.pasteur.fr) (Sen *et al*., 2008), CAI Calculator 2 (http://www.evolvingcode.net/codon/CalculateCAIs.php) (Wu *et al*., 2005a, b) and ACUA (Umashankar *et al*., 2007).

The software Codon W was used to analyze G or C in the third position of codons (GC3s), effective number of codons (Nc) (Ghosh *et al*., 2004) and the frequency of optimal codons (Fop) (Sur *et al*., 2008). The effective number of codons (Nc) is a simple gauge of codon bias (Ghosh *et al*., 2004). Nc values range from 20 (when only one codon is per amino acid) to 61(when all codons are used in equal probability). Under random codon usage the expected value of Nc was calculated by the following formula:

$$Nc = 2 + S + \{29/ [S^2 + (1-S)^2]\} \quad (1)$$

where, S denotes GC3 values.

Fop is the portion of synonymous codons that are optimal. Its value ranges form 0 (implying a gene has no optimal codons) and 1.0 (when a gene is wholly composed of optimal codons).

The codon adaptation index (CAI) (Wu *et al*., 2005a) was determined using a web-based application: The CAI Calculator 2 (http://www.evolvingcode.net/codon/cai/cais.php) (Wu *et al*., 2005a, b) taking the ribosomal genes as a reference. It is a quantum of relative adaptiveness of a gene's codon usage towards the codon usage of highly expressed genes. The relative adaptiveness of each codon is the ratio of the usage of each codon, to that of the most abundant codon within the same synonymous family. The CAI value vary from 0 to 1.0 with higher CAI values indicating that the gene of interest has a codon usage pattern more similar to that in the reference genes.

In order to detect whether the values for the abovementioned indices in nitrogen fixing genes, ribosomal protein genes and photosynthesis related genes appreciably vary from the protein coding genes Z test was executed (Sur *et al*., 2008).

A study of the horizontally transferred photosynthesis related genes and nitrogen fixing genes in CCM and CTE were done to detect whether they are present in all the strains or native to a particular strain. First of all the nitrogen fixing genes and photosynthesis related genes acquired by horizontal gene transfer mechanisms in CCM and CTE were sorted out. Using the Integrated Microbial Genomes database (www.img.jgi.doe.gov) (Markowitz *et al*., 2006), the sorted genes for each strain were subjected to IMG Genome BLAST against the studied strains to find out the sequence homologs. The minimum percent identity was set at different levels like 90, 80 and 70% and the maximum E value 1e-2.

AT and GC skewness was calculated using ACUA (Umashankar *et al*., 2007). Correspondence analysis (COA) on codon count was performed using Codon W (http://bioweb2.pasteur.fr). This method explores the major trends in codon and amino acid variations among the genes.

The three *Chlorobium* genomes were subjected to whole genome alignment using MAUVE (Darling *et al*., 2004) to find out the degree of conservation among them.

The research was started in the fall of 2007. The research was virtually done in three laboratories. Studies of codon usage patterns and expression level prediction of genes were done at Bioinformatics Facility, NBU while we executed studies of horizontally transferred genes at Raiganj college and whole genome alignment at Department of Microbiology, UNH.

RESULTS AND DISCUSSION

Codon Usage Patterns for Three *Chlorobium* genomes

The primary aim in this study on the codon usage patterns among the three *Chlorobium* genomes was to estimate the level of heterogeneity in codon use. A good number of bacteria with a reasonable AT/GC content have a significant amount of codon heterogeneity (Sen *et al*., 2008). Codon heterogeneity is normally coupled with gene expression level. Hence, highly expressed genes contain higher frequencies of codons that are translationally optimal (Lafay *et al*., 2000). The GC3s and Nc values for all the genes in these genomes were calculated to determine if codon heterogeneity exists among genes of the three *Chlorobium* genomes. The results from this analysis are shown in Fig. 1.

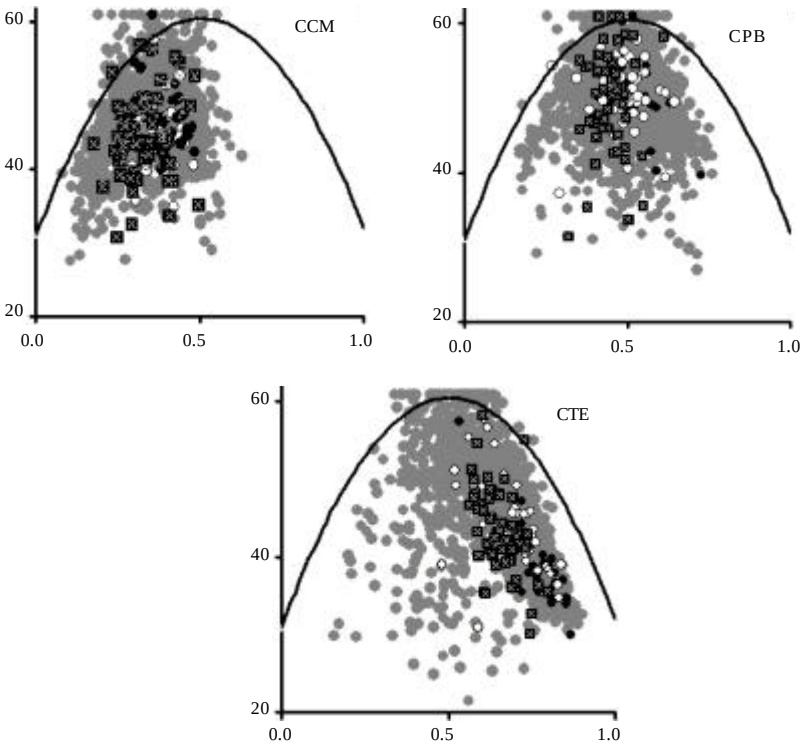

Fig. 1: Effective number of codons (Nc) (Axis Y) plotted against the GC content at the synonymous third position (Axis X) in the *Chlorobium* genomes. The continuous curve symbolizes the null hypothesis that the GC bias at the synonymous site is solely due to mutation, but not selection. The protein coding genes are represented by gray circles, ribosomal protein genes in black squares with cross, nitrogen fixing genes in black circles and photosynthesis related genes in white circles

These Nc vs. GC3 plots have been recommended to be an effective means to detect the codon usage variations among the genes in the same genome. The Nc values range from 21±1 to 61±0 in CTE and 27±2 to 61±0 in CPB and CCM demonstrating heterogeneity of codon usage existing in these genomes as anticipated. Ribosomal protein genes, known to be highly expressed during rapid cell growth, were recognized and shown in the Nc vs GC3 plots. Most of the ribosomal protein genes are loosely clustered in the *Chlorobium* genomes and remain below the expected curve. This is unlike the results obtained for *Streptomyces* (Wu *et al*., 2005a), *Xanthomonas*, *Frankia* (Sen *et al*., 2007, 2008) and *Azotobacter* (Sur *et al*., 2008) where the ribosomal protein genes are strongly clustered at low ends of the plot. This indicates lower codon bias in these genes. The location of the genes associated with nitrogen fixation and photosynthesis, are also shown in the Nc plots (Fig. 1). The photosynthetic genes remain clustered more or less strongly in all the *Chlorobium* genomes. Contrary to CPB and CTE genomes, nitrogen-fixing genes in CCM remain scattered. In cases were synonymous codon bias is completely dictated by GC3s, Nc values fall on the expected curve in the Nc vs GC3 plot (Ghosh *et al*., 2004). However, in the *Chlorobium* genomes it is seen that barring a few, majority of the genes were below the expected curve (Fig. 1). This is a reflection of the fact that synonymous codon bias in majority of the genes in *Chlorobium* is quite independent of the overall base compositions.

Table 1: Mean values± Standard deviation of Nc, GC, GC3, Fop and CAI indices for several gene groups in the three *Chlorobium* genomes

Strain	Gene	Mean Nc	Mean GC (%)	Mean GC3 (%)	Mean fop	Mean CAI
CCM	PCG	46.76±4.78	44.3±0.05	37.1±0.08	0.400±0.05	0.642±0.04
	NFG	46.44±4.44	46.4±0.02	40.3±0.05	0.432±0.05	0.645±0.03
	RPG	44.46±5.88	42.6±0.02	32.0±0.07	0.476±0.06	0.700±0.04
	PG	44.86±4.37	45.8±0.02	37.6±0.04	0.435±0.06	0.662±0.04
	HTG	48.82±5.60	43.3±0.05	35.1±0.08	0.396±0.06	0.627±0.05
CTE	PCG	44.63±7.00	55.7±0.06	67.2±0.12	0.518±0.06	0.648±0.09
	NFG	38.74±5.04	59.4±0.03	77.4±0.06	0.579±0.05	0.736±0.06
	RPG	43.23±5.83	53.7±0.03	65.7±0.05	0.541±0.04	0.720±0.05
	PG	44.15±6.00	56.4±0.03	68.9±0.09	0.553±0.05	0.689±0.06
	HTG	51.18±6.59	48.0±0.06	51.5±0.10	0.454±0.06	0.539±0.08
CPB	PCG	51.66±4.99	48.7±0.04	49.0±0.08	0.448±0.05	0.709±0.03
	NFG	50.75±5.01	50.4±0.02	51.5±0.06	0.477±0.06	0.739±0.02
	RPG	50.15±6.98	46.3±0.03	44.7±0.05	0.460±0.06	0.760±0.02
	PG	50.39±4.57	48.9±0.03	49.0±0.07	0.481±0.05	0.729±0.03

PCG: Protein Coding Genes, NFG: Nitrogen Fixing genes, RPG: Ribosomal Protein Genes, PG: Photosynthesis related genes, HTG: Horizontally Transferred Genes

Table 1 shows the mean values of different indices used to study codon usage patterns. The ribosomal protein genes, nitrogen fixing genes and photosynthesis related genes had a lower mean Nc value than the protein coding genes. Horizontally Transferred Genes (HTG) for CCM and CTE had high Nc values. This was expected since they were acquired by horizontal gene transfer mechanisms during evolution. Majority of horizontally transferred genes in CTE and CCM included, hypothetical proteins, replication and repair proteins, restriction system specificity proteins some transcriptional elongation factors, reductases, NADH dehydrogenases, chlorosome envelope proteins etc. Since, no data were obtained for horizontally transferred genes in CPB; HTGs were not tabulated in CPB. Cameron and Aguade (1998) pointed out that, genes having Nc values less than 40 have a stronger codon bias controlled by mutational pressure. In the *Chlorobium* genomes, barring the NFGs of CTE none of the other categories of genes had mean Nc values less than 40, indicating that translational selection may be acting. High effective number of codons indicated low codon bias. The ribosomal protein genes, nitrogen fixing genes and photosynthesis related genes had higher Fop values compared to the protein coding genes. The elevated Fop values signifying the presence of higher fraction of optimal codons in these genes. If mutational bias exclusively controlled codon usage, these genes would have had a low Fop value. As that was not the case for these *Chlorobium* genomes, there may be other factors acting on codon bias. The values of the nitrogen fixing genes, ribosomal protein genes and photosynthesis related genes were tested for any significant difference with that of the protein coding genes. The z values for NFGs and PGs revealed minor differences from PCGs for GC3 and GC in case of CPB. The Z values of Nc, GC3 and CAI for RPGs, NFGs and PGs for CTE varied from PCGs. These observations entail that some level of disparity exists in the characteristics of the genes even though they belong to the same genome. However, no significant variations were noticed in CCM.

Rocha (2004) correlated codon usage bias with tRNA content in bacterial genomes. We obtained the optimal generation times of the three *Chlorobium* genomes from personal communications with researchers working in the field of *Chlorobium* biology and investigated whether codon bias could be an important factor in these bacteria. *Chlorobium phaeobacteroides* DSM 266 and *Chlorobium chlorochromatii* CaD3 had optimal generation times of 0.5-1 h while *Chlorobium tepidium* TLS had optimal generation time below 2 h. On the basis of Rocha's (2004) observations, *Chlorobium* genomes could be regarded as fast growers. He illustrated that fast growers have a median of 61 tRNA genes compared to 44 for slow growers and the former tend to have stronger codon usage bias contrary to the latter. However, CPB, CTE and CCM had 47, 50 and 45 tRNA genes respectively. This is unlike *Synecococcus*, *Desulfovibrio* etc. surviving in stratified aquatic environments, which followed Rocha's (2004) observations. The studied *Chlorobium* strains had 40 unique anticodon tRNA genes i.e., they

have less diverse tRNAs. So, these strains are specialized to use a small set of anticodons in spite of maintaining a moderate number of tRNAs. The ribosomal protein genes of these *Chlorobium* strains, which are highly expressed, showed moderate codon bias. The translation machinery of *Chlorobium* probably did not co-evolve with higher codon usage in highly expressed genes even though it is a fast growing bacterium. This may be one reason why *Chlorobium* genes did not show strong codon usage bias.

Analysis of the Horizontally Transferred Nitrogen Fixing and Photosynthesis Related Genes

The *Chlorobium* strains contained 242 and 297 horizontally transferred genes for CCM and CTE, respectively. Among these the number of photosynthesis related genes and those related to nitrogen fixation were 4 and 3 for CCM; 7 and 2 for CTE. So it is seen that very few genes associated with these processes were acquired by horizontal gene transfer mechanisms. Most of them were hydrogenases, ferredoxin, chlorosome enveolope proteins and bacteriochlorophylls. IMG genome BLAST results revealed homologs having sequence identity with few similar proteins in other strains. In CCM the ferredoxin gene showed 71.94% sequence homology with similar protein in CPB. The other two horizontally transferred nitrogen fixing genes showed with different gene homologs (percent identities ranging from 70-72). The photosynthesis related genes like bacteriochlorophyll A protein, chlorosome enveolope protein A, cytochrome b6-f complex and cytochrome cbb3 oxidase found 5 homologs (percent identities ranging from 71.94-96.25%) in CTE and CPB. In CTE the two horizontally transferred nitrogen fixing genes found 2 homologs (percent identities ranging from 70-74%) in CPB. The photosynthesis related genes like bacteriochlorophyll A protein, photosystem P840 reaction centre cytochrome c-551, chlorosome enveolope protein A and cytochrome c biogenesis protein found 5 homologs (percent identities ranging from 67.6-96.25%) in CCM and CPB. From the analysis we found that only few homologs were acquired by horizontal gene transfer mechanisms. This indicated that the rest were native to those strains and warded off the selective pressures of nature. The horizontally transferred homologs on the other hand were gained from other organisms and the high percent identity within the strains indicated that these genes evolved together within the genus.

Correspondence Analysis

Multivariate statistical analysis is an extensively used method to study the dissimilarity in codon usage among the genes in different organisms (Ghosh *et al*., 2000). Synonymous codon usage by nature is multivariate, hence it is indispensable to analyse this data with multivariate statistical techniques. Correspondence analysis is one of the important multivariate techniques. It is an ordination method that recognizes the major trends in the variation of the data and distributes genes along continuous axes in harmony with these inclinations (Basak *et al*., 2004). Correspondence analysis was performed on simple codon count. Figure 2 shows the positions of the genes on the first and second major axis. We accounted on the first two axes, since subsequent axes yielded little information. Although we recognized genes on the leading strand and lagging strands, there was a rather large overlap between the two clouds of genes in CTE and CCM, while in CPB the genes in the lagging stand remain dispersed. The positions of the genes in the leading strands as well as lagging strands of the first major axis of variation had very low correlations with GC skew and AT skew in all the three genomes. These results suggest that leading or lagging strand did not have any effect on codon usage variations among the genes in the three *Chlorobium* genomes. However, our analysis revealed some interesting trends. The position of the genes on the first major axis of variation showed strong positive correlation with GC3 ($r = 0.958$, $r = 0.775$; $p<0.001$) for CTE and CCM and a strong negative correlation with GC3 ($r = -0.801$; $p<0.001$) for CPB. This suggests that variations in synonymous GC3 composition play a significant role in codon usage variations among the genes in these genomes and strongly expressed genes have higher GC content at the synonymous third positions. The positions of the genes on the

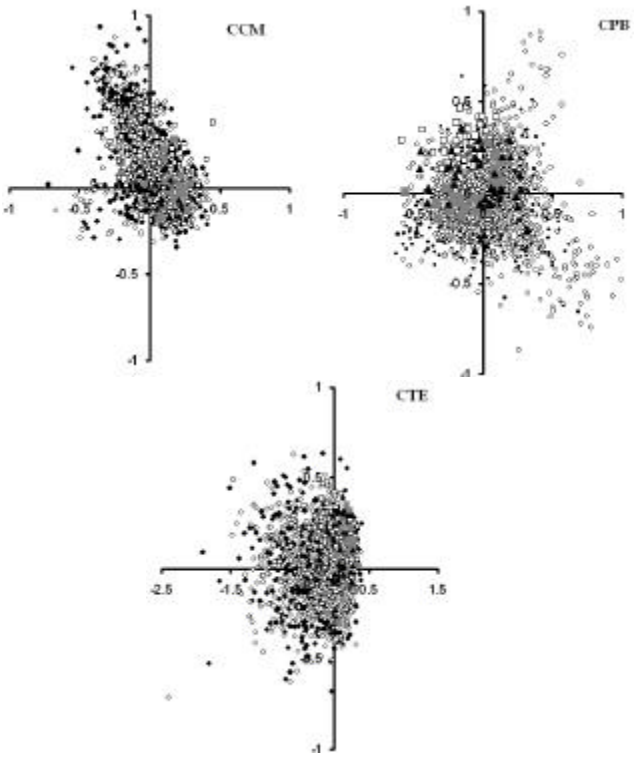

Fig. 2: Correspondence analyses on codon count for the *Chlorobium* genomes. For each plot, the X and Y-axis correspond to axis 1 and 2 of the analysis. Genes in leading strands are represented by white circles, lagging strands by black circles; ribosomal protein genes in white squares, nitrogen-fixing genes in grey squares and photosynthesis related genes in black triangles

first major axis showed positive correlations with G3 ($r = 0.476$, $r = 0.582$; $p<0.001$ for CTE and CCM) and C3 ($r = 0.643$, $r = 0.502$; $p<0.001$ for CTE and CCM) and significant negative correlations with A3 ($r = -0.771$, $r = -0.621$; $p<0.001$ for CTE and CCM) and T3 ($r = -0.835$, $r = -0.464$; $p<0.001$ for CTE and CCM). In case of CPB, the positions of the genes on the first major axis of variation revealed strong negative correlations with C3 and T3 ($r = -0.926$, $r= 0.825$; $p<0.001$) and insignificant correlations with G3 and A3. Unlike in CPB and CCM, we found significant positive correlation of the first major axis of variation with CAI values in CTE ($r = 0.897$, $p>0.001$). This result showed that expression levels of the genes had a role to play in dictating codon usage variation among the genes in CTE. Nitrogen fixing genes, ribosomal protein genes and photosynthesis related genes remain clustered in the core region in CTE and CCM and roughly scattered in CPB. The genes which are situated away from the centre of the axes included a large number of hypothetical protein genes, cold shock proteins, heat shock proteins; translation elongation factor Tu, etc.

Conserved Nature of the *Chlorobium* Genomes

Whole genome alignment of three *Chlorobium* genomes resulted in a global alignment of the locally collinear blocks (LCBs) having sequence elements conserved among the genomes under study. The alignment identified 850 LCBs with a minimum weight of 16. Each LCB represented a homologous region shared by the *Chlorobium* genomes devoid of any rearrangements of homologous sequence. These 850 LCBs contained sequence elements that are conserved among the *Chlorobium* genomes. A

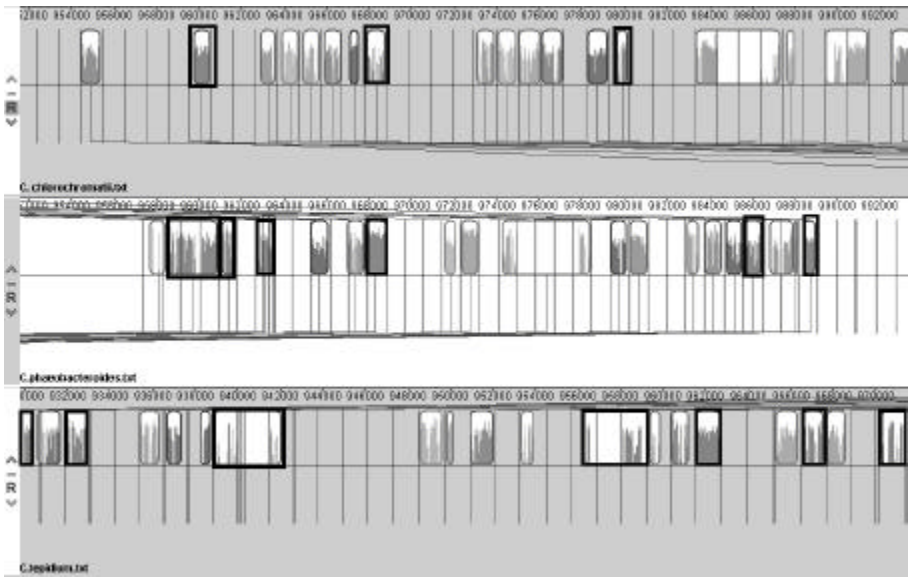

Fig. 3: MAUVE visualization of locally collinear blocks (LCBs) identified between the three *Chlorobium* genomes. Homologous segments are shown in colored blocks connected across the genomes. Conserved areas among all the three genomes are marked by dark blocks. Lines between genomes mark out the orthologous LCB through every genome

large number of regions were identified that were conserved among all the three *Chlorobium* genomes. They are represented by dark shaded blocks in Fig. 3. The regions that were conserved only among the subsets of the studied genomes had different color. We obtained four inverted regions in CTE that were in reverse complement orientation relative to the first sequence i.e., CCM.

Identification of Potentially Highly Expressed Genes in *Chlorobium*

The codon adaptation index is a gauge of directional codon bias. The index uses a reference set of highly expressed genes from a species to evaluate the relative qualities of each codon and a score for a gene is determined from the frequency of use of all codons in that gene. In that respect, CAI values have been a quantitative indicator for gene expression levels (Sen *et al*., 2007). Studies on *Streptomyces coelicolor* and *Streptomyces avermitilis* (Wu *et al*., 2005a) confirmed the interrelationships between CAI values and expression levels and were subsequently authenticated experimentally that CAI predicted potentially highly expressed genes certainly are highly expressed. CAI values for these *Chlorobium* genomes were checked to recognize the potentially highly expressed genes.

Figure 4 shows the distribution of CAI values for all the genes in the three *Chlorobium* genomes. These CAI values ranged from 0.561-0.844, 0.274-0.846 and 0.460-0.786 for CPB, CTE and CCM, respectively. The median CAI values for the genes were 0.712, 0.664 and 0.641, while mode CAI values were 0.699, 0.657 and 0.670 for CPB, CTE and CCM respectively. The plot of the frequency distribution of CAI values for the three *Chlorobium* genomes exposed different distribution patterns (Fig. 4). The pattern for CPB had a peak in the 0.70-0.75 CAI range that rose and fell abruptly, whereas the distribution of CTE and CCM CAI values had a peak in the range of 0.65-0.70 and 0.60-0.65, respectively. CAI values for CTE were broadly dispersed whereas that for CCM rose and fell progressively. CTE CAI values had a lower peak value (24.42%) compared to CPB and CCM (50.93 and 40.63% correspondingly). The CAI distribution patterns for CPB and CCM showed

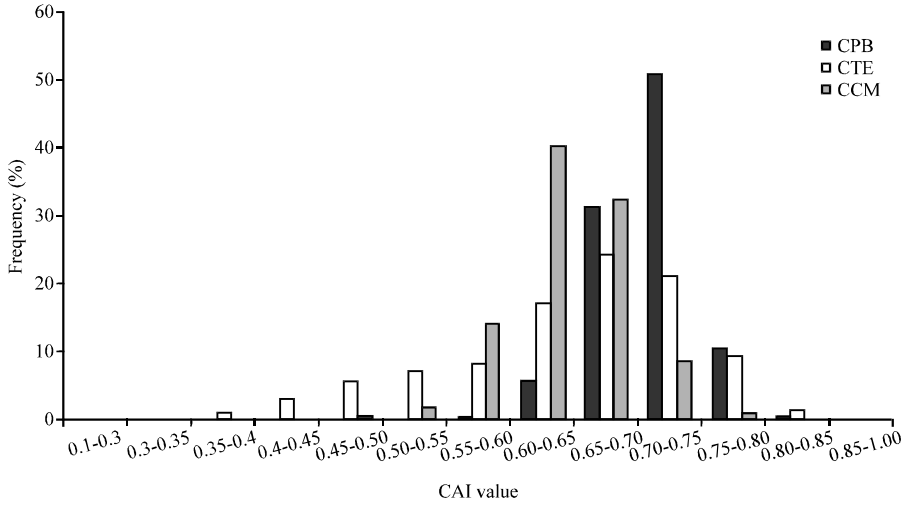

Fig. 4: Frequency distribution of CAI values for all coding genes in *Chlorobium* genomes

inclination towards higher CAI values. These results advocate that CPB and CCM have more highly expressed genes in their genomes in contrast to CTE. However, the CCM genome had lower number of genes compared to the genomes of CPB and CTE. Vogl (2006) reported that CCM existed in phototropic consortia and studies relating to *in silico* subtractive hybridization identified 188 ORFs unique for CCM most of them coding for hypothetical proteins without any homology to sequences in free living green-sulfur bacteria. He further stated that this number is quite less compared to other niche-specific low and high light adapted strains and adaptation to the association in phototropic consortia did not require large number of additional genes. This may be the reason behind the distribution of CAI values of CCM in Fig. 4. The adaptation of CPB to extremely low light and anoxic sulfur containing habitat probably needs higher metabolic rate and increased energy transfer efficiency and thus the inclination of its genes towards higher CAI values.

Wu *et al*. (2005a) portrayed that the top 10% of the genes, in terms of CAI values, were the predicted highly expressed genes (PHX) and corresponded to CAI cutoffs of 0.754, 0.755 and 0.700 for CPB, CTE and CCM, respectively. The CPB genome had 258 PHX genes including 32 RPGs, 5 NFGs and 12 PGs while the CTE genome had 227 PHX genes with 14 RPGs, 15 NFGs and 6 PGs and the CCM genome had 201 PHX genes with 27 RPGs, 3 NFGs and 7 PGs. The top 20 PHX genes are shown in a separate table as supplementary material.

Functional Analysis of the PHX Genes

In order to comprehend the functional distribution of the PHX genes among the three *Chlorobium* genomes the Clusters of Orthologous Groups of proteins were studied. The COG types consist of proteins or groups of paralogs from 3 lineages and match up with ancient phylogenetic lineages. For these *Chlorobium* genomes, 21 COG categories were recognized. Figure 5 shows the allotment of the PHX into each COG category based on the total PHX genes and the total genes within that COG group. To assist the investigation, each of the COG categories were horded in the following 4 COG groups: Information and storage processing consisting of COGs related to transcriptions, translation, RNA processing, DNA replication recombination and repair (group 1); Cellular processess comprising of COGs related to cell division, defense mechanism, signal transduction, cell envelope biogenesis, cell motility, intracellular structures and post translational modification (group 2); Metabolism including energy production and conversion, carbohydrate transport, amino acid transport, lipid transport and

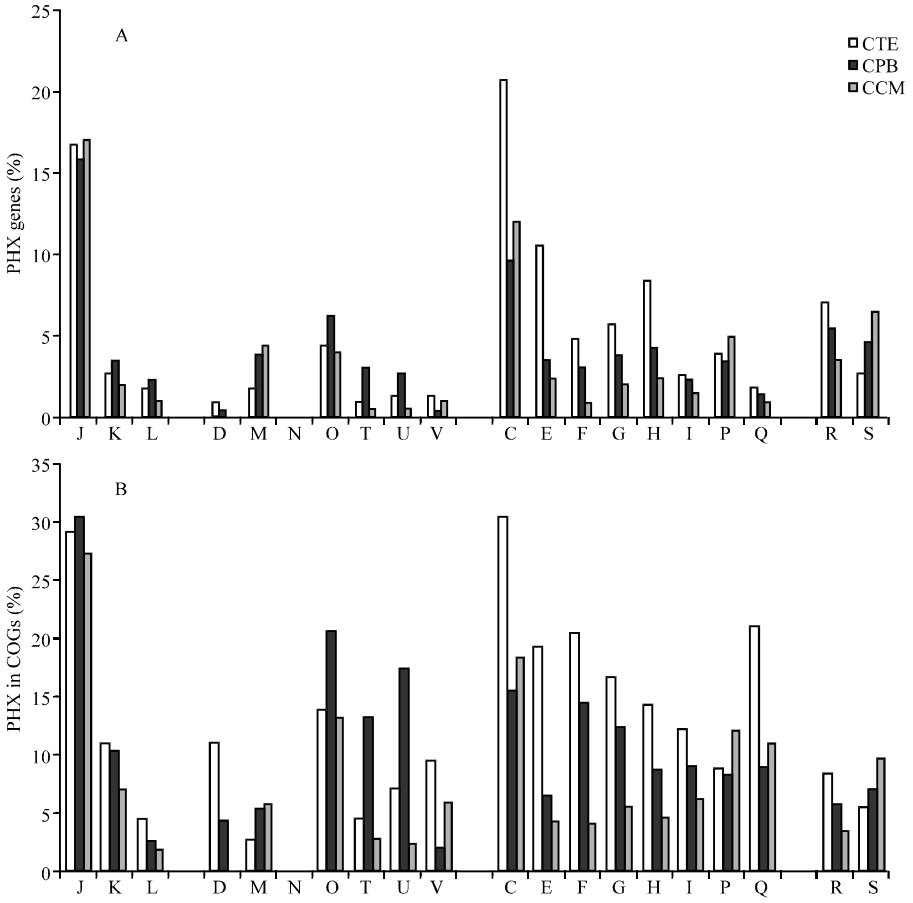

Fig. 5: Allocation of PHX genes in various COG functional groups of *Chlorobium* genomes. COG functional groups are the following: (1) Information and storage processing: J-Translation, ribosomal structure and biogenesis; K-Transcription; L-DNA replication, recombination and repair, (2) Cellular processes: V-Defense mechanisms; T-Signal transduction; M-Cell enveolope biogenesis, outer membrane; N-Cell motility and secretion; U-Intracellular trafficking, secretion and vesicular transport and O-Post translational modification, protein turnover and chaperones. (3) Metabolism: C-Energy production and conversion; G-Carbohydrate transport and metabolism; E-Amino acid transport and metabolism; L-Lipid metabolism and Q-Secondary metabolite biosynthesis, transport and catabolism and (4) Poorly characterized: R-General function and S-Unknown function

metabolism, nucleotide transport, coenzyme metabolism, inorganic ion transport and secondary metabolite biosynthesis (group 3); General function prediction and unknown function (group 4). CAI values of all the genes present in different COG groups were determined and the PHX genes were recognized as per the cut off values of various *Chlorobium* genomes. Figure 5a illustrates the percentage of PHX genes in different COG categories grouped in the four COG functional groups. The *Chlorobium* genomes had the following distribution of PHX in the COG functional groups: CPB 27.18, 20.87, 39.32 and 12.62%; CTE 21.14, 10.57, 58.59 and 9.69%; CCM 29.41, 15.44, 40.44 and 14.78% for the COG functional groups (1-4), respectively.

The analysis of the distribution pattern of PHX genes based on percentages within COG categories for these genomes showed other features. It has been found that the percentage of PHX genes in category 1 for CPB and CTE and category 3 for CPB, CTE and CCM are well above the expected value of 10%, while both COG functional groups 2 and 4 contained less than 10% PHX genes. This means that the genes in these categories are comparatively better expressed than the rest in the genomes. Functional analysis showed that the COG functional group 3 (metabolism) included the highest number of PHX genes in all the genomes especially CTE where it amounts to 58.59%. The distribution patterns of the PHX genes in the different COG groups were more or less same in all the three strains. The top five COG categories for the *Chlorobium* strains were the following: translation, energy production and conversion, post translational modification, general function prediction and unknown functions for CPB; energy production and conversion, translation, amino acid transport and metabolism, coenzyme metabolism and general function prediction for CTE; translation, energy production and conversion, unknown function, inorganic ion transport and metabolism and cell envelope biogenesis for CCM. Intriguingly, some nitrogen fixing genes and photosynthesis related genes in the *Chlorobium* genomes have been found to fit into PHX category. Table 2 shows the PHX genes linked to nitrogen fixation and photosynthesis. All these provide significant insights into the genes required for maintaining the lifestyle of three *Chlorobium* genomes in different habitats. Huge

Table 2: PHX genes associated with nitrogen fixation and photosynthesis

Locus tag	Description	CAI value
Chlorobium phaeobacteroides		
Nitrogen fixing genes		
Cpha266_0742	Nitrogen regulatory protein P-II	0.780
Cpha266_0973	Nickel dependent hydrogenase, large subunit	0.776
Cpha266_0363	PTS IIA like nitrogen regulatory protein PtsN	0.772
Cpha266_0972	Hydrogenase (NiFe) small subunit HydA	0.762
Cpha266_0975	Hydrogenase maturation protease	0.759
Photosynthetic genes		
Cpha266_2299	Bacteriochlorophyll C binding protein	0.844
Cpha266_2463	Cytochrome C class I	0.803
Cpha266_2300	Chlorosome enveolope protein C	0.798
Cpha266_0202	Chlorosome enveolope protein B	0.794
Cpha266_0974	Ni/Fe-hydrogenase, b-type cytochrome subunit	0.776
Cpha266_1888	Bacteriochlorophyll A protein	0.774
Cpha266_1871	Proto-chlorophyllide reductase 57 kD subunit	0.768
Cpha266_0310	C-type cytochrome	0.761
Cpha266_0430	Cytochrome b/b6, N-terminal domain	0.757
Cpha266_2464	Cytochrome C, class I	0.757
Cpha266_0171	C-type cytochrome	0.756
Cpha266_0148	Cytochrome C, class I	0.754
Chlorobium tepidium		
Nitrogen fixing genes		
CT1533	Nitrogenase iron protein subunit nif H	0.818
CT1786	Nif U protein	0.810
CT1536	Nitrogenase molybdenum iron protein alpha chain	0.806
CT1892	Hydrogenase /sulfur reductase gamma subunit	0.803
CT1537	Nitrogenase molybdenum iron protein beta chain	0.801
CT1891	Hydrogenase /sulfur reductase beta subunit	0.796
CT1893	Hydrogenase /sulfur reductase delta subunit	0.794
CT1538	Nitrogenase iron molybdenum cofactor biosynthesis protein NifE	0.785
CT1539	Nitrogenase iron-molybdenum cofactor biosynthesis protein NifN	0.785
CT1534	Nitrogen regulatory protein PII	0.783
CT1540	Nif B protein	0.769
CT1798	Hydrogenase accessory protein HypB	0.766
CT1247	Hydrogenase, methyl-viologen-reducing type, delta subunit	0.760
CT0474	Sulfide dehydrogenase (flavoprotein) subunit SudB	0.759
CT1894	Hydrogenase/sulfur reductase, alpha subunit	0.757

Table 2: Continued

Locus tag	Description	CAI value
Photosynthetic genes		
CT1777	Bacteriochlorophyll c8 methyltransferase	0.786
CT0075	Cytochrome C-555	0.779
CT0303	Cytochrome b-c complex, cytochrome b subunit	0.775
CT1422	2-desacetyl-2-hydroxyethyl bacteriochlorophyllide A dehydrogenase	0.766
CT1818	Cytochrome d ubiquinol oxidase, subunit I	0.763
CT1499	Bacteriochlorophyll A protein	0.759
Chlorobium chlorocromatii		
Nitrogen fixing genes		
Cag_1583	Heterodisulfide reductase, subunit A/hydrogenase, delta subunit	0.711
Cag_1566	Hydrogenase/sulfur reductase, beta subunit	0.704
Cag_1244	Nitrogenase iron protein subunit NifH	0.703
Cag_0917	Cytochrome C-555	0.778
Photosynthetic genes		
Cag_0191	Chlorosome envelope protein B	0.746
Cag_0394	Cytochrome b6-f complex, iron-sulfur subunit	0.734
Cag_1172	Chlorosome envelope protein I	0.729
Cag_0219	Chlorosome envelope protein A	0.716
Cag_0975	Chlorosome envelope protein B	0.710
Cag_1358	probable cb-type cytochrome C oxidase subunit III	0.704

number of PHX genes related to metabolism in these genomes, especially in CTE presumably aid in their capability to subsist in extreme environments like hot springs and carry out essential life processes. Increased number of PHX genes in the COG groups like translation, energy production and conversion, general function prediction probably help them to compete with other bacteria in their habitat. CCM exists as a symbiont in phototrophic consortium. Higher amount of PHX genes related to inorganic ion transport and metabolism, translation and cell envelope biogenesis possibly enhance their ability to coordinate cell division and interspecific association between partners and regulate gene expression while living in such association. CTE exhibited somewhat different codon usage patterns compared to CCM and CPB. Although there is similarity in 16S rRNA patterns among the three strains yet the different codon usage pattern in CTE may be attributed to its biogeographical isolation in New Zealand.

As mentioned earlier, *Chlorobium* survive in anoxic sulfur rich environments and fix nitrogen as well carry out photosynthesis. Majority of the nitrogen fixing genes in the PHX category were hydrogenases, nitrogen regulatory proteins and nitrogenase iron protein complexes. These are extremely important for carrying out the process efficiently and the anoxic environment favors it. CTE had higher number of NFGs in the PHX category in contrast to CPB and CCM. Its survival in stressed environments like hot springs probably necessitated the need so that nitrogen fixation is carried out adeptly. The greater part of the photosynthesis related genes in the PHX category included bacteriochlorophylls, chlorosomes and cytochromes. Given the fact that these *Chlorobium* strains survive in stratified aquatic environments and highly reducing conditions of light, higher expression levels are indispensable for carrying out photosynthesis.

CONCLUSION

The *Chlorobium* genes are moderately biased and it is attributed to the failure of the translation machinery to co-evolve with higher codon usage. Among the *Chlorobium strains*, CTE genes are more biased and exhibited somewhat different codon usage patterns compared to CPB and CCM. There exists a degree of heterogeneity among the genes in these bacteria. GC3 compositional constraint plays a crucial role in the codon usage variation among the genes in the studied genomes of *Chlorobium*. Unlike CPB and CCM, expression level of the genes plays a key role in shaping codon usage variation

among the genes in CTE. Correlation of codon usage bias with tRNA content in *Chlorobium* genomes revealed the failure of the translation machinery of *Chlorobium* to co-evolve with higher codon usage in highly expressed genes resulting in moderate bias. Correspondence analysis reveals conserved nature of the nitrogen fixing genes and photosynthesis related genes. Positions of the genes in leading strand and lagging strand of replication had practically no role in shaping synonymous codon usage variation among the genes in these bacteria. Whole genome alignment of the *Chlorobium* genomes revealed their conserved nature. The CAI allocation patterns for CPB and CCM showed inclination towards higher CAI values. Majority of the potentially highly expressed genes in the COG categories were associated with metabolism. Presence of a number of genes associated with nitrogen fixation and photosynthesis in PHX provides them a selective advantage. These observations offer notable insights into the lifestyle of these bacteria surviving in anoxic, stratified environments in presence of reduced light and carrying out two very essential processes in nature. However, further studies involving whole genome DNA microarray and detailed proteomic analysis should be applied in the near future to gain further insights into the core architecture of the *Chlorobium* genomes.

ACKNOWLEDGMENTS

The authors are grateful to the Department of Biotechnology (DBT), Government of India, for providing financial help (GRANT NO. BT/BI/04/055/2001 dated 22/09/06) in setting up of Bioinformatics Facility in the Department of Botany, University of North Bengal. Arnab Sen acknowledges the same for providing DBT Overseas Associateship (Grant No. BT/HRD/03/01/2002).

REFERENCES

Banerjee, T., S. Basak, S.K. Gupta and T.C. Ghosh, 2004. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. J. Biomol. Struct. Dyn., 22: 13-24.

Basak, S., T. Banerjee, S.K. Gupta and T.C. Ghosh, 2004. Investigation on the causes of codon and amino acid usages variation between thermophilic *Aquifex aeolicus* and mesophilic *Bacillus subtilis*. J. Biomol. Struct. Dyn., 22: 205-214.

Cameron, J. and M. Aguade., 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol., 47: 268-274.

Darling, A.C.E., B. Mau, F.R. Blattner and N.T. Perna, 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res., 14: 1394-1403.

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: 6976-6985.

Eisen, J.A., K.E. Nelson, I.T. Paulsen, J.F. Heidelberg and M. Wu *et al*., 2002. The complete genome sequence of the green sulfur bacterium *Chlorobium tepidum*. Proc Natl. Acad. Sci. USA., 99: 9509-9514.

Garrity, G.M. and J.G. Holt, 2001. Phylum BXI. In: Chlorobi phy. nov. Bergey's Manual of Systematic Bacteriology. 2nd Edn., Vol. I. Boone, D.R. and R.W. Castenholz (Eds.). Springer, New York, ISBN: 978-0-387-98771-2 pp: 601-623.

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

Imhoff, J.F., 2003. Phylogenetic taxonomy of the family Chlorobiaceae on the basis of 16 s rRNA and fmo (Fenna-Matthews-Olson protein) gene sequences. Int. J. Syst. Evol. Microbiol., 53: 941-951.

Knight, R.D., S.J. Freeland and L.F. Landweber, 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol., 2: 0010.1-0010.13.

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

Markowitz, V.M., N. Ivanova, K. Palaniappan, E. Szeto and F. Korzeniewski *et al*., 2006. An Experimental metagenome data management and analysis system. Bioinformatics, 22: 359-367.

Overmann, J., 2000. The Family Chlorobiaceae. In: The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community, Release 3.1. Dworkin, M., S. Falkow, E. Rosenberg, K.H. Schleifer and E. Stackebrandt (Eds.). 3rd Edn. Springer-Verlag, New York.

Overmann, J. and K. Schubert, 2002. Phototrophic consortia: Model systems for symbiotic interrelations between prokaryotes. Arch. Microbiol., 177: 201-208.

Rocha, E.P.C., 2004. Codon usage bias from tRNAs point of view: Redundancy, specialization and efficient decoding for translation optimization. Genome Res., 14: 2279-2286.

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* Biol., 7: 547-558.

Sen, A., S. Sur, A.K. Bothra, D. Benson, P. Normand and L.S. Tisa, 2008. The implication of lifestyle on codon usage patterns and predictably highly expressed genes for three *Frankia* genomes. Anton. Van. Leewen., 93: 335-346.

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

Tsirigos, A. and I. Rigoutsos, 2005. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. Nucleic Acids Res., 33: 3699-3707.

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

Van Gemerden, H. and J. Mas, 1995. Ecology of Phototrophic Sulfur Bacteria. In: Anoxygenic Photosynthetic Bacteria Blankenship, R.E., M.T. Madigan and C.E. Bauer (Eds.). Kluwer Academic Publishers, Dordrecht, pp: 49-85.

Vogl, K., J. Glaeser, K.R. Pfannes, G. Wanner and J. Overmann, 2006. *Chlorobium chlorochromatii* sp. nov., a symbiotic green sulfur bacterium isolated from the phototrophic consortium *Chlorochromatium aggregatum*. Arch. Microbiol., 185: 363-372.

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: 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: 135-146.