3. Materials and Methods

3.1 Codon usage analysis

3.1.1 Retrieval of sequences

Genome sequences for the studied microorganisms listed in Table 3.1 were obtained from the IMG database (www.img.jgi.doe.gov) (Markowitz *et al.,* 2006) using their respective genome accession numbers. All of the protein coding genes, ribosomal protein genes which are considered to be highly expressed and nitrogen fixation related genes were considered for the analysis. In addition, the sequences for COGs (Cluster of orthologous genes) (Tatusov *et al.,* 2003) functional groups for the studied microorganisms and their tRNA copy numbers were also obtained from the aforesaid database.

3.1.2 Analysis of codon usage variation

To analyze the codon usage patterns of the studied organisms the software Codon W (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?Form=codonw) (Peden, 1999) was used. The parameters such as GC content, GC3 content, Nc, CBI, Fop, GRAVY and aromaticity were calculated. CAI values were calculated using the e-CAI server (Puigbo *et al.,* 2008). All these parameters reduce the codon usage data to a useful summary and enlighten about the factors affecting codon usage patterns in microorganisms.

The GC content estimates the amount of the guanine cytosine in the nucleotide sequences. The GC3 content determines the frequency of either G or C nucleotides present in the third position of the synonymous codon. It however excludes methionine, tryptophan and the termination codons.

The effective number of codons used in a gene (Nc) is an important parameter that can measure overall codon bias of synonymous codons (Wright, 1990). Its

Table 3.1: List of organisms' genomes and proteomes undertaken for the study

Group	Organism Name	CODE	Accession #
Cyanobacteria	*Anabaena variabilis* ATCC 29413	ANVAR	NC_007413
	Cyanothece sp. PCC 7424	C7424	NC_011729
	Cyanothece sp. PCC 7425	C7425	NC_011884
	Cyanothece sp. PCC 8801	C8801	NC_011726
	Cyanothece sp. PCC 51142	C51142	NC_010546
	Cyanothece sp. CCY0110	CCY0110	NZ_AAXW00000000*
	Lyngbya sp. PCC8106	LYN	NZ_AAVU00000000*
	Nostoc sp. 7120	N7120	NC_003272
	Nostoc sp. 73102	N73102	NC_010628
	Nodularia spumigena CCY9414	NOSPUM	NZ_AAVW00000000*
	Synechococcus sp.JA-2-3B'a(2-13)	SYNE	NC_007776
	Trichodesmium erythraeum IMS101	TRICHOS	NC_008312
Proteobacteria	*Azoarcus sp. BH72*	AZOR	NC_008702
	Azotobacter vinelandii AvOP	AZOTO	NZ_AAAU00000000*
	Azorhizobium caulinodans ORS 571	AZORHI	NC_009937
	Beijerinckia indica subsp. indica ATCC 9039	BEI	NC_010581
	Bradyrhizobium sp. BTAI1	BRBTAI	NC_009485
	Bradyrhizobium japonicum USDA 110	BRUSDA	NC_004463
	Bradyrhizobium sp. ORS278	BRORS	NC_009445
	Gluconacetobacter diazotrophicus PAL5	GLUCO	NC_010125
	Klebsiella pneumoniae pneumoniae MGH78578	KLEB	NC_009648-53
	Mesorhizobium loti MAFF303099	MESMAF	NC_002678
	Mesorhizobium sp. BNC1	MESOBN	NC_008242
	Rhizobium etli CFN 42	RHICFN4	NC_004041
	Rhizobium etli CIAT 652	RHICIAT	NC_010994
	Rhizobium leguminosarum bv. trifolii WSM1325	RHIWSM	NC_011366
	Rhizobium leguminosarum bv. viciae 3841	RHIVICIA	NC_008378
	Rhodobacter sphaeroides 2.4.1	RHODO	NC_007493
	Sinorhizobium meliloti 1021	SMELILO	NC_003037
	Sinorhizobium medicae	SNOMEDI	NC_009620
Actinobacteria	*Frankia alni ACN14a*	ACN14A	NC_008278
	Frankia sp. CcI3	CcI3	NC_007777
	Frankia sp. EAN1pec	EAN1PEC	NC_009921
Firmicutes	*Clostridium acetobutylicum ATCC 824*	CACET	NC_003030
	Clostridium beijerinckii NCIMB 8052	CBEI	NC_009617
Green sulfur bacteria	*Chlorobium chlorochromatii CaD3*	CCHROM	NC_007514
	Chlorobium ferrooxidans DSM 13031	CFEROX	NZ_AASE00000000*
	Chlorobium limicola DSM 245	CLIMI	NC_010803
	Chlorobium phaeobacteroides DSM 266	CPHAEO	NC_008639
	Chlorobium phaeobacteroides BS1	CPEO1	NC_010831
	Chlorobium tepidum TLS	CTEPI	NC_002932
Methanogenic Archaea	*Methanobacterium thermoautotrophicum str. Delta H*	MTHER	AE000666
	Methanococcus maripaludis strain S2	MMARI	BX950229
	Methanosarcina mazei strain Goe1	MMAZ	AE008384
	Methanospirillum hungatei JF-1	MHUN	NC_007796

*Draft sequence

value represents the number of equal codons that would generate the same codon usage bias observed (Sen *et al.*, 2008).Values for the effective number of codons range from 20 (when only one codon is per amino acid) to 61 (when all codons are used in equal probability). The Nc value is influenced by mutational biases and or/selection for particular codons. However, Wright (1990) has pointed out that in organisms where mutational bias absolutely determines synonymous codon usage, the Nc value ranges between 31-61 depending upon the genomic GC content. While calculating the effective number of codons at first *F*^ (F caret) is calculated in each of the synonymous groups:

$$\hat{F} = \frac{\left(n_{aa}\sum_{i=1}^{j} p^{2}\right) - 1}{n_{aa} - 1} \qquad (1)$$

where, p symbolizes the fraction of usage of a codon i inside its synonymous cluster of size j, and n_{aa} the total usage of that synonymous group. The average of *F*^ for synonymous groups of same size (i.e. 2, 4, and 6) is also determined. Nevertheless, in lack of isoleucine residues *F*^ is calculated as an average of $F^{\wedge av2}$ and $F^{\wedge av4}$ and Nc value is determined using the following formula:

$$Nc = 2 + 9/\hat{F}^{av2} + 1/\hat{F}^{3} + 5/\hat{F}^{av4} + 3/\hat{F}^{av6} \qquad (2a)$$

In order to include the influence of G+C

Table 3.2: Major COG groups and their categories

COG GROUP	MAJOR CATEGORIES
1. Information storage and 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 envelope biogenesis, outer membrane
	N—Cell motility and secretion
	U—Intercellular trafficking, secretion and vesicular transport
	O—Posttranslational modification, protein turnover, and chaperones
3. Metabolism	C—Energy production and conversion
	G—Carbohydrate transport and metabolism
	E—Amino acid transport and metabolism
	F—Nucleotide transport and metabolism
	H—Coenzyme metabolism
	P—Inorganic ion transport and metabolism
	I—Lipid metabolism
	Q—Secondary metabolites biosynthesis, transport and catabolism
4. Poorly characterized	R—General function prediction only
	S—Unknown function

bias the following equation is used to determine the expected value of Nc under random codon usage:

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

Here, S represents GC3 values.

The codon adaptation index (CAI) is a commonly used gauge to determine synonymous codon usage in prokaryotes as well as eukaryotes. It is a measure of codon usage within a gene relative to reference sets of genes that are known to be highly expressed (Sharp and Li, 1987). Before calculating the codon adaptation index it is essential to determine the relative synonymous codon usage (RSCU) values from the set of highly expressed genes in an organism as determined by Sharp and Li (1987):

$$RSCU_{ij} = \frac{x_{ij}}{1/n_i \sum_{j=1}^{n_i} x_{ij}} \quad (3)$$

here x_{ij} signifies the number of occurrence of the *j*th codon for *i*th

amino acid, and n_i symbolizes the size of the synonymous group for the *i*th amino acid (i.e., 2,3,4 or 6). The CAI value for the gene is then determined by the geometric mean of the relative adaptiveness values of each of the codons present in the genes. CAI (Sharp and Li, 1987) is calculated using the formula:

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L} \ln \omega k\right) \quad (4)$$

where, ω_k signifies the relative adaptedness of the k^{th} codon and L represents the number of synonymous codons in the gene. In the e-CAI server the CAI values were determined using codon usage table of the particular organism as reference. CAI values vary from 0 to 1 with higher CAI values indicating that the gene of interest has a codon usage pattern more similar to that of highly expressed genes (Sen *et al.*, 2007).

The codon bias index (CBI) (Bennetzen and Hall, 1982) is a gauge of directional codon bias and determines the level to which a gene uses a subset of optimal codons. The codon bias index values ranges from 0 to 1. It is calculated as follows:

$$CBI = N_{opt} - N_{ran} / N_{tot} - N_{ran} \quad (5)$$

where N_{opt}= number of optimal codons; N_{tot}= total number of synonymous codons; N_{ran}=expected number of optimal codons in cases where codons are assigned randomly.

The frequency of optimal codons (Fop) (Ikemura, 1985) is the fraction of synonymous codons that are optimal codons. If rare codons are identified there is a stipulation for determining the original Fop index (Equation 6a) or the modified Fop index (Equation 6b). All the negative values arising while determining Fop are adjusted to zero. Fop is calculated as follows:

$$Fop = N_{oc} / N_{sc} \quad (6a)$$

$$Fop_{(\text{mod})} = N_{oc} - N_{rc} / N_{sc} \quad (6b)$$

where N represents the frequency of each codon type used, whereas N_{oc}, N_{rc} and N_{sc} stand for optimal codons, rare codons and synonymous codons respectively. Fop values ranges from 0 to 1. In case where Fop values are 1 the genes are said to be made entirely of optimal codons (Ikemura, 1985).

Laa determines the length of the amino acids. GRAVY (Kyte and Doolittle, 1982) is the gauge of the hydrophobicity of the proteins coded by the genes. Positive GRAVY (hydrophobic) and negative GRAVY (hydrophilic) scores are the arithmetic mean of the sum of hydropathic (solubility) indices of the amino acids. Aromaticity determines amino acid usage, given that disparity in amino acid composition can be relevant for evaluating codon usage (Lobry and Gautier, 1994).

In order to test whether the values of the aforesaid indices in nitrogen fixing genes, ribosomal protein genes significantly differ from that of the protein coding genes, Z test was performed.

3.1.3 Correspondence analysis of codon and amino acid usage

The Codon W software (Peden, 1999) was used to calculate the correspondence analysis of codon count and amino acid usage frequencies. Owing to some fallacies associated with correspondence analysis of RSCU it was not considered for the analysis. Correspondence analysis is a multivariate statistical technique that 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). 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

Table 3.3: List of organisms used for analysis of TTA codons

. Organisms	Accession number
Azoarcus sp. BH72	NC_008702
Azotobacter vinelandii AvOP	NZ_AAAU00000000*
Azorhizobium caulinodans ORS 571	NC_009937
Bradyrhizobium sp. BTAil	NC_009485
Bradyrhizobium sp. ORS278	NC_009445
Frankia alni ACN14a	NC_008278
Frankia sp. CcI3	NC_007777
Frankia sp. EAN1pec	NC_009921
Gluconacetobacter diazotrophicus PAL5	NC_010125
Rhodobacter sphaeroides 2.4.1	NC_007493

*Draft sequence

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). Correspondence analyses on amino acid usage for the studied organisms were performed for the protein coding genes to recognize the apparent forces in characterizing adaptation of the expressed proteins.

3.1.4 Correlation of different indices

The aforementioned parameters were correlated amongst themselves and with the principal axis of variations for correspondence analysis of codon count and amino acid usages to get further insights into their role in manipulating the diversity of codon usage patterns in the studied microorganisms.

3.1.5 Correlating codon usage bias with tRNA content

Eduardo Rocha (2004) studied the correlation between synonymous codon usage and tRNA content in some bacterial genomes. The optimal generation times for the studied microorganisms were obtained and they were correlated with the tRNA copy numbers for the microorganisms.

3.1.6 Prediction of potentially highly expressed genes

The CAI values obtained for protein coding genes in the studied microorganisms are useful for predicting the level of expression of a gene (Sharp and Li, 1987). Wu *et al.,* (2005) analyzed proteome results and validated the correlation between CAI values and expression levels showing experimentally that CAI predicted potentially highly expressed genes indeed are highly expressed. Jansen *et al.,* (2003) also confirmed this finding with yeast genome. CAI values for these studied genomes were examined to identify the predicted highly expressed genes. As defined by Wu *et al.*, (2005, 2005a), the top 10% of the genes, in terms of CAI values, were classified to be predicted highly expressed genes.

3.1.7 Functional analysis of predicted highly expressed genes

Clusters of orthologous groups of proteins (COG) were used to understand the functional distribution of the predicted highly expressed genes among the studied genomes. To help the analysis, each of the COG functional categories was clustered into four COG functional groups. The major COG functional groups are tabulated in Table 3.2. The functional analyses of COGs based on potentially highly expressed genes in the studied organisms were used to understand their role in influencing the lifestyle of the organisms.

3.2 Analysis of TTA codons from GC rich diazotrophs

3.2.1 Sequence retrieval

The nucleotide sequences of the high GC containing microorganisms enlisted in Table 3.3 were considered for the analysis and were retrieved from the IMG database (www.img.jgi.doe.gov) (Markowitz *et al.*, 2006).

3.2.2 Determination of TTA codons

The server TTA Lynx (http:// ttalynx.bio.lnu.edu.ua) (Zaburannyy *et al.*, 2009) was used to determine the number of TTA codons and their corresponding gene sequences. The studied sequences were loaded in FASTA formats and contained the coding sequences (CDS). The aforesaid server was then used to compute the codon wise statistics of the TTA codons.

3.2.3 Codon usage analysis of TTA codon containing gene sequences

The predicted TTA containing gene sequences were analyzed for codon usage patterns with respect to the parameters like GC content, GC3 content, Nc, CAI, CBI, Fop, GRAVY and aromaticity so as to ascertain their characteristics. The software used are as per section 3.1.2.

3.2.4 Functional analysis of TTA codon containing genes

Clusters of orthologous groups of proteins (COG) were used to study the functional distribution of the TTA codon containing genes in order to establish their role in influencing the lifestyle of the GC rich organisms.

3.3 Proteome analysis

3.3.1 Retrieval of sequences

The sequences of the total proteomes and that for the nitrogen fixation related protein sequences of the microorganisms are listed in Table 3.1 along with their code names. These sequences were retrieved from (IMG) Integrated Microbial Genome database (www.img.jgi.doe.gov) (Markowitz *et al.*, 2006). The sequences of proteins in COGs (Cluster of orthologous genes) (Tatusov *et al.*, 2003) functional groups from studied proteomes were also retrieved from the IMG database. Symbiotic plasmids for *Rhizobium etli* CFN 42, *Bradyrhizobium* sp. BTAI1 and *Sinorhizobium meliloti* were also considered for the analysis.

3.3.2 Determination of protein isoelectric point, energetic costs and amino acid adaptation index

The protein isoelectric points (pI), protein energetic costs and amino acid adaptation index (AAAI) were calculated using the software DAMBE (http://dambe.bio.uottawa.ca) (Xia and Xie, 2001). The allocation of isoelectric points (pI) in a proteome is one of the vital characteristics of proteins. The isoelectric point (pI), is defined as the pH at which a particular molecule or surface has no net electrical charge i.e. the negative and positive charges are equal (Nandi *et al.*, 2005). Amino-acid adaptation index (AAAI) is used for assessing the adaptation of proteins and for making comparisons of amino acid usage in different organisms (Xia and Xie, 2001). The mean amino acid composition inertly reveals random variation of the genetic code. It appears that the most crucial thing shaping frequency of amino acids in a protein is the number of codons coding for it instead of particular optimal amino acids (Xia and Xie, 2001). The energetic cost of protein accounts for a noteworthy amount in metabolism. Energy in the form of high-energy phosphate bonds and reducing power are utilized in carrying out reactions and more is required for alteration of available starting metabolites to amino acids (Akashi and Gojobori, 2002). Hence, determination of energetic costs of proteins is vital for metabolic efficiency in proteomes of microorganisms.

3.3.3 Correlation of different indices

It has been reported that environment and GC content play a role in influencing amino-acid usage in organisms (Tekaia and Yeramian, 2006). On the basis of this GC content, GC3 composition (amount of G or C codons in the third position), hydropathicity (GRAVY score) and aromaticity determined using Codon W software (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?Form=codonw) (Peden, 1999) for protein coding genes of the studied microorganisms previously were correlated with isoelectric point (pI), amino acid adaptation index (AAAI) and energetic costs of proteins. Protein isoelectric point (pI) values were further correlated with protein energetic costs and aromaticity.

3.3.4 Variations of isoelectric point across COGs

Variations of isoelectric points (pI) across the studied nitrogen fixing microorganisms using COGs (Tatusov *et al.,* 2003) functional groups were investigated. The COGs functional groups were clustered into four classes for convenience (Sur *et al.,* 2008) as mentioned earlier. In the first step the percentage of amino acids in the isoelectric point category for each COG functional group was determined. In the next step percentage of amino acids for each COG functional group amongst the total COGs were determined. Chi-square tests for the COG groups in the acidic dataset as well as basic dataset were performed to identify significant differences.

3.4 Phylogenetic analysis of *nif* genes

3.4.1 Using condensed matrix method

3.4.1.1 Retrieval of sequences

The nucleotide sequences of 16S rRNA, *nif*H, *nif*K, *nif*D genes and complete genomes from microorganisms coming from cyanobacteria, proteobacteria, gram positive bacteria, green sulfur bacteria and archaea are listed in Table 3.4. The sequences were retrieved from the Integrated Microbial genomes (IMG) database (www.img.jgi.doe.gov) (Markowitz *et al.,* 2006).

3.4.1.2 Frequency of triplets of nucleic acid bases

It is well known that a DNA sequence of four letters consists of 64 possible triplets (subsequences of length 3) starting from AAA, AAT, AAG, AAC, ATA, ATT, ATG, ATC, AGA, AGT, AGG, AGC, ACA, ACT, ACG, ACC etc. (Randic *et al.,* 2001). The triplets contain all the relevant information for polypeptide synthesis. A program developed in Turbo C++ was used to count each and every potential triplet from nucleotide sequences of 16S rRNA, *nif*H, *nif*K, *nif*D genes and whole genomes from the studied organisms.

Table 3.4: List of organisms and sequences used for phylogenetic analysis using condensed matrix based method

Organism	Accession#
Cyanobacteria	
Anabaena variabilis ATCC 29413	(NC_007413)
Cyanothece sp. PCC 8801	(NC_011726)
Cyanothece sp. ATCC 51142	(NC_010546)
Cyanothece sp. CCY0110	(NZ_AAXW00000000)*
Cyanothece sp. PCC 7424	(NC_011729)
Cyanothece sp. PCC 7425	(NC_011884)
Lyngbya sp. PCC 8106	(NZ_AAVU00000000)*
Nodularia spumigena CCY9414	(NZ_AAVW00000000)*
Nostoc sp. PCC 7120	(NC_003272)
Nostoc punctiforme PCC 73102	(NC_010628)
Trichodesmium erythraeum IMS101	(NC_008312)
Synechococcus sp. JA-2-3B'a(2-13)	(NC_007776)
Proteobacteria	
Azoarcus sp. BH72	(NC_008702)
Azotobacter vinelandii AvOP	(NZ_AAAU00000000)*
Azorhizobium caulinodans ORS 571	(NC_009937)
Beijerinckia indica subsp. indica ATCC 9039	(NC_010581)
Bradyrhizobium sp. BTAi1	(NC_009485)
Bradyrhizobium japonicum USDA 110	(NC_004463)
Bradyrhizobium sp. ORS278	(NC_009445)
Gluconacetobacter diazotrophicus PAL5	(NC_010125)
Klebsiella pneumoniae	(NC_009648-53)
Mesorhizobium loti MAFF303099	(NC_002678)
Rhizobium etli CFN 42	(NC_004041)
R. etli CIAT 652	(NC_010994)
R. leguminosarum bv. trifolii WSM1325	(NC_011366)
R. leguminosarum bv. viciae 3841	(NC_008378)
Rhodobacter sphaeroides	(NC_007493)
Sinorhizobium meliloti 1021	(NC_003037)
Actinobacteria	
Frankia alni ACN14a	(NC_008278)
Frankia sp. CcI3	(NC_007777)
Frankia sp. EAN1pec	(NC_009921)
Clostridia	
Clostridium acetobutylicum ATCC 824	(NC_003030)
Clostridium beijerinckii NCIMB 8052	(NC_009617)
Green sulphur bacteria	
Chlorobium chlorochromatii CaD3	(NC_007514)
Chlorobium ferrooxidans DSM 13031	(NZ_AASE00000000)*
Chlorobium limicola DSM 245	(NC_010803)
Chlorobium phaeobacteroides DSM 266	(NC_008639)
Chlorobium phaeobacteroides BS1	(NC_010831)
Chlorobium tepidum TLS	(NC_002932)
Methanogenic archaea	
Methanobacterium thermoautotrophicum str. Delta H	(AE000666)
Methanococcus maripaludis strain S2	(BX950229)
Methanosarcina mazei strain Goe1	(AE008384)
Methanospirillum hungatei JF-1	(NC_007796)

*Draft sequences

The introduction of a 4×4×4 cubic matrix was necessary to go for further analysis. The 4×4×4 cubic matrix comprising of 64 possible entries helps in resolving the occurrence of the probable 64 triplets in a DNA sequence.

In case of a cubic matrix, it is possible to obtain three groups of 4×4 matrices (Sur *et al.,* 2009) each of which containing all entries of the cubic matrix (Randic *et al.,* 2001). In most cases, the group {M_1, M_2, M_3, M_4} represents the cubic matrix. The matrices were formed using all the triplets for all the studied DNA sequences. Condensed categorization of the primary sequence is derived from 4×4 matrices, whose rows and columns are related with the A, G, C and T bases. The four matrices enclose information about the frequencies of occurrence of all possible triplets of the DNA sequence along with the information regarding the frequency of occurrence of pairs of each and every letter within a DNA sequence (Randic *et al.,* 2001). In our method the codon positions 1, 2 and 3 were given equal weight subsequently addition or deletion of bases during the course of evolution were given due care so as to incorporate their influence. The methodology depicts DNA by condensed a matrix counting the rate of presence of adjoining base pairs (Randic, 2000).

3.4.1.3 Calculation of eigen value and construction of phylogram

Leading eigenvalues were calculated using MATLAB (version 5.0.0.4069) software. These eigenvalues are a special set of scalars associated with a linear system of equations, usually matrix equations that are often regarded as characteristic roots, characteristic values (Hoffman and Kunze, 1971) and proper values or latent roots (Marcus and Minc, 1988). Evaluation of DNA sequences for similarity or dissimilarity is normally aided by the convenience of leading eigenvectors calculated by this method. Diversity between eigenvalues was used to study sequence similarity/ dissimilarity keeping in mind the characterization of a sequence by leading eigenvalue (Nandy *et al.,* 2006). Matrices linked to each sequence are estimated and the leading eigen values computed. Variations in leading eigen values concurrent to the string are estimated and the relationships between genes investigated. Distance matrixes of the studied sequences were constructed by summing up the square of the difference of eigen values. Phylograms were built by cluster analysis of the similarity matrix using PHYLIP (Ver 3.65) (Felsenstein, 1989) and drawn with PHYLODRAW (Ver 0.8).

The nucleotide triplet based condensed matrix phylogeny has been successfully applied in addressing the evolutionary scenario of amino-acyl tRNA synthetases in three domains of life

(Mondal *et al.*, 2008) and H1N1 viruses (Sur *et al.*, 2009).

3.4.2 Composition vector (CV) approach in whole genomes and proteomes

The CVTree server (http:// tlife.fudan.edu.cn/cvtree) (Qi *et al.*, 2004) was used to examine the phylogeny of whole genomes from the studied nitrogen fixing microorganisms. This technique thwarts the vagueness of choosing genes for phylogenetic restoration evading alignment of sequences (Qi *et al.*, 2004). This method infers evolutionary relatedness of the genomes from frequency of amino acid K-strings in their whole proteomes. Complete DNA and proteome sequences were analyzed by feeding them into the CVTree server.

3.4.3 Evolution of *nif* genes as a unit

Given the fact that, *nif*H, *nif*K and *nif*D genes act in tandem and are expected to evolve at consistent rates owing to their mutual interdependence; other important parameters such as, GC (guanine cytosine) content, Nc (effective number of codons) (Peden, 1999) GC3 (guanine cytosine content at 3rd position), Laa (length of the amino acids) are also estimated to be consistent. Taking this fact into consideration the next aim was to investigate whether these genes evolved in a homogenous manner and evolved as a unit.

3.5 Homology modeling of NifH proteins from *Frankia*

3.5.1 Retrieval of the target sequence

The amino acid sequences of the NifH protein from *Frankia* sp. CcI3 and *Frankia* sp. EAN1pec bearing accession numbers NC_007777 and NC_009921 were obtained from the IMG database (www.img.jgi.doe.gov) (Markowitz *et al.,* 2006) and that for *Frankia* sp. ACN14a was obtained from the *Frankia alni* ACN14a Genome page (http:// cmr.tigr.org/tigr-scripts/CMR/ GenomePage.cgi?org=ntfa01webpage) bearing TIGR locus NT01FA6905 and GenBank accession number CAJ65436. The sequences were thoroughly scanned and it was confirmed that the 3D structure of the proteins were not available in Protein Data Bank (http:// www.rcsb.org/pdb/home/home.do), consequently the current work of constructing the 3D model of the nitrogenase iron proteins of the three *Frankia* strains were initiated.

3.5.2 Template selection

The preliminary task in homology modelling technique is to recognize protein structures linked to the target sequence and subsequently select those that will be used as templates (Centeno *et al.,* 2005). Position specific iterative BLAST i.e., PSI-BLAST (Altschul *et al.,* 1997) was carried out against database specification of PDB proteins which were available at the National Centre for Biotechnology Information (NCBI) Web server (http:// www.ncbi.nlm.nih.gov/blast/) to find out remote similarities. The appropriate template was selected on the basis of the quality of the experimental template structure, environmental likeness and phylogenetic similarity.

3.5.3 Alignment of the target and template

An optimal alignment between the target sequence and template is required to construct a 3D model of the target protein, after the template sequence has been recognized. Multiple sequence alignments were performed using Clustal W 1.83 (Thompson *et al.,* 1994) using default settings and the aligned sequences were extracted in (.) PIR format (Thompson *et al.,* 1994). The aligned sequences were converted into (.) ALI format (Sali and Blundell, 1993). The acquired alignments were crucially assessed in terms of number, length and position of the gaps to make it more reliable. Secondary structures were predicted using HNN (Hierarchical Neural Network method) (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html).

3.5.4 Construction of the rough models

The rough 3D models of the NifH protein from three *Frankia* strains were constructed by MODELLER 9v4 program (Sali and Blundell, 1993) using the alignment between the NifH proteins and the template proteins. The technique is based upon the satisfaction of the spatial restraints acquired from the alignment (Centeno *et al.,* 2005). The method is theoretically comparable to that used in resolving protein structures from NMR-derived restraints. These restraints are usually acquired by assuming that the resultant distances between aligned residues in the template and the target structures are alike. These restraints obtained on the basis of homology, are generally improved by stereochemical restraints on bond lengths, bond angles, dihedral angles, and non-bonded atom-atom contacts that are attained from a molecular mechanics force field. After reducing the violation of all the restraints the model is finally obtained (Sali and Blundell, 1993).

3.5.5 Refinement of the models

The models obtained by the homology modelling technique often contain certain amount of errors and become decisive if the concerned residues are associated with the protein function (Centeno *et al.*, 2005). To overcome this problem, refinement of the models is necessary. During the refinement process, the constructed protein models were subjected to constraint energy minimization with a harmonic constraint of 100 kJ/mol/Å^2, using the steepest descent (SD) and conjugate gradient (CG) method to remove any existing bad sectors between the protein atoms and regularizing the protein structure geometry. All of the computations were done in vacuo with the GROMOS96 43B1 parameters set using the Swiss-Pdb Viewer package (http://expasy.org/ spdv/program/spdv37sp5.zip) (Kaplan and Littlejohn, 2001). GROMOS is a commonly used molecular dynamics computer simulation program for investigating biomolecular systems and applied for examining conformations acquired by experimental or computer simulations (Kaplan and Littlejohn, 2001). Hydrogen bonds were not considered.

3.5.6 Evaluation of Refined Models

For evaluation of their internal quality and reliability, the refined models were subjected to the following tests: ProSA (Wiederstein and Sippl, 2007) analysis was performed to assess the accuracy and reliability of the modelled structures and check the 3D models for potential errors. VERIFY3D (Eisenberg *et al.*, 1997) was used to validate the refined structures. Here, the 3D structures of the protein models are compared to its own amino-acid sequence taking into consideration a 3D profile calculated from the atomic coordinates of the structures of correct proteins (Eisenberg *et al.*, 1997). The constructed models of the NifH proteins were evaluated for their backbone conformation using a Ramachandran plot (Ramachandran *et al.*, 1963). The Auto Deposition Input Tool (ADIT) (http://deposit.pdb.org/ validate) was used to inspect the favorable and unfavorable properties of the modelled structures. SAVES (Structure analysis and verification server) (http://nihserver.mbi.ucla.edu/ SAVS/) was used to carry out the verifications of the models with PROVE and ERRAT. Presence of pockets in the structures was predicted using CASTp server (Dundas *et al.*, 2006). The refined models were submitted to ProFunc (http://www.ebi.ac.uk/thronton -srv/databases/ProFunc) (Laskowski *et al.*, 2005) to recognize the functional region in the proteins. Since there are no data on the site-directed mutagenesis, site-directed mutagenesis predictions were carried out using the server SDM (http://www-cryst.bioc.cam.ac.uk/~sdm/ sdm.php). Stability changes associated with mutations were assessed with I-Mutant 2.0 (http:// gpcr.biocomp.unibio.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi.).

3.5.7 Studying intrinsic dynamics of the protein models

An understanding of structural dynamics of the proteins is essential to gain greater insights into their important biological functions (Yang *et al.*, 2006). The studies on the structural dynamics of the three protein models from *Frankia* were performed using the WEBnm (http://www.bioinfo.no/tools/ normalmodes) program (Hollup *et al.*, 2005) that calculates the slowest modes and related deformation energies; ElNemo (http://igs-server.cnrs-mrs.fr/ elnemo/index.html.) program (Suhre and Sanejouand, 2004) which calculates the normal mode analysis of the proteins contributing to the corresponding protein movement, and the MolMovDB (http://molmovdb.org/) program that determines the five lowest frequency modes (Alexandrov *et al.*, 2005). Normal mode analysis forecasts the probable movements of the proteins and is the method of selection for exploring the slowest activity of choice in proteins (Hollup *et al.*, 2005). Solvent accessibility of the amino-acid residues in the modeled proteins was determined using ASA-view (http:// gibk26.bse.kyutech.ac.jp/~shandar/ netasa/asaview/) software (Ahmad *et al.*, 2004).