4. Results and Discussion

4.1.The synonymous codon usage patterns

4.1.1 Synonymous codon usage pattern and heterogeneity

The primary aim in this study was to detect the degree of codon usage. Considerable number of microorganisms with a balanced AT/GC genome has codon heterogeneity (Sen *et al.*, 2007). This is linked to gene expression level. Highly expressed genes contain higher percentage of codons that are translationally optimal (Ikemura, 1985). The studied organisms were analyzed for codon heterogeneity using the GC content, GC3 content and Nc values for all of the genes. The effective number of codons versus GC3 plots has been suggested to be an important means to investigate the codon usage variations among the genes in the same genome (Peden, 1999). In most of the high GC rich and moderately GC rich organisms it was found that the Nc (effective number of codons) values ranged from 21±2 to 61 suggesting that these genomes exhibited considerable amount of codon heterogeneity.

In the clostridial genomes it is seen that the GC content and subsequently the GC3 content is very low suggesting that they are AT rich. The pattern is similar in all categories of genes and the Nc values indicated that the organisms are moderately biased. However the ribosomal protein genes for *Clostridium beijerinckii* showed high bias. It is seen from the Nc/GC3 plots depicted in Figure 4.1.1.1 that the ribosomal protein genes (RPGs) for *Clostridium beijerinckii* are clustered together at one end of the plot. The continuous curve in the Nc/GC3 plot indicated the factor influencing codon usage bias. In cases where synonymous codons are absolutely determined by GC3 values, Nc values would have had fallen on the

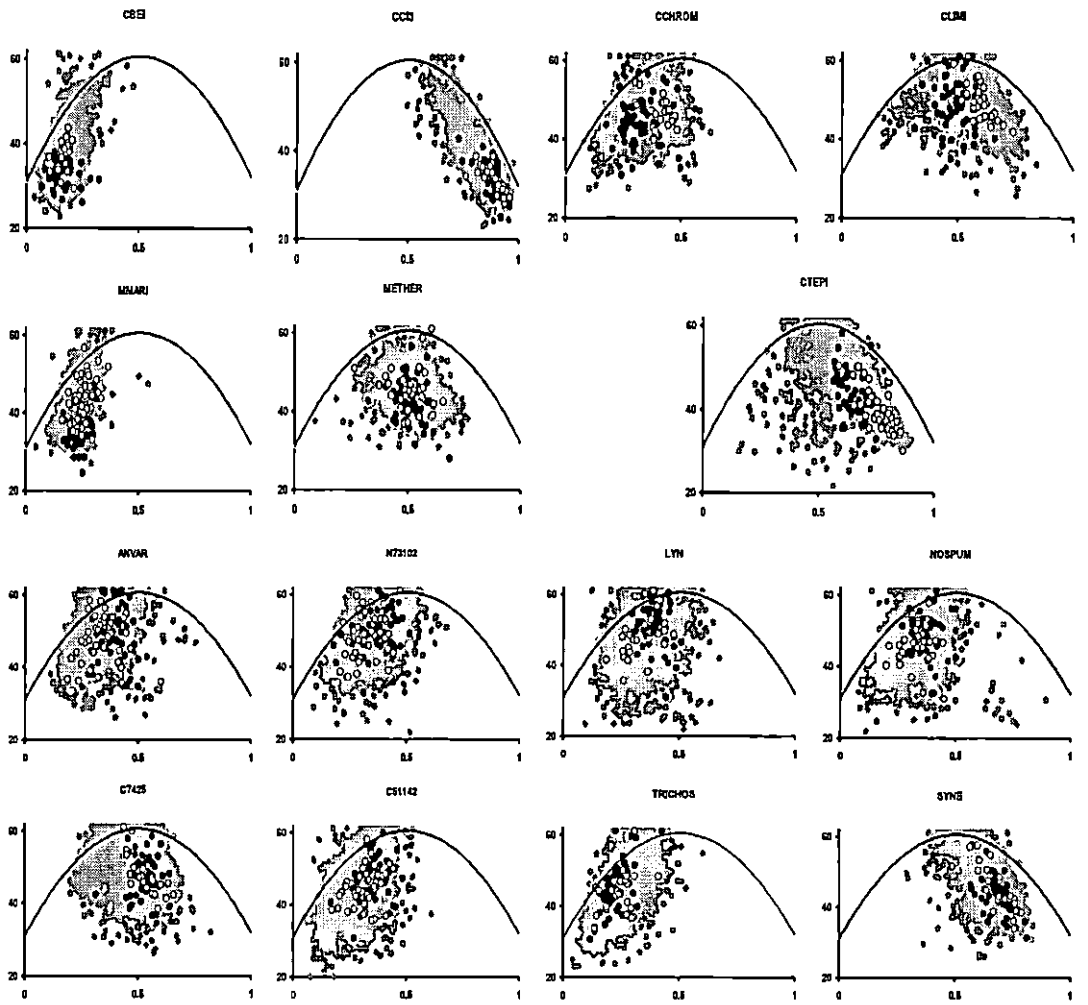

Figure 4.1.1.1: Nc/GC3 plots for CBEI, CCI3, MMARI, MTHER, CCHROM, CLIMI, CTEPI, ANVAR, N73102, LYN, NOSPUM, C7425, C51142, TRICHOS and SYNE. Protein coding genes are represented by grey circles, ribosomal protein genes by black circles and nitrogen fixing related genes by white circles. The continuous curve represents the null hypothesis that GC bias at the synonymous site is solely due to mutation and not selection. X and Y axis are represented by GC3 and Nc respectively

expected curve. However, barring very few genes most of the values obtained for majority of the genes were well below the expected values. This particular trend has been similar in all the studied diazotrophs. The CAI values for the clostridial strains are high however as expected the RPGs had higher CAI values compared to the protein coding genes. The CBI values in both the strains are negative indicating that the numbers of optimal codons are less in comparison to the random codons (Peden, 1999). The mean Fop values are low indicating mutational bias may be acting on these genomes. In *Clostridium beijerinckii*, the mean Fop values of the RPGs are comparatively higher than other categories of genes signifying that these genes have higher

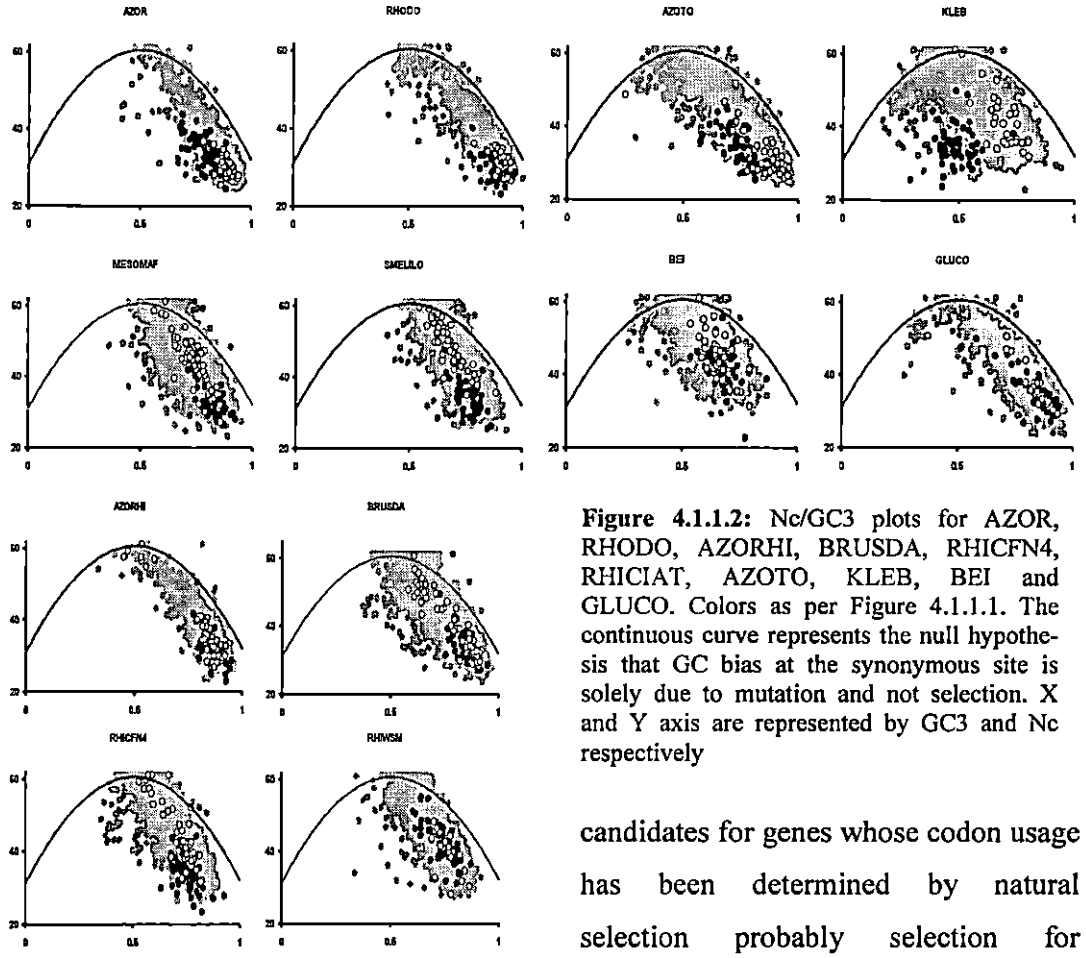

Figure 4.1.1.2: Nc/GC3 plots for AZOR, RHODO, AZORHI, BRUSDA, RHICFN4, RHICIAT, AZOTO, KLEB, BEI and GLUCO. Colors as per Figure 4.1.1.1. The continuous curve represents the null hypothesis that GC bias at the synonymous site is solely due to mutation and not selection. X and Y axis are represented by GC3 and Nc respectively

proportion of optimal codons. The GRAVY and aromaticity values are low. Z scores for the studied indices did not reveal any significant differences. Correlation of CAI values with Nc, GC3 and GC values did not reveal any significant differences.

In the actinobacterial *Frankia* strains, the GC, GC3 content are very high as a result their Nc values are low indicating a high degree of codon bias. Genes having Nc values less than 40 have stronger codon bias and are probably under the influence of mutational bias (Sen *et al.*, 2008). They are good candidates for genes whose codon usage has been determined by natural selection probably selection for translational efficiency (Peden, 1999). The *Frankia* genomes had a large number of genes with Nc values less than 40 including almost all of the ribosomal protein genes. It has been reported (Sen *et al.*, 2008) that CcI3 had sustained strong reducing evolutionary pressures for the genes with strong codon bias. The location of different categories of genes for CcI3 has been shown in the Nc/GC3 plot (Figure 4.1.1.1). Ribosomal protein genes are strongly clustered at lower end of the plots indicating a significant strong codon bias in these genes. The CAI

Table 4.1.1: Correlation of CAI with Nc, GC and GC3 values

Organism	Nc/CAI	GC/CAI	GC3/CAI
ANVAR	-0.13	-0.27	-0.28
C7424	0.013	-0.04	-0.015
C7425	0.05	-0.10	-0.11
C8801	-0.02	0.48	0.54
C51142	0.008	-0.008	0.003
CCY0110	0.004	-0.001	0.006
LYN	-0.006	-0.003	-0.02
N7120	-0.09	-0.22	-0.19
N73102	-0.10	-0.17	-0.18
NOSPUM	0.006	-0.02	-0.03
SYNE	0.07	-0.05	-0.07
TRICHOS	-0.06	0.51	0.64
MHUN	0.004	0.56	0.69
MMARI	-0.42	0.45	0.16
MMAZEI	0.06	0.31	0.35
MTHER	-0.27	0.26	0.34
CACET	0.01	-0.007	-0.006
CBEI	-0.04	0.04	-0.05
CCHROM	0.033	0.16	0.19
CLIMI	-0.27	0.39	0.52
CPHAEO	-0.03	0.45	0.65
CPEO1	-0.04	0.19	0.25
CTEPI	-0.001	-0.04	-0.02
CFEROX	-0.20	0.43	0.70
ACN14A	-0.19	0.18	0.23
Ccl3	-0.53	0.32	0.61
EAN1PEC	-0.53	0.38	0.58
MESOBN	-0.46	0.27	0.66
MESOMAF	-0.22	0.18	0.23
SMELILO	-0.45	0.15	0.53
SNOMEDI	-0.10	0.05	0.10
RHODO	-0.28	0.17	0.36
AZOR	0.01	-0.22	-0.07
AZOTO	-0.32	0.34	0.35
AZORHI	-0.22	-0.15	0.31
BEI	0.14	-0.33	-0.19
BRBTAI	-0.20	0.14	0.22
BRUSDA	-0.19	0.13	0.20
BRORS	-0.10	0.02	0.52
RHICFN4	-0.27	0.20	0.29
RHICIAT	-0.54	0.34	0.70
RHIWSM	0.21	0.28	0.43
RHIVICIA	-0.46	0.19	0.58
GLUCO	-0.61	0.17	0.51
KLEB	-0.14	0.08	0.10

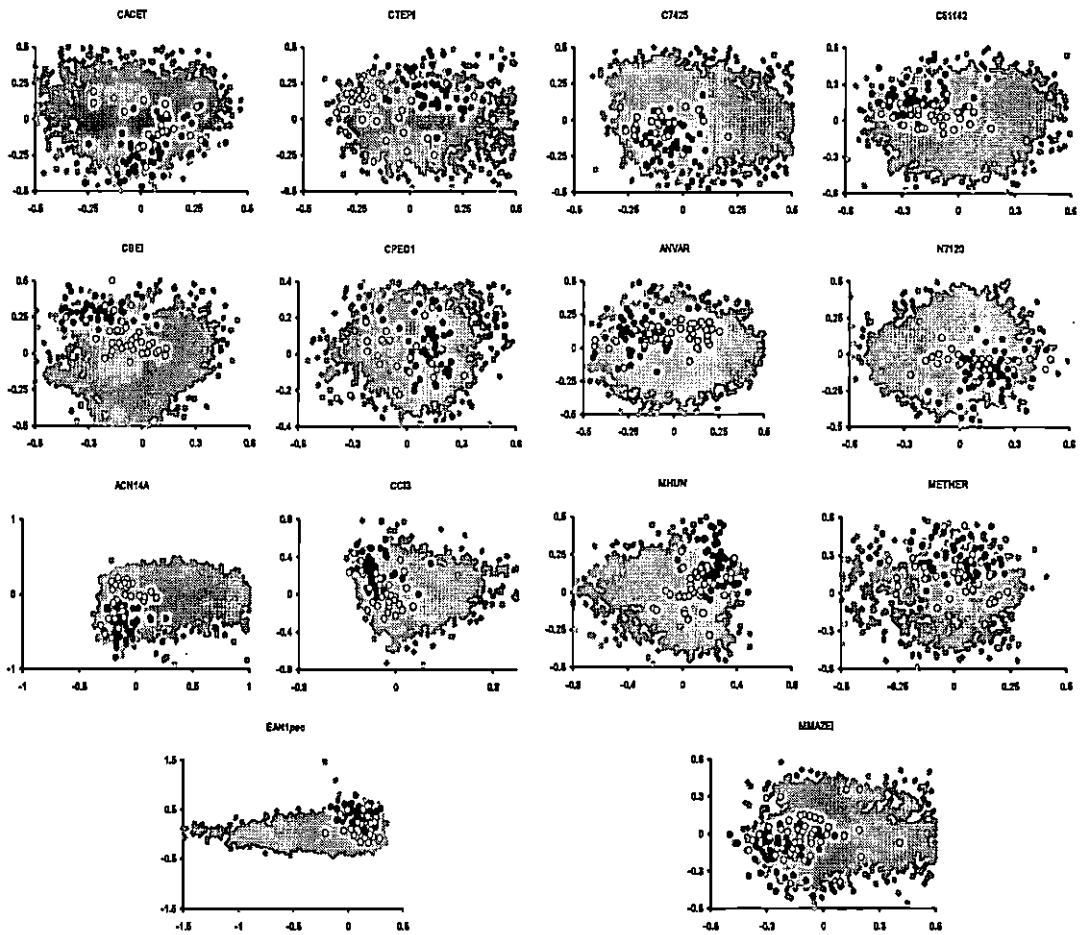

Figure 4.1.2.1: Correspondence analysis of codon count in CACET, CBEI, C7425, C51142, CTEPI, CPEO1, ANVAR, N7120, ACN14A, CCI3, EAN1PEC, MTHER, MHUN and MMAZEI. Colors as per Figure 4.1.1.1. X and Y axis are represented by Axis 1 and Axis 2 respectively

values for *Frankia* genomes are quite high with the RPGs having higher mean CAI values compared to the protein coding genes. There is not much difference in the CBI and Fop values amongst the strains. However, moderate Fop values indicated that other than the mutational pressure other factors may be playing a role in effecting codon usage variation in these genomes. The RPGs showed high negative mean GRAVY values compared to the other category of genes whilst the aromaticity values were more or less similar in amongst the *Frankia* genomes. Z scores revealed minor differences amongst the CAI and GRAVY values but no major difference amongst the other parameters. CAI values did not reveal any significant correlation with Nc and GC content, however *Frankia* CcI3 and EAN1pec revealed moderate positive correlation of CAI with GC3 content ($r=0.58$, $r=0.61$) at ($P<0.001$) level implying that highly expressed genes have higher GC3 content.

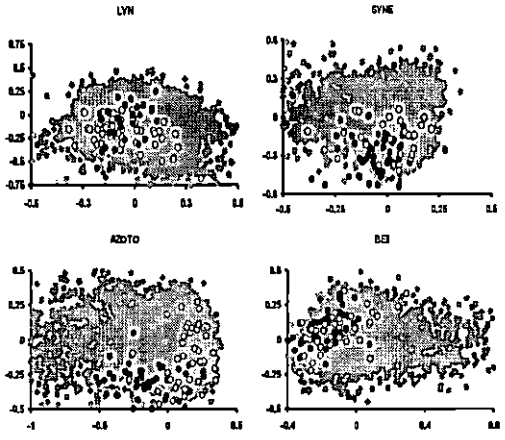

Figure 4.1.2.2: Correspondence analysis of codon count in LYN, SYNE, AZOTO and BEI. Color as per Figure 4.1.1.1. X axis and Y axis are represented by Axis 1 and Axis 2 respectively.

The methanogenic archaeal strains revealed low GC, GC3 content and high Nc values implying low degree of bias. *Methanococcus maripaludis* had much lower GC and GC3 content compared to that of the other studied archaeal strains. The RPGs of *M. maripaludis* had lower Nc values indicating high bias. The codon usage pattern of *M. maripaludis* is some what different in comparison to the other studied methanogens. While others were moderate in their GC content *M. maripaludis* was AT rich. Barring *M. maripaludis* the RPGs of other methanogens remain scattered indicating dissimilar codon bias while in *M. maripaludis* it was clustered at one end of the Nc/GC3 plot (Figure 4.1.1.1). The CAI values for the studied methanogenic genomes were moderate with the exception of RPGs which are highly expressed as expected. Barring *Methanococcus maripaludis* the RPGs for the other strains had lower codon bias. The CBI values are mostly negative in the strains with the exception of few were they are positive but very low. This is an indication that these strains had less optimal codons and are lowly biased. Moderate Fop values for these genomes support the aforesaid fact. RPGs had comparatively higher Fop values than the protein coding genes as expected. RPGs had higher negative GRAVY scores compared to other sets of genes portraying its role in influencing variations in ribosomal protein genes. This difference has been detected by Z scores too. Aromaticity values are more or less similar in most of the strains. Low aromaticity scores signified less number of aromatic amino acids in these genes. It is seen from Table 4.1.1 that with the exception of GC and GC3 content of *Methanospirillium hungatei* the CAI values did not have any noteworthy correlations in the other strains. CAI revealed moderate positive correlations ($r=0.56$, $r=0.61$) with GC and GC3 respectively at $P<0.001$ level demonstrating that highly expressed genes have elevated share of GC/GC3 content.

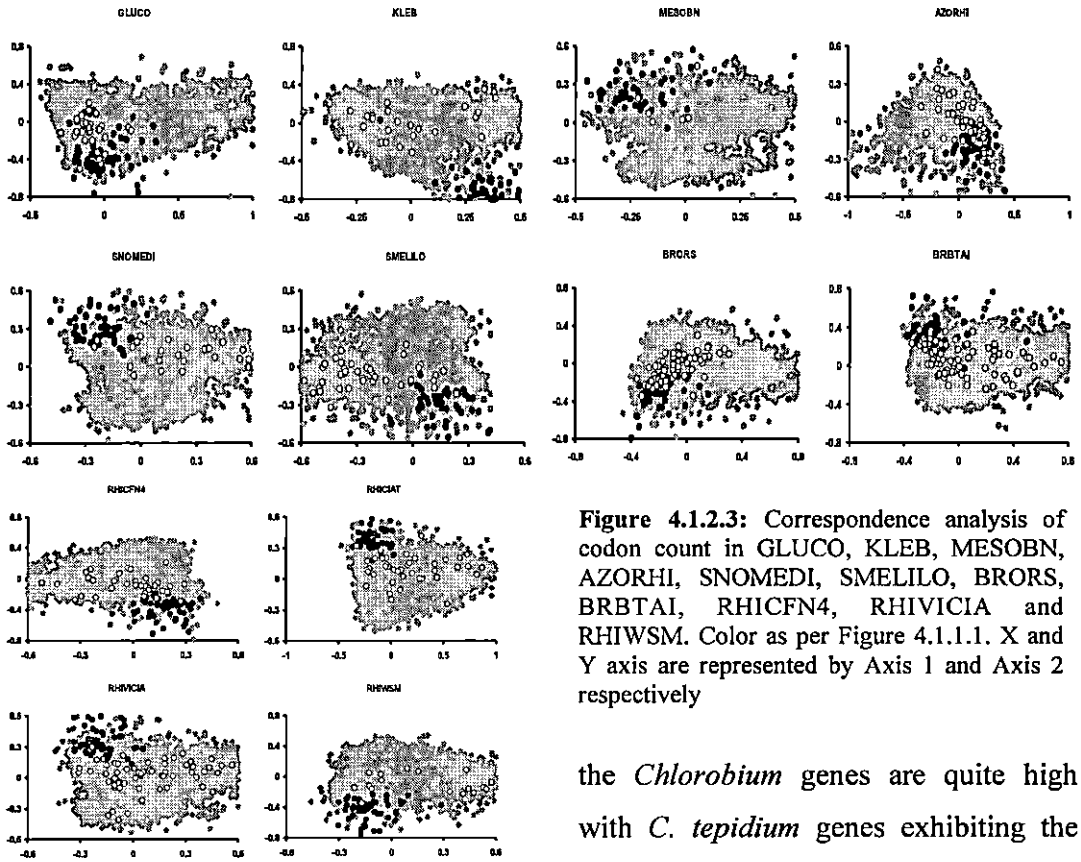

Figure 4.1.2.3: Correspondence analysis of codon count in GLUCO, KLEB, MESOBN, AZORHI, SNOMEDI, SMELILO, BRORS, BRBTAI, RHICFN4, RHIVICIA and RHIWSM. Color as per Figure 4.1.1.1. X and Y axis are represented by Axis 1 and Axis 2 respectively

The *Chlorobium* genomes exhibited wide ranging heterogeneity in codon usage patterns with Nc values ranging from 21±1 to 61±0 in *C. tepidium*, C. *limicola* and 24±2 to 61±0 in the others. The GC, GC3 content are moderate in these organisms. The Nc values are high symbolizing low codon usage bias. Figures 4.1.1.1 illustrated the Nc/GC3 plots for some representative chlorobial strains. Low codon bias is clearly visible. Barring the NFGs of *C. tepidium* none of the NFGs in the other strains had Nc values less than 40 indicating that translational selection may be acting. The expression levels of the *Chlorobium* genes are quite high with *C. tepidium* genes exhibiting the highest expression level amongst the strains. CBI values are very low indicating low codon bias. The Fop values are moderate however; RPGs had higher Fop values indicating higher percentage of optimal codons in these genes. GRAVY scores for RPGs are negative and higher in comparison to other gene categories. This portrays the effect of hydrophilic proteins in influencing codon usage bias. Aromaticity values did not reveal any thing significant. The analysis of z scores entail that some disparity existed in the characteristics of the genes even though they belonged to the same genome. However, no significant

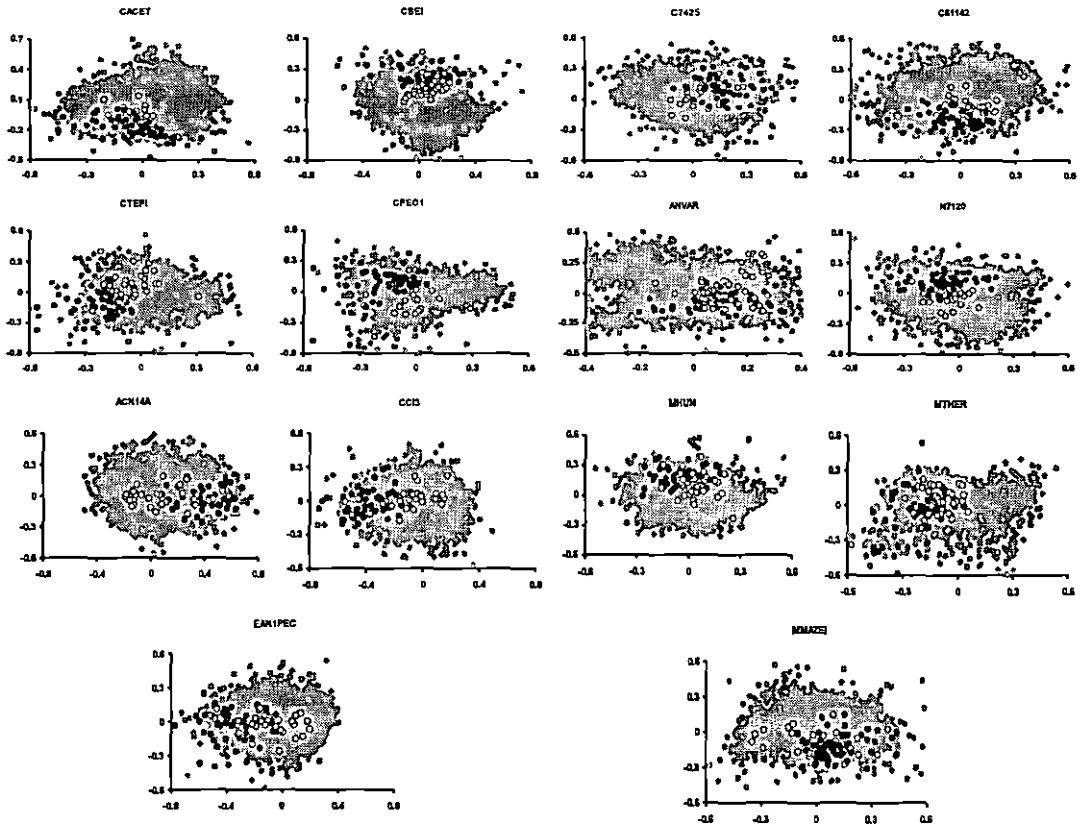

Figure 4.1.2.4: Correspondence analysis of amino acid usage in, CACET, CBEI, CTEPI, CPEO1, ACN14A, CCI3, EAN1PEC, C7425, C51142, ANVAR, N7120, MHUN, MTHER, MMAZEI. Color as per Figure 4.1.1.1. X and Y axis are represented by Axis 1 and Axis 2 respectively

variations were noted for *C. chlorocromatii*. In case of *C. phaeobacteroides* and *C. ferroxidans* CAI values revealed strong positive correlation ($P<0.001$) with GC3 thus entailing its effect on GC3 compositions.

In the cyanobacterial strains, the GC, GC3 and Nc values varied in different species. Over and above the studied cyanobacterial strains were moderate to low GC containing and have high Nc values indicating low codon bias. The GC and GC3 content of *Synechococcus* varied considerably compared to the other cyanobacterial strains. Figures 4.1.1.1 demonstrated the Nc/GC3 plots for some cyanobacterial strains. Barring *Synechococcus* and *Trichodesmium*, ribosomal protein genes for most of the strains remained scattered indicating that they had dissimilar codon bias while in the former two they were clustered at one ends of the plots. The CAI values for cyanobacterial strains revealed high expression levels and varied at the interspecific level. In most cases ribosomal protein genes had higher CAI values as expected. Low CBI and in some cases negative CBI

values indicated low bias as also lower levels of optimal codons. The aforesaid fact has been substantiated by the moderate Fop values. However, in majority of the strains the RPGs had higher Fop values as observed for most of the organisms. The aromaticity values were low in all the strains. However, GRAVY scores for the ribosomal protein genes barring *Anabaena* were comparatively higher than the values obtained for other gene categories indicating the influence of hydrophilic proteins in codon usage bias in these organisms. Z scores revealed minor differences among the different sets of genes. In *Trichodesmium* CAI values showed positive correlations with GC and GC3 content portraying that highly expressed genes had higher GC and GC3 content. The codon usage pattern revealed by *Synechococcus* is comparatively different from the other studied strains. It may due to its ecological niche, biogeographical location and habitat specificities.

It is seen from Table A4.1.1 and Table A4.1.2 (See appendix) that in case of proteobacterial members there is a great deal of variation amongst them with respect to codon usage patterns. As evident from Table A4.1.1 (See appendix), most of the proteobacterial members have moderate GC content with some being GC rich. In these members increase in GC3 content resulted in decrease in Nc values and vice-versa. Heterogeneity in codon usage patterns have been observed in majority of the members. Nc values were moderate in the rhizobial, bradyrhizobial and sinorhizobial members indicating moderate bias. Moderate codon bias associated with most of the protein coding genes may be due to their symbiotic nature. However, the RPGs are more or less highly expressed and substantiates the fact that codons linked to plentiful tRNAs are inclined towards highly expressed genes. The difference in codon bias for the RPGs compared to the protein coding genes is clearly visible from Table A4.1.1 (See appendix). However, in the free-living members the Nc values were low implying strong degree of codon usage bias. The Nc/GC3 plots for some representative proteobacterial members are depicted in Figures 4.1.1.2. In *Azoarcus*, *Rhodobacter*, *Gluconacetobacter*, *Azorhizobium*, *Bradyrhizobium japonicum* USDA the RPGs are clustered at one lower end of the plot and they are highly biased. In other members of rhizobia, bradyrhizobia, sinorhizobia,

mesorhizobia and the free-living *Beijerinckia* and *Klebsiella* the ribosomal protein genes remain moderately clustered. As mentioned earlier the genes having Nc values less than 40 have stronger codon bias are influenced by mutational bias and are potential candidates for whose codon usage is determined by translational efficiency (Peden, 1999). As seen from Table A4.1.2 (See appendix) the CAI values for the proteobacterial members are quite high. This signifies that they are highly expressed. As expected RPGs are more highly expressed compared to other studied categories of genes but in majority of cases the difference is little. As evident from Table A4.1.2 (See appendix) the CBI values in some members are higher than the others. This is an indication of the fact that they are rich in optimal codons and have higher bias. There is a good deal of variation with respect to the Fop values in the proteobacterial members. Fop values are moderate to high amongst the members. However, the Fop values for ribosomal protein genes are higher than the other genes and *Klebsiella* has the highest of 0.76 amongst the studied members. So, ribosomal protein genes have the lion's share of optimal codons as seen in other genomes. Interestingly in majority of the members ribosomal protein genes have high GRAVY scores as observed earlier in most of the other genomes from different lineages signifying that genes associated with the hydrophilic proteins are favored by translationally optimal codons. There is not much difference in the aromaticity values in proteobacterial genomes. The correlation of CAI with Nc values did not reveal any significant result in most of the cases. However, in *Rhizobium etli* CIAT 652 and *Gluconacetobacter* significant ($P<0.001$) negative correlations were obtained. Significant positive correlations were obtained while correlating CAI with GC3 content in *Mesorhizobium* sp. BNC1, *Rhizobium etli* CIAT 652 and R. *leguminosarum bv. viciae 3841*. This implied that variations in GC3 content are manipulated by expression levels in these genomes.

4.1.2 Correspondence analysis of codon and amino acid usages

In clostridial strains, Figures 4.1.2.1 and 4.1.2.4 show the scatter plots for the correspondence analysis (COA) of codon count (CU) and amino- acid frequencies (AAU) for the studied categories of genes. In the clostridia strains majority of the genes were clustered in the centre of the axis of

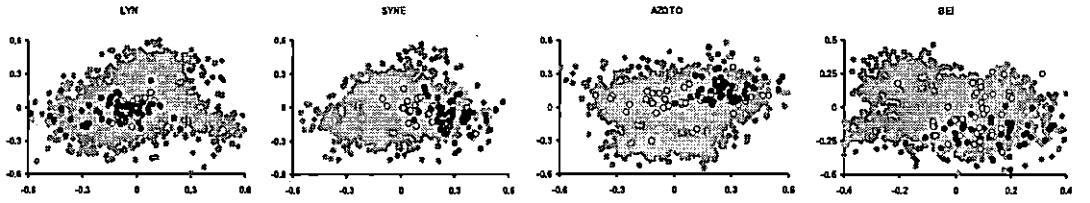

Figure 4.1.2.5: Correspondence analysis of amino acid usage in LYN, SYNE, AZOTO and BEI. Color as per Figure 4.1.1.1. X and Y axis are represented by Axis 1 and Axis 2 respectively

correspondence analysis of codon count within the range -0.3 to +0.3. A few genes were located away from the core block. While in *Clostridium acetobutylicum* ribosomal genes were located in the central core in *C. beijerinckii* they remained clustered at one end. Genes away from the central core included hypothetical proteins, some ribosomal protein genes, transcription factors, redox factors etc. As seen from Table 4.1.2 correlation of Axis 1 of correspondence of codon count for *C. beijerinckii* revealed significant negative correlation ($P<0.001$) with GC content indicating the influence of GC content on codon usage variations while *Clostridium acetobutylicum* did not reveal any important correlation. COA of AAU also revealed positive correlations with aromaticity signifying the role played by aromatic amino acids on amino acid usage.

Correspondence analysis of codon count and amino-acid frequencies for the *Frankia* strains are depicted in Figures 4.1.2.1 and 4.1.2.4 respectively. Codon usage variations in these genomes were not influenced by the leading and lagging strands. In case of COA of CU majority of the genes remained clustered in the core region and the ribosomal protein genes were confined at one ends of the plots. *Frankia* CcI3 revealed an ascending horn where ribosomal protein genes were confined. Genes located away from the core regions included the hypothetical proteins, genes coding for monooxygenase components, amidohydrolase, acyktransferases, cold shock proteins, transposases, transcriptional regulators, redox factors, chaperones, elongation factors etc. It is seen from Table 4.1.2 that the first major axis of variation revealed strong significant correlations ($P<0.001$) with T3, C3, A3, Nc, GC and GC3 in most of the *Frankia* strains. Strong negative as well as positive correlations with GC3 suggest that local variations in GC3 content do have a role in synonymous codon selection. Hence, the strongly

Table 4.1.2 Correlation of the major axis of COA on CU and AAU with different indices

Organism	Axis 1 CU/T3	Axis 1 CU/G3	Axis 1 CU/C3	Axis 1 CU/A3	Axis 1 CU/Nc	Axis 1 CU /GC	Axis 1 CU/ GC3	Axis 1 CU/ CAI	Axis 1 CU/ Fop	Axis 1 AAU / Aromaticity	Axis 1 AAU / Hydropathicity
ANVAR	0.66	-0.24	-0.74	0.64	-0.34	-0.89	-0.84	0.277	-0.520	-0.70	-0.11
C7424	-0.49	0.31	0.41	-0.60	0.19	0.82	0.65	-0.08	0.08	0.72	0.10
C7425	0.68	-0.62	-0.65	0.81	0.35	-0.86	-0.89	0.07	-0.28	-0.17	-0.76
C8801	-0.70	0.37	0.82	-0.72	0.43	0.93	0.90	-0.02	0.44	0.73	-0.003
C51142	0.61	-0.38	-0.78	0.70	-0.47	-0.93	-0.86	0.01	-0.43	0.03	0.73
CCY0110	0.62	-0.35	-0.71	0.66	0.48	-0.86	-0.88	0.06	0.30	0.05	0.36
LYN	0.35	-0.24	-0.37	0.68	-0.14	-0.77	-0.57	0.03	-0.25	-0.13	0.43
N7120	-0.57	0.13	0.69	-0.61	0.30	0.86	0.76	-0.21	0.50	0.61	0.03
N73102	0.62	-0.37	-0.68	0.72	-0.41	-0.92	-0.83	0.18	-0.49	-0.01	-0.59
NOSPUM	0.46	-0.19	-0.57	0.63	-0.31	-0.82	-0.69	0.01	-0.44	-0.08	0.71
SYNE	-0.68	0.55	0.43	-0.78	-0.61	0.77	0.88	-0.08	0.30	-0.26	-0.64
TRICHOS	-0.43	0.24	0.74	-0.54	0.60	0.85	0.77	-0.07	0.46	0.38	-0.32
MHUN	-0.68	0.53	0.75	-0.79	0.001	0.93	0.87	0.55	0.48	-0.01	0.73
MMARI	0.65	0.55	-0.59	-0.24	0.45	-0.76	-0.20	-0.80	-0.64	0.01	0.75
MMAZEI	0.73	-0.38	-0.78	0.63	-0.15	-0.91	-0.85	-0.34	-0.56	-0.82	-0.35
MTHER	-0.59	0.49	0.67	-0.57	-0.38	0.73	0.77	0.26	0.17	0.44	0.72
CACET	-0.20	0. 13	-0.008	-0.01	0.001	0.29	0.27	0.30	0.10	-0.39	-0.02
CBEI	0.34	0.55	0.007	0.01	0.20	-0.69	0.35	-0.07	-0.11	0.66	-0.13
CCHROM	-0.44	0.58	0.48	-0.60	0.13	0.67	0.75	0.13	0.19	0.11	-0.41
CLIMI	0.79	-0.13	-0.84	0.54	0.39	-0.80	-0.90	-0.53	-0.67	-0.14	-0.61
CPHAEO	-0.82	-0.07	0.92	-0.22	-0.07	0.68	0.80	0.77	0.74	-0.06	-0.71
CPEO1	0.81	0.05	-0.91	0.21	0.18	-0.60	0.77	-0.29	-0.71	0.41	0.88
CTEPI	0.83	-0.47	-0.86	0.77	0.53	-0.87	-0.95	0.82	-0.70	0.51	0.84
CFEROX	-0.52	0.17	0.74	-0.67	-0.33	0.68	0.81	0.55	0.51	0.14	0.67
ACN14A	0.86	-0.30	-0.72	0.87	0.84	-0.53	-0.95	-0.23	-0.54	0.41	-0.14
Ccl3	0.71	-0.39	-0.52	0.83	0.82	-0.41	-0.92	-0.62	-0.39	-0.36	0.25
EAN1PEC	-0.84	0.44	0.56	-0.89	-0.90	0.68	0.96	0.59	0.44	-0.46	0.29

Table 4.1.2 Contd. next page

Contd. from previous page

Organism	Axis 1 CU/T3	Axis 1 CU/G3	Axis 1 CU/C3	Axis 1 CU/A3	Axis 1 CU/Nc	Axis 1 CU /GC	Axis 1 CU/ GC3	Axis 1 CU/ CAI	Axis 1 CU/ Fop	Axis 1 AAU / Aromaticity	Axis 1 AAU / Hydropathicity
MESOBN	0.67	-0.40	-0.80	0.80	0.83	-0.55	-0.90	-0.58	-0.55	0.44	0.92
MESOMAF	0.81	-0.43	-0.81	0.84	0.86	-0.61	-0.93	-0.24	-0.66	-0.34	-0.86
SMELILO	-0.74	0.30	0.84	-0.82	-0.86	0.51	0.90	0.50	0.60	-0.40	-0.93
SNOMEDI	0.54	-0.19	-0.79	0.76	0.77	-0.35	-0.80	-0.12	-0.56	-0.41	-0.90
RHODO	0.85	-0.49	-0.58	-0.92	0.87	-0.60	-0.95	-0.33	-0.36	-0.20	0.37
AZOR	0.82	-0.13	-0.76	0.67	0.90	-0.36	-0.89	0.005	-0.72	0.08	-0.35
AZOTO	-0.89	0.50	0.82	-0.87	-0.90	0.76	0.96	0.35	0.81	0.20	-0.08
AZORHI	-0.71	0.18	0.78	-0.83	-0.87	0.22	0.88	0.29	0.51	-0.37	-0.30
BEI	0.48	-0.32	-0.62	0.69	0.66	-0.72	0.66	0.21	-0.45	-0.45	-0.41
BRBTAI	0.71	-0.41	-0.67	0.79	0.79	-0.50	-0.84	-0.22	-0.55	0.21	0.58
BRUSDA	0.38	-0.23	-0.36	0.40	0.41	-0.32	-0.42	-0.37	-0.31	-0.13	-0.27
BRORS	0.77	-0.38	-0.76	0.83	0.86	-0.42	-0.90	-0.11	-0.56	-0.32	-0.80
RHICFN4	-0.67	0.19	0.83	-0.80	-0.8	0.52	0.85	0.31	0.62	-0.44	-0.87
RHICIAT	0.72	-0.24	-0.84	0.84	0.86	-0.59	-0.89	-0.64	-0.65	-0.44	-0.88
RHIWSM	0.56	-0.27	0.71	0.75	0.84	-0.51	0.55	-0.48	-0.63	-0.39	-0.81
RHIVICIA	0.68	-0.21	-0.84	0.82	0.86	-0.58	-0.88	-0.52	-0.66	0.40	0.91
GLUCO	0.40	-0.28	0.65	0.62	0.43	-0.46	-0.71	0.29	0.51	0.13	0.69
KLEB	0.87	-0.70	-0.76	0.91	0.69	-0.88	-0.96	-0.12	-0.60	0.30	0.88

expressed genes should have a high GC content at the synonymous third positions. Strong positive correlation and negative correlation of the Axis 1 with Nc portrays the increase in codon bias among the genes lying towards the left of Axis 1 and the right of Axis 1 respectively. Moderate correlations of Axis 1 with CAI in *Frankia* CcI3 and EAN1pec indicate the influence of expression level on codon usage variations. The role of optimal codon containing genes is also evident from the moderate correlations of Fop with Axis 1 for ACN14a. In case of COA of AAU most of the genes were confined mostly in the range of -0.4-0.4 of the major axes of variations. Aromaticity and GRAVY did not reveal any significant correlation with principal axis of AAU.

In the studied methanogens, COA of CU majority of the genes remained confined in the core region. However, the ribosomal protein genes remained clustered at one end in case of *Methanosarcina mazei* and *Methanospirillium hungatei*. More or less comparable results were obtained in COA of

AAU with the RPGs remaining scattered in *Methanobacterium thermoautotrophicum* signifying their somewhat different codon usage patterns. Table 4.1.2 reveal strong correlation of the principal axis of variation of CU with T3, A3 and C3 in most of the studied methanogens. Again significantly ($P<0.001$) strong positive as well as negative correlations with GC implies the role of GC content in influencing codon selection. Barring *Methanococcus maripaludis* the other strains showed significant ($P<0.001$) correlation with GC3 content thus establishing the fact that local variations of GC3 content influence codon selection. In *Methanococcus maripaludis*, CAI values exposed strong negative correlation with the principal axis of variation thus portraying its role in codon usage bias in *Methanococcus maripaludis*. Moderate correlations were observed for Fop values. The principal axis of COA of AAU in *Methanosarcina mazei* illustrated strong negative correlation thus exposing the part played by aromatic amino acids in amino acid usages. With the exception of *Methanosarcina mazei* all the other strains showed significant ($P<0.001$) positive correlations of GRAVY with the principal axis of variation. This entails that genes associated with hydrophilic proteins are favored by the translationally optimal codons. Leading and lagging strands seldom influenced codon usage variations in these genomes.

Figures 4.1.2.1-4.1.2.4 demonstrated the COA of CU and AAU for some representative strains of *Chlorobium*. Although genes in the leading strand and lagging strands were recognized, there has been a large overlap between the two clouds of genes in the studied genomes. Leading and lagging strands had very little effect on codon usage variations amongst the *Chlorobium* genomes. Most of the genes were confined in the core region with some genes like that of the hypothetical proteins, cold shock proteins, heat shock proteins and translation factors remaining away from the centre of the axis. However, the analysis revealed some interesting trends. The position of the genes on the principal axis of variation of COA of CU showed significant ($P<0.001$) positive and negative correlations (Table 4.1.2) with GC as well as GC3 content suggesting that variations in GC content and synonymous GC3 composition play an important role in codon usage variations and strongly expressed genes have

higher GC content at the synonymous third positions. It is evident from Table 4.1.2 that the position of the genes on the principal axis of variation of COA of CU also revealed correlations with T3, C3 and A3 respectively. The first major axis of variations of COA of CU in *C. phaeobacteroides* DSM 266 and *C. tepidium* also showed significant positive correlations with CAI implying that expression levels of these genes dictate codon usage variations amongst the genes in these genomes. With the exception of *C. chlorocromatii*, the major axis of variations of COA of CU had significant ($P<0.001$) correlations with Fop values indicating that genes containing higher percentage of optimal codons manipulate codon usage patterns. Significant correlations of the major axis of variations of COA of AAU with hydropathicty implies that genes linked with hydrophilic proteins are favored by translationally optimal codons.

The correspondence analysis of the CU and AAU for some representative cyanobacterial strains are depicted in Figures 4.1.2.1-4.1.2.2 and 4.1.2.4-4.1.2.5 respectively. It is evident from the figures that majority of the genes are clustered in the core region for both COA of CU and COA of AAU. Ribosomal protein genes for *Cyanothece sp.* 7425, *Cyanothece sp.* 51142, *Anabaena*, *Nostoc* sp. 7120 and *Synechococcus* are confined at one ends of the plot. Leading strand and lagging strand do not play an important role in codon usage variations in the cyanobacterial strains. The main axis of variation of correspondence analysis for CU showed strong significant correlations ($P<0.001$) with T3, C3 and A3 thus highlighting their importance in codon usage variations. In *Trichodesmium* and *Synechococcus* the major axis of variation of COA for CU showed significant correlations ($P<0.001$) with Nc indicating the role of effective number of codons in these organisms. As observed from Table 4.1.2 significant correlations ($P<0.001$) of the major axis of variation for COA of CU with GC and GC3 content implied the role of compositional bias in effecting codon usage variations in these genomes. Moderate correlations of main axis of variation of correspondence analysis for CU with Fop highlighted the effect of high optimal codon containing genes in influencing codon usage patterns. The principal axis of variation in COA of AAU for *Anabaena*, *Cyanothece* sp. 7424, *Cyanothece* sp. 8801 and Nostoc

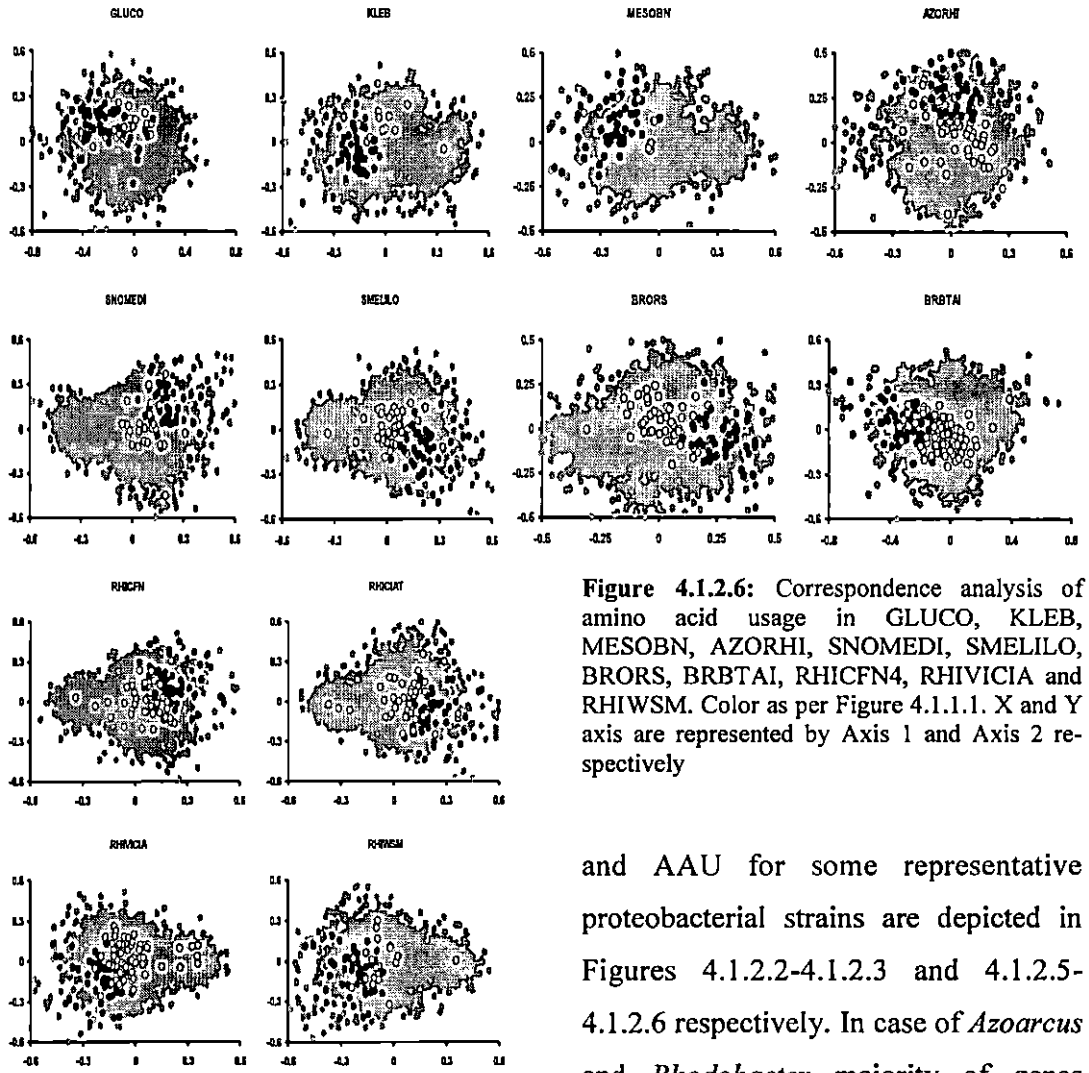

Figure 4.1.2.6: Correspondence analysis of amino acid usage in GLUCO, KLEB, MESOBN, AZORHI, SNOMEDI, SMELILO, BRORS, BRBTAI, RHICFN4, RHIVICIA and RHIWSM. Color as per Figure 4.1.1.1. X and Y axis are represented by Axis 1 and Axis 2 respectively

sp. 7120 showed correlations with aromaticity highlighting that aromatic amino acids influence codon usage variations in these genomes. Barring *Cyanothece sp.* 7425, *Cyanothece sp.* 51142, *Nodularia* and *Synechococcus* principal axis of variation in COA of AAU revealed significant ($P<0.001$) correlations with hydropathicity thus indicating its influence in codon usage variations for these genomes.

The correspondence analysis of the CU and AAU for some representative proteobacterial strains are depicted in Figures 4.1.2.2-4.1.2.3 and 4.1.2.5-4.1.2.6 respectively. In case of *Azoarcus* and *Rhodobacter* majority of genes including the RPGs lie in the core region in COA of CU. However, in COA of AAU RPGs lie away from the core regions. In *Azotobacter* and *Gluconacetobacter* a large number of genes including the RPGs lie in the core region. The ribosomal protein genes in *Azotobacter* and *Gluconacetobacter* and *Klebsiella* lie in the descending horns while in *Beijerinckia* they remain clustered at one end. Clustering of the highly expressed ribosomal protein

genes in most of them indicated similar codon bias and the influence of translational selection. In *Azorhizobium*, *Mesorhizobium sp.* BNC1, *Sinorhizobium meliloti* and *S. medicae* clear cut ascending and descending horns have been noticed for COA of CU as well as AAU in addition to the core region. The highly expressed ribosomal protein genes are located in these horns. The clustering of highly expressed ribosomal genes point out the fact that they share similar codon bias different from rest of the genes of the genomes. These results point out that translational selection is quite strong in these organisms to ward off the influence of mutational bias. Similar types of results were obtained for bradyrhizobial members where there is a core region and the location of ribosomal protein genes in the ascending and descending horns. Genes away from the core regions were hypothetical proteins, translational factors etc. The *Rhizobium* strains too revealed a thick core regions and presence of ascending and descending horns. As observed for bradyrhizobial and sinorhizobial strains the highly expressed RPGs were clustered at the corner of the horns and had similar codon bias. They were also under the influence of translational selection. This particular fact has been previously observed in nitrogen fixing related genes of *Bradyrhizobium* (Sur *et al.*, 2005). Correlation of the principal axis of COA of CU with T3, C3 and A3 in the proteobacterial members revealed significant ($P<0.001$) results entailing the role played by T3, C3 and A3 content in manipulating codon usage variations in proteobacterial members. As seen from Table 4.1.2 proteobacterial members reveal strong positive and negative correlations with Nc values. Negative correlations of the positions of the genes in Axis 1 produced by COA of CU with Nc values in some proteobacterial members depicted in Table 4.1.2 and is attributed to the decrease in codon bias among the genes lying towards the left of Axis 1 while positive correlations indicate the increase in codon bias. While the major axis of COA in CU showed moderate correlations with GC content in most of the proteobacterial members, strong correlations ($P<0.001$) were obtained with GC3 in majority of the strains implying the role of GC3 compositions in influencing codon usage variations across the genomes of these diazotrophs. The major axis of COA in CU for *Mesorhizobium sp.* BNC1 and *Rhizobium etli* CIAT652 revealed

Table 4.1.3 Optimal generation time and tRNA content for the studied organisms

Organism	tRNA Content	Optimal Generation time
Clostridium acetobutylicum ATCC 824	73	1.5 hrs
C. beijerinckii NCIMB 8052	94	0.5 hrs
Frankia alni ACN14a	46	24-48 hrs
Frankia sp. CcI3	46	24-48 hrs
Frankia sp. EAN1pec	47	24-48 hrs
Chlorobium ferrooxidans DSM 13031	46	2 hrs
C. limicola DSM 245	48	2 hrs
C. tepidum TLS	50	2 hrs
C. chlorochromatii CaD3	45	0.5-1 hrs
C. phaeobacteroides DSM 266	47	0.5-1 hrs
C. phaeobacteroides BS1	46	2 hrs
Anabaena variabilis ATCC 29413	48	18-19 or 24 hrs
Cyanothece sp. PCC 7424	44	21-26 hrs
Cyanothece sp. PCC 7425	47	21-26 hrs
Cyanothece sp. PCC 51142	43	21-26 hrs
Cyanothece sp. PCC 8801	43	21-22 hrs
Cyanothece sp. CCY0110	45	21-24 hrs
Nostoc sp. 73102	88	21-22 hrs
Nostoc sp. 7120	70	21-22 hrs
Nodularia spumigena CCY9414	38	5 Days
Lyngbya sp. PCC8106	41	60-270 hrs
Synechococcus sp. JA-2-3B'a(2-13)	45	less than 24 hrs
Trichodesmium erythraeum IMS101	38	2-4 Days
Sinorhizobium meliloti 1021	55	1.5-2 hrs
S. medicae	53	1.5-2 hrs
Mesorhizobium sp. BNC1	52	4-15 hrs
M. loti MAFF303099	53	4.5 hrs
Azotobacter vinelandii AvOP	64	2.4 hrs
Azoarcus sp. BH72	56	1.8 hrs
Azorhizobium caulinodans ORS 571	53	1.5 hrs
Bradyrhizobium japonicum USDA 110	51	7.7-9.9hrs
Bradyrhizobium sp. ORS278	50	7.7-9.9hrs
Bradyrhizobium sp. BTAI1	52	7.7-9.9hrs
Klebsiella pneumoniae pneumoniae MGH78578	86	40mins-1hr
Gluconacetobacter diazotrophicus PAL5	55	1 hr
Beijerinckia indica subsp. indica ATCC 9039	52	13 hrs
Rhizobium etli CFN 42	50	1-1.3 hrs
R. etli CIAT 652	51	1-1.3 hrs
R. leguminosarum bv. trifolii WSM1325	53	1-1.3 hrs
R. leguminosarum bv. viciae 3841	52	1-1.3 hrs
Rhodobacter sphaeroides 2.4.1	54	1.6-2.1 hrs
Methanospirillum hungatei JF-1	51	2-3 Days
Methanococcus maripaludis strain S2	38	2-3 Days
Methanosarcina mazei strain Goe1	57	7.7 hrs
Methanobacterium thermoautotrophicum str. Delta H	39	3 Days

negative correlations with CAI thereby indicating the role of gene expression level in these cases. On correlating the Fop values with major axis of COA in CU significant results were obtained in majority of the members. This points out that higher percentage of optimal codons play a vital role in effecting codon usage variation amongst the proteobacterial genomes. While aromaticity values showed less significant correlation with the major axis of variation of COA of AAU in the studied members strong positive and negative correlation with the hydropathicity i.e., GRAVY indicated the influence of hydrophilic and hydrophobic amino acid residues in synonymous codon usage variations in proteobacterial members.

4.1.3 Codon usage bias and tRNA content

Rocha (2004) in his classic work correlated the codon usage bias with the tRNA content in bacterial genomes. The tRNA content and the optimal generation times for the genomes of studied organisms are depicted in Table 4.1.3. As per the observations of Rocha (2004) fast growers have a median of 61 tRNA genes compared to 44 for the slow growers and the former tend to have stronger codon bias compared to the latter. The optimal generation times for the clostridial genomes were obtained from personal communications with Eduardo Rocha. From the table it is clear that these organisms are fast growers and as per Rocha's observations they had high number of tRNA and they had a small number of anticodons implying that they possessed less diverse tRNAs. So, they are specialized to use a small set of anticodons while maintaining a high number of tRNA. Codon bias for the ribosomal protein genes for these organisms were also high signifying that codons associated with high tRNAs have an inclination to be overrepresented in highly expressed genes like the ribosomal protein genes.

In the actinobacterial *Frankia* strains low number of tRNAs were expected since they are slow growers. However, all the three strains had 43 unique anticodon tRNA genes. The optimal generation time for these genomes was obtained from personal communications with Prof. Louis Tisa. Genome reduction in obligate symbionts and obligate pathogens is well known and they show low codon bias and a decreased number of tRNA genes. Strain CcI3 showed a slightly reduced codon bias compared to the other two

strains. However, genome reduction in CcI3 did not result in a reduction of the total tRNA genes (Sen *et al.*, 2008).

The studied methanogenic archaeal strains showed some variations with respect to tRNA copy numbers. The generation times for these strains were obtained from Eduardo Rocha. All of the studied strains are slow growers but *Methanosarcina mazei* had a comparatively high tRNA copy number. Over and above barring *Methanosarcina mazei* the findings are in line with that of Rocha (2004) regarding the nature of tRNA and their genomes. High copy number in *Methanosarcina mazei* probably assisted in the translation mechanism of the organism in its environment.

In *Chlorobium* strains, the optimal generation times were obtained from personal communications with researchers working in the field of *Chlorobium* biology. An interesting fact is that inspite of them being fast growers they had a much lower tRNA content. This is unlike other organisms 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, they are well equipped to use a small set of anticodons despite maintaining moderate number of tRNAs. The ribosomal protein genes of these strains showed moderate codon bias. The translation machinery of *Chlorobium* probably did not co-evolve with higher codon usage in ribosomal protein genes even though they are fast growers. This may be one of the reasons why *Chlorobium* genes did not show strong codon usage bias (Sur *et al.*, 2008).

The slow growing *Cyanothece* strains supported the observations of Rocha (2004) with respect to the tRNA copy number. As expected they had low codon bias. Comparatively the *Nostoc* strains had much higher tRNA copy number inspite of them being slow growers. Interestingly, they had a low codon bias. However, the ribosomal protein genes for these organisms have high CAI values. The most probable explanation for this phenomenon may be that the translation apparatus possibly evolved with low codon bias in the highly expressed genes like ribosomal protein genes for *Nostoc*. In case of other cyanobacterial strains like *Nodularia*, *Anabaena*, *Trichodesmium*, *Lynbya* and *Synechococcus* the low tRNA copy number is commensurate with the optimal generation time for the

organisms. All the information regarding the generation times for the cyanobacterial strains were obtained from personal communications with E. Rocha.

In case of sinorhizobial and mesorhizobial members the results are more or less in line with Rocha's (2004) observations. The results for the fast growing strains of *Rhizobium* support the observations of Rocha (2004) and the explanation is same like that of sinorhizobial and mesorhizobial members. However, the tRNA copy number is slightly lower than the median for fast growers. In the slow growing bradyrhizobial members the tRNA copy number is slightly higher than the median for slow growers and the highly expressed ribosomal protein genes show elevated level of codon bias probably owing to the co-evolution of the translational apparatus with high codon bias in them. Translational efficiency plays a significant role in influencing codon usage variation in the genomes of bradyrhizobia. In other proteobacterial members like *Azoarcus*, *Azorhizobium*, *Rhodobacter*, *Gluconoacetobacter*, *Azotobacter*, *Beijerinckia* and *Klebsiella* the results support the observations by Rocha (2004). However, *Klebsiella* had huge number of tRNAs while maintaining small set of anticodons. The ribosomal protein genes for this organism also showed higher codon bias that varied widely from the protein coding genes of the organism and translational efficiency plays a vital role in effecting codon usage variation. The information on generation times for proteobacterial members were obtained from E. Rocha (personal communication).

4.1.4 Identification of potentially highly expressed genes

Wu *et al.,* (2005) analysed the proteome results and validated the correlation between CAI values and expression levels showing experimentally that CAI predicted potentially highly expressed genes are indeed highly expressed. As per Wu *et al.,* (2005, 2005a), the top 10% of the genes in terms of CAI values were classified as potentially highly expressed (PHX). The top 5 PHX genes for each of the studied organisms are represented in Table A4.1.4 (See appendix).

In the clostridial strains the corresponding CAI cut-off was 0.795 and 0.71 for *Clostridium acetobutylicum* and *C. beijerinckii* respectively. *C. acetobutylicum* had 383 PHX genes including 30 ribosomal protein genes. *C. beijerinckii* had 508 PHX genes

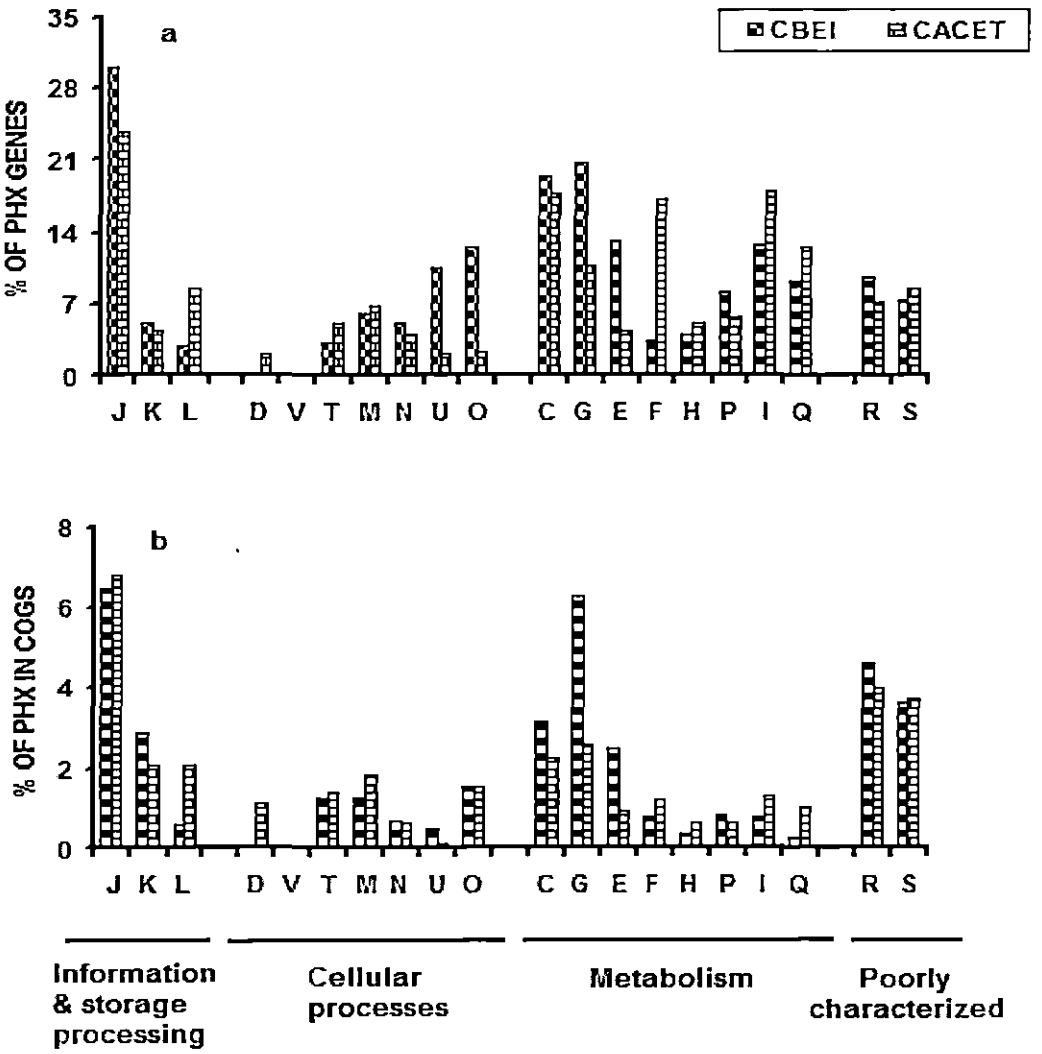

Figure 4.1.5.1: Distribution of potentially highly expressed genes in clostridial strains within COG functional groups

including 39 ribosomal protein genes. The CAI cut-off for the *Frankia* strains CcI3, ACN14a and EAN1pec were 0.795, 0.842 and 0.832. CcI3, ACN14a and EAN1pec had 451, 671 and 717 PHX genes that included 6, 12 and 45 ribosomal protein genes respectively. In the methanogens, the CAI cut-off for *Methanospirillum hungatei*, *Methanococcus maripaludis*, *Methanosarcina mazei* and *Methanobacterium thermoautotrophicum* were 0.639, 0.615, 0.668 and 0.600 correspondingly. *Methanospirillum hungatei*, *Methanococcus maripaludis*, *Methanosarcina mazei* and *Methanobacterium thermoautotrophicum* had 323, 173, 338 and 185 PHX genes incorporating 19, 51, 41 and 17 ribosomal protein genes respectively. The CAI cut-off for the

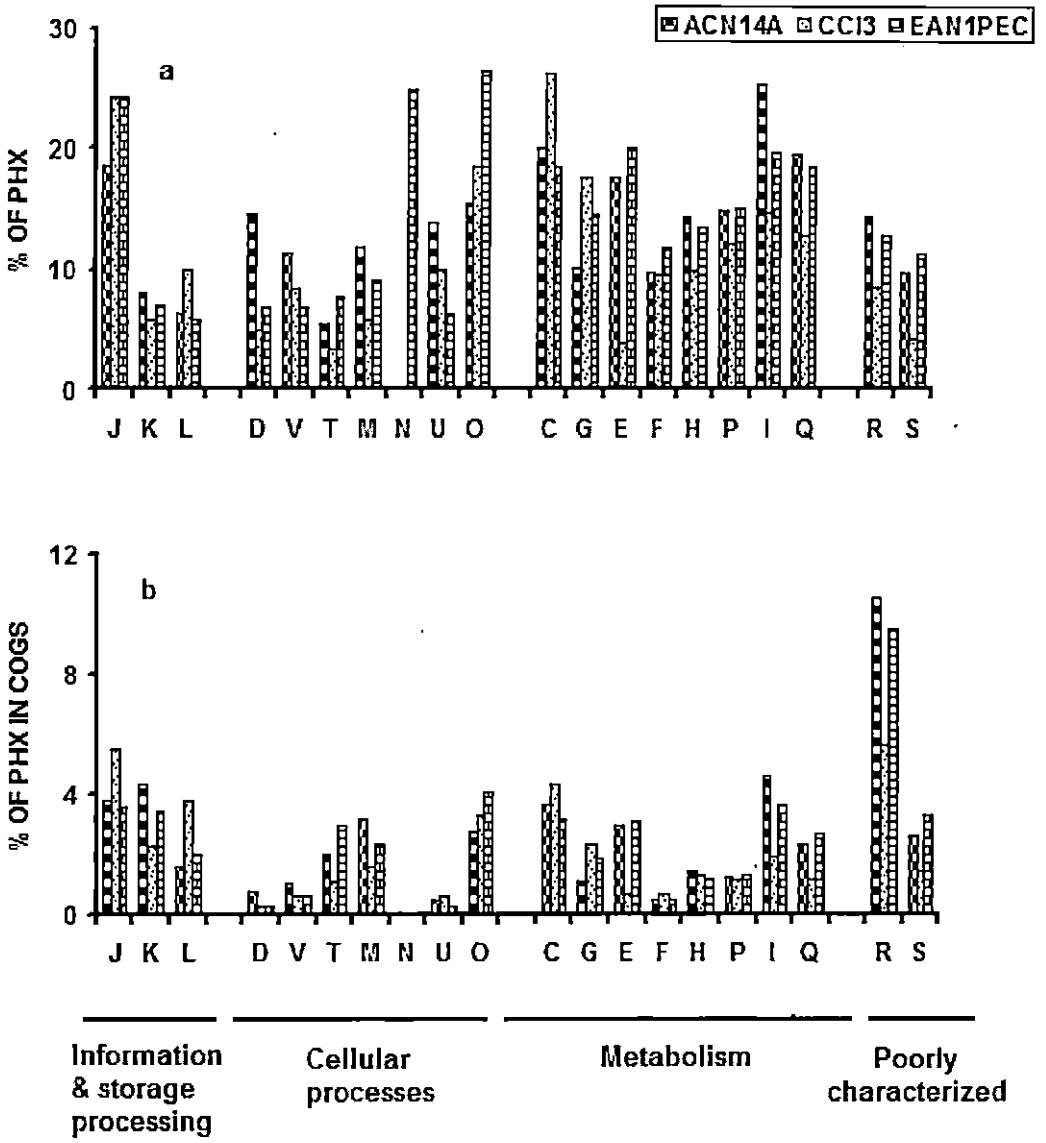

Figure 4.1.5.2: Distribution of potentially highly expressed genes in *Frankia* strains within COG functional groups

chlorobial genomes were 0.764, 0.724, 0.739, 0.722, 0.78 and 0.634 for *Chlorobium limicola*, *C. phaeobacteroides* DSM 266 , *C. phaeobacteroides* BS1, C. *ferredoxidans*, *C. tepidium* and C. *chlorocromatii* respectively. On the basis of this, *Chlorobium limicola*, *C. phaeobacteroides* DSM 266 , *C. phaeobacteroides* BS1, C. *ferredoxidans*, *C. tepidium* and C. *chlorocromatii*had 251, 274, 255, 214, 221 and 250 PHX genes. The number of ribosomal protein genes in the PHX was 4, 6, 5, 5, 8 and 12 for the chlorobial strains.

In case of the cyanobacterial strains, CAI cut-off was 0.807 and 0.809 for the

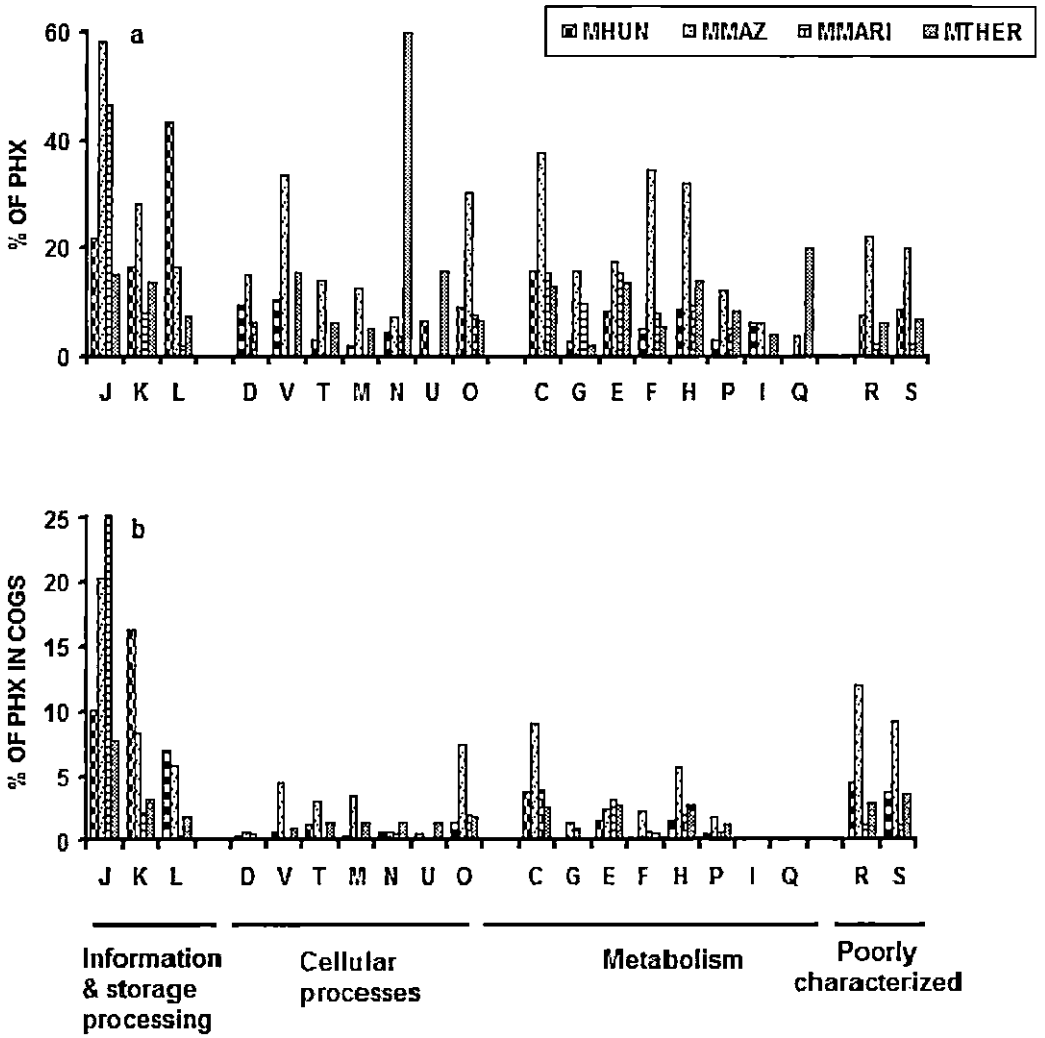

Figure 4.1.5.3: Distribution of potentially highly expressed genes in methanogens within COG functional groups

Nostoc strains N7120 and N73102. N7120 had 612 PHX genes including 5 ribosomal protein genes while N73102 had 667 PHX genes that housed 5 ribosomal protein genes. In case of *Synechococcus*, *Trichodesmium*, *Nodularia* and *Lyngbya* the CAI cut-offs were 0.692, 0.842, 0.809 and 0.796. On the basis of this *Synechococcus*, *Trichodesmium*, *Nodularia* and *Lyngbya* had 286, 506, 481 and 608 PHX genes. *Synechococcus*, *Trichodesmium*, *Nodularia* and *Lyngbya* had 1, 11, 4 and 3 ribosomal protein genes in the PHX category. In the *Cyanothece* strains the CAI cut-offs were 0.800, 0.739, 0.799, 0.814 and 0.829 for C7424, C7425, C8801, C51142 and CCY0110 respectively. This accounted for 585, 540, 452, 522 and 595 PHX genes in the aforesaid strains. C7424, C7425, C8801, C51142 and CCY0110 had 4, 0, 3, 2

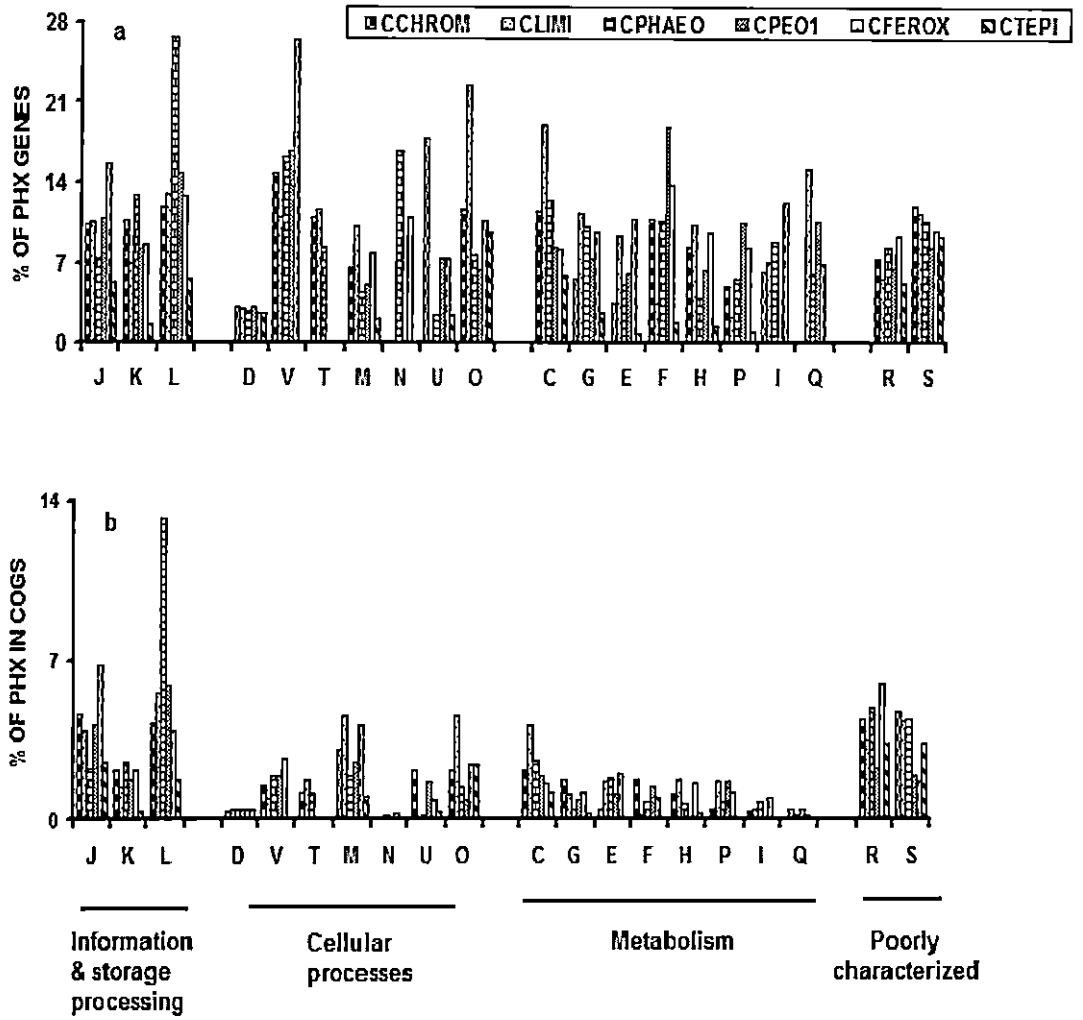

Figure 4.1.5.4: Distribution of potentially highly expressed genes in chlorobial strains within COG functional groups

and 2 ribosomal protein genes in the PHX. In *Anabaena* the CAI cut-off was 0.807. It resulted in having 568 PHX genes including 4 ribosomal protein genes.

The CAI cut-off for the proteobacterial members like *Azotobacter*, *Beijerinckia*, *Klebsiella* and *Azoarcus* were 0.741, 0.831, 0.758 and 0.846 correspondingly. *Azotobacter*, *Beijerinckia*, *Klebsiella* and *Azoarcus* had 498 including 7, 378 including 13, 518 including 42 and 397 including 22 ribosomal protein genes in the PHX category. In the *Rhizobium* strains the CAI cut-offs corresponded to 0.789, 0.750, 0.749 and 0.793 in RHIZCFN4, RHIVICIA, RHIWSM and RHICIAT respectively. This accounted for 603 including 9, 726 including 13, 372 including 22 and 605 including 6 ribosomal protein genes in the PHX category. Among the bradyrhizobial and

sinorhizobial members the CAI cut-off corresponded to 0.766, 0.817, 0.73, 0.741 and 0.755 for *Bradyrhizobium sp. ORS278*, *Bradyrhizobium sp. BTAII*, *Bradyrhizobium japonicum*, *Sinorhizobium medicae* and *Sinorhizobium meliloti* correspondingly. On the basis of this 673 including 11, 770 including 7, 830 including 15, 650 including 11 and 621 including 11 ribosomal protein genes were obtained for *Bradyrhizobium sp. ORS278*, *Bradyrhizobium sp. BTAII*, *Bradyrhizobium japonicum*, *Sinorhizobium medicae* and *Sinorhizobium meliloti*. In case of the mesorhizobial members like *Mesorhizobium* sp. BNC1 and *Mesorhizobium loti* the CAI cut-offs were 0.779 and 0.808 respectively. This resulted in the detection of 461 comprising 10 and 725 comprising of 6 ribosomal protein genes in the PHX category. In the other members like *Rhodobacter*, *Azorhizobium* and *Gluconacetobacter* the CAI cut-off corresponded to 0.764, 0.746 and 0.733 correspondingly. This resulted in the prediction of 431 including 1, 470 including 6 and 392 including 6 ribosomal protein genes for the PHX genes category in these organisms.

4.1.5 Functional analysis of potentially highly expressed genes

In the clostridial strains, Figure 4.1.5.1 shows the percentage of potentially highly expressed genes in various COG categories. The pattern of distribution is quite similar in both the strains. COG functional group 3 (Metabolism) had the most number of PHX genes. The major COG categories were J, G, R and S. When the percentage of total PHX genes within the COG group (Figure 4.1.5.1b) was considered it was observed that *Clostridium beijerinckii* had higher share compared to that of *C. acetobutylicum*. Table A4.1.4 (See appendix) depicts the top 5 PHX genes in COGs for the studied genomes. The elevated levels of PHX in the translation and carbohydrate transport and metabolism COGs for these clostridial strains probably assists in their survival in soil environment and carry out carbohydrate related metabolism more efficiently.

Figure 4.1.5.2a shows the percentage of potentially highly expressed genes for *Frankia* strains in various COG categories clustered into four COG functional groups. As expected, metabolism group towered over others for the three *Frankia* genomes. However, CcI3 had a different distribution pattern compared to

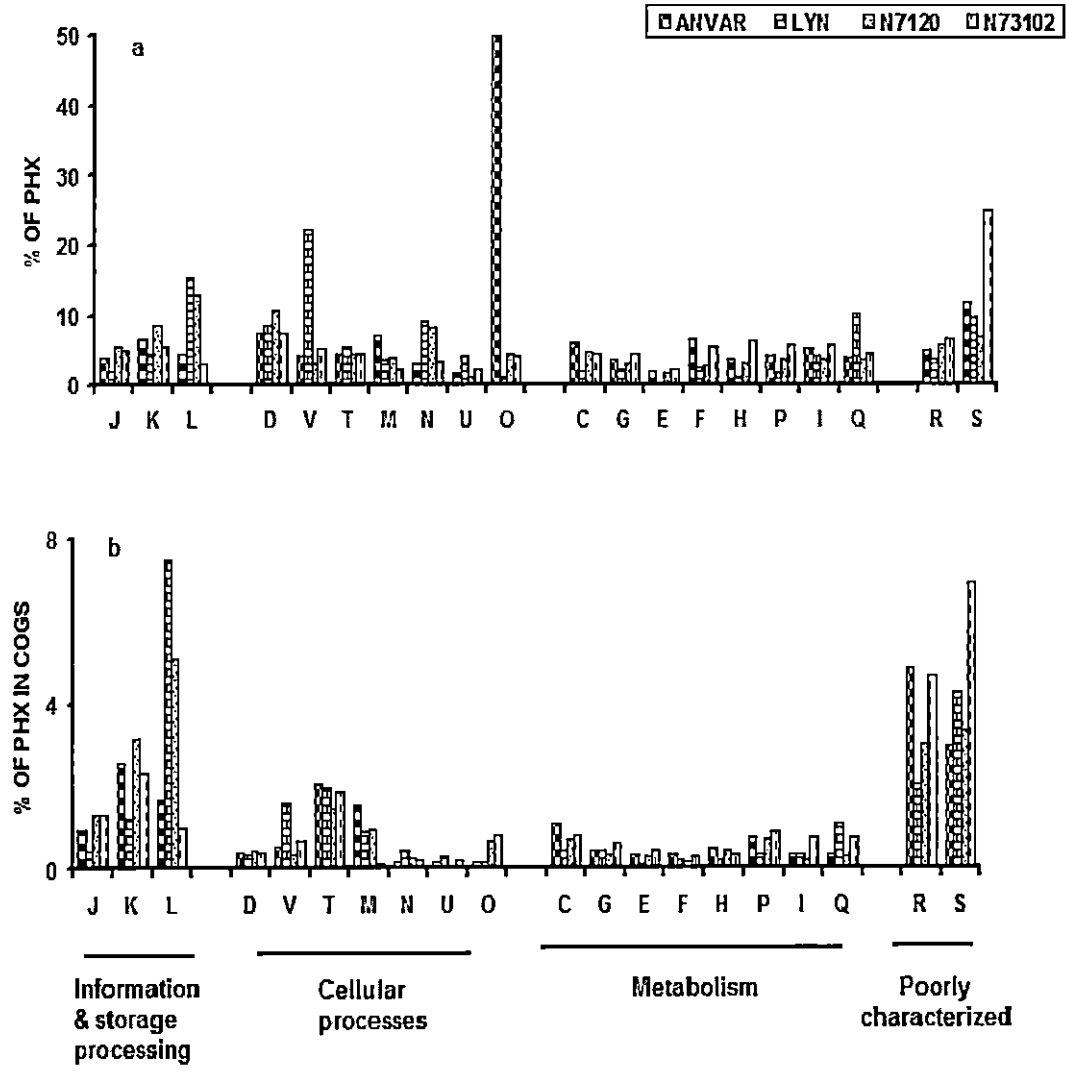

Figure 4.1.5.5: Distribution of potentially highly expressed genes in the cyanobacterial members AN-VAR, LYN, N7120 and N73102 within COG functional groups

ACN14a and EAN1pec, which were similar in patterns and numbers. *Frankia* CcI3 had higher percentage of PHX in the Information and storage processing COG group. The top COG categories for ACN14a and EAN1pec were R, I, C, E and Q, while in CcI3 it was J, L, O, C and G. Analysis of the distribution pattern of PHX genes based on percentages within the COGs shows another perspective. The overall distribution pattern was different although metabolism contained the highest percentage of PHX genes. CcI3 had more PHX genes in the Information and storage processing category and fewer in categories like I, Q, P and R compared to *Frankia* EAN1pec and ACN14a. EAN1pec had more PHX genes in K, T and O. These provide some insight on the genes required for the lifestyles and habitats for the three

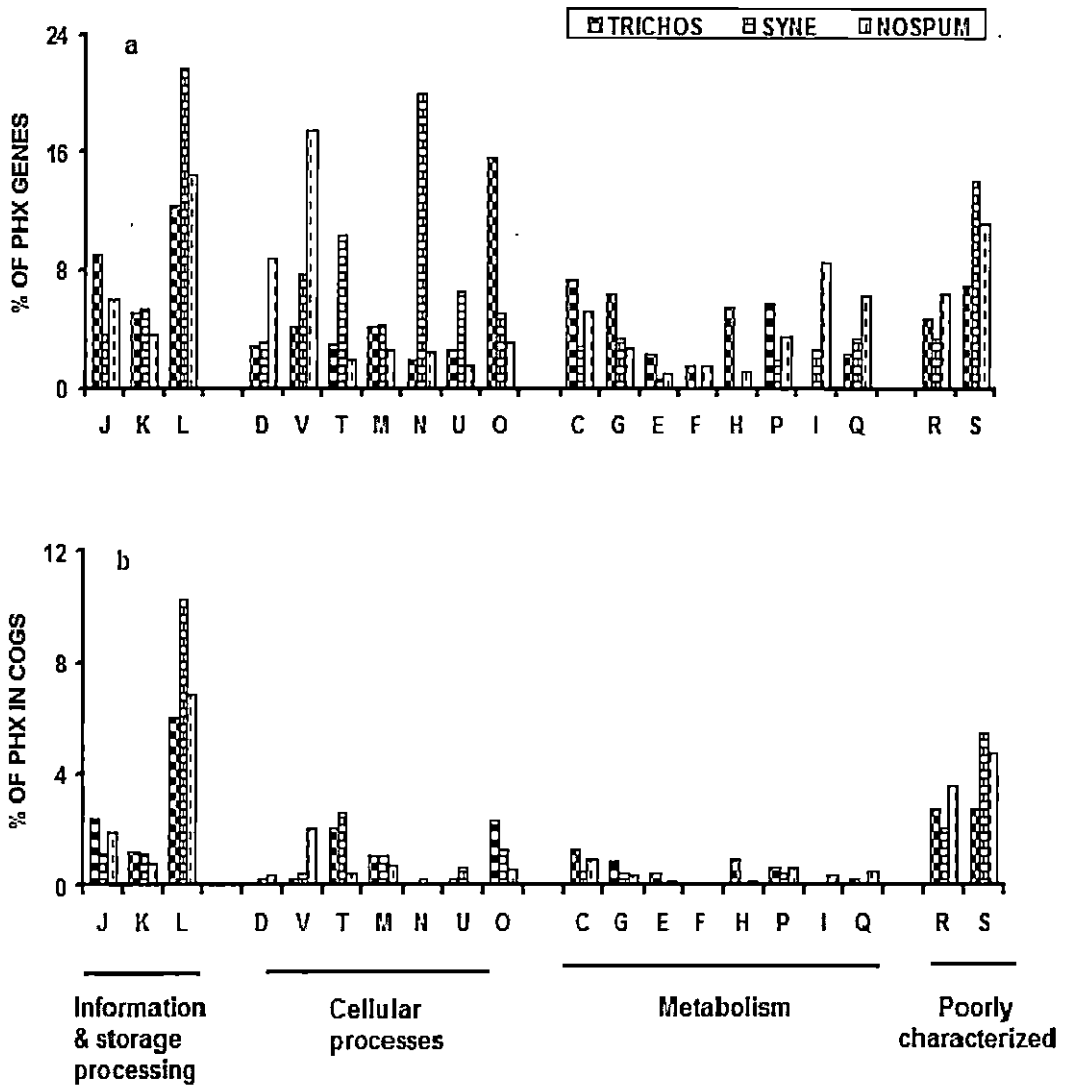

Figure 4.1.5.6: Distribution of potentially highly expressed genes in the cyanobacterial members TRICHOS, SYNE and NOSPUM within COG functional groups

Frankia strains. Multiple lines of evidence indicate that *Frankia* CcI3 and other *Casuarina* strains evolved as specialist (Normand *et al.,* 2007). Besides having a reduced genome comparable to obligate symbionts CcI3 have a narrow host range and are absent in soils outside the scope of their host plants. But *Frankia* ACN14a and EAN1pec have wider host plant ranges and are found world-wide and also without host plants. Increased number of PHX genes in several COG categories may aid in the ability of *Frankia* ACN14a and EAN1pec to survive and compete in the soil outside the host plant. Increased levels of PHX genes in K and T for *Frankia* EAN1pec would be advantageous in soil environment by increasing its ability to respond to signals and regulate gene expression.

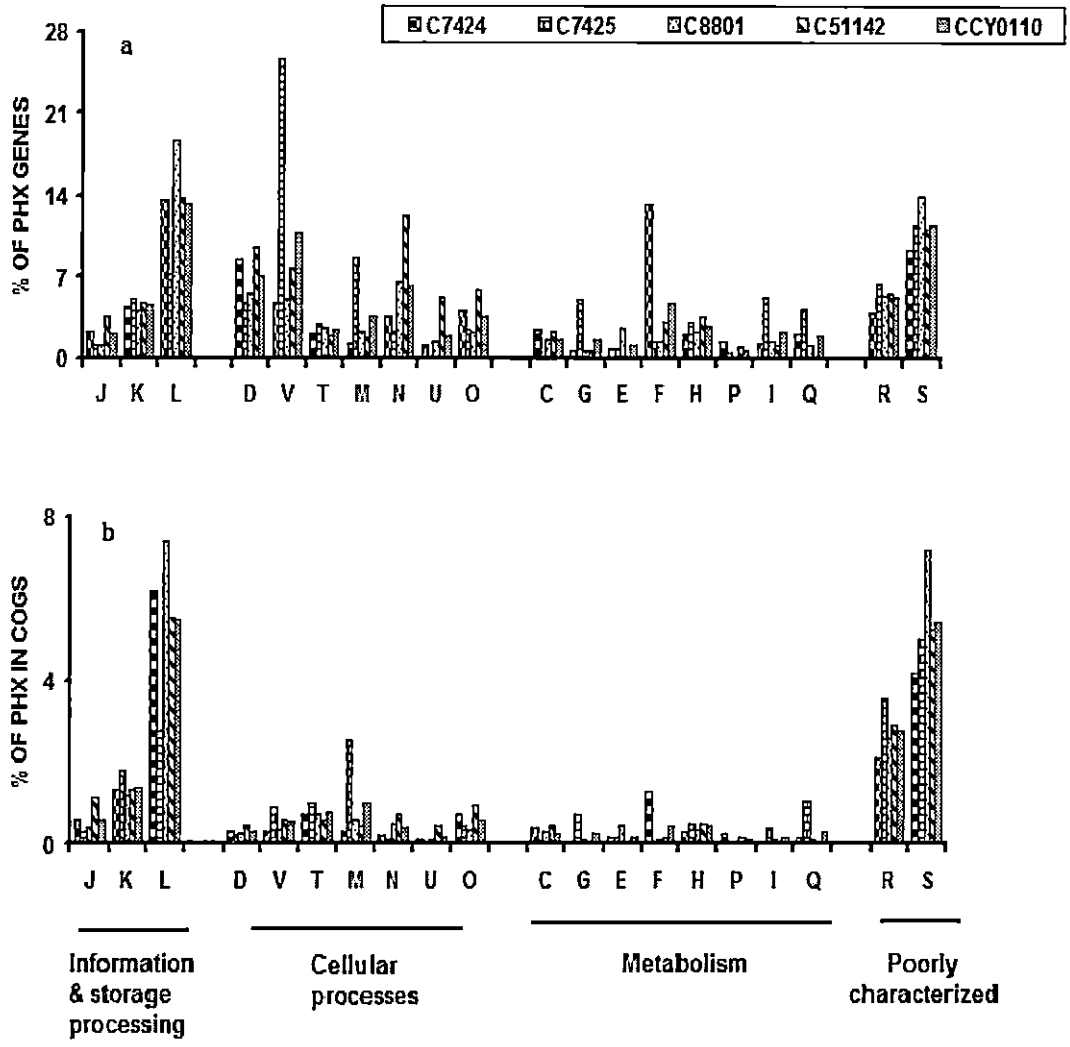

Figure 4.1.5.7: Distribution of potentially highly expressed genes in the cyanobacterial members C7424, C7425, C8801, C51142 and CCY0110 within COG functional groups

In the methanogens, as revealed in Figure 4.1.5.3 Information and storage processing COG group appears to be the most important group housing a large number of PHX genes. *Methanosarcina mazei* showed some difference with respect to other strains with regards to its distribution in metabolism categories as also R and S categories. A wide diversity is noted with respect to the distribution of PHX genes in COGs. However, J, K. C, R and S appear to be the leading categories. Elevated levels of PHX in COGs for translation (J), transcription (K), energy production and conversion (C) and general function prediction (R) throw light upon their lifestyle in sewages and sludges and carry out day to day activities as well as fix nitrogen.

Figure 4.1.5.4 demonstrates the allotment of the PHX into each COG

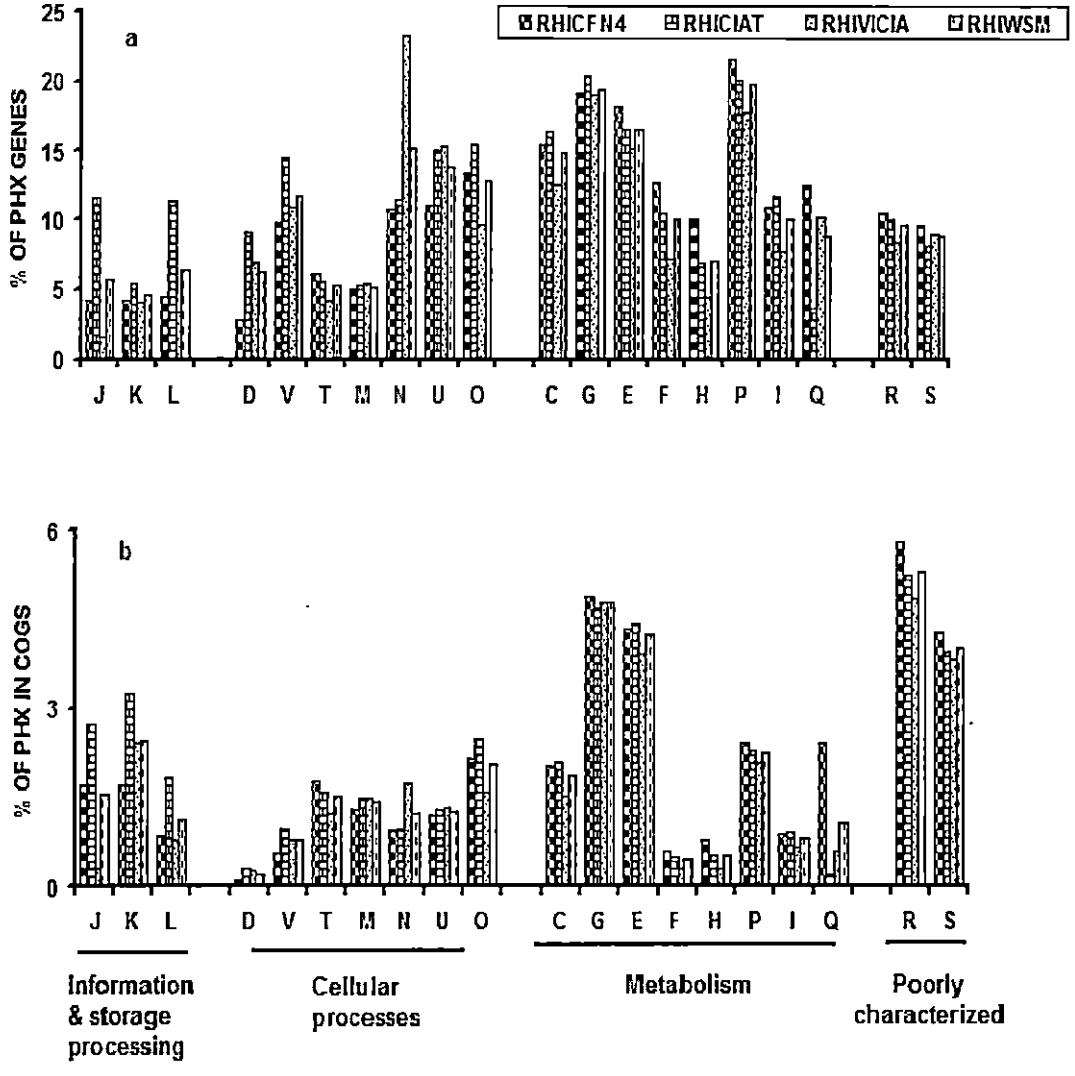

Figure 4.1.5.8: Distribution of potentially highly expressed genes in the proteobacterial members RHICFN4, RHICIAT, RHIVICIA and RHIWSM within COG functional groups

category for the chlorobial strains. The analysis of the distribution of PHX genes based on percentages within the COG categories for these genomes showed some difference with respect to the allotment in individual COGs. When individual COGs were considered it was seen that the metabolism and Information storage processing COG group had the maximum share. The top five COG categories for the *Chlorobium* strains were: translation, replication recombination repair, defense mechanisms, post translational modifications, energy production and conversion. All these provide significant insights into the genes required for maintaining the lifestyle of the *Chlorobium* genomes in different habitats. Huge number of PHX genes in COGs aid in the capability of *C. tepidium* to subsist in extreme

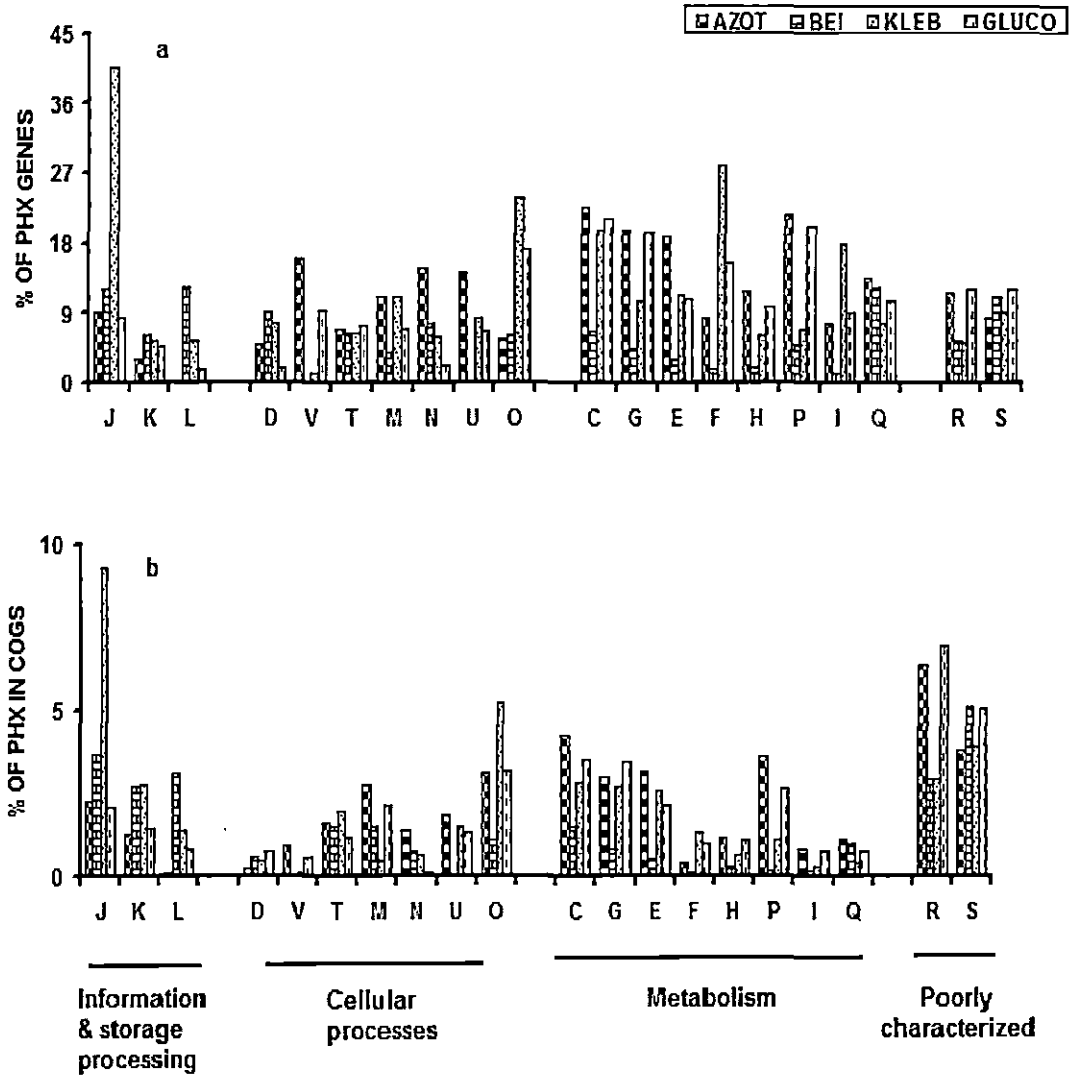

Figure 4.1.5.9: Distribution of potentially highly expressed genes in the proteobacterial members AZOTO, BEI, KLEB and GLUCO within COG functional groups

environments like hot springs and carry out essential life processes particularly nitrogen fixation. Increased presence of PHX genes in COGs like translation, energy production and metabolism, general function prediction probably help chlorobial strains surviving in anoxic, stressed habitats in competing with other bacteria. *C. chlorocromatii* exists as symbionts in phototrophic consortium (Sur *et al.,* 2008). Higher amount of PHX genes in translation, replication recombination repair and cellular process enhance their ability to coordinate cell division and interspecific association between partners and regulate gene expression while living in such an association. The difference in the functional categorization amongst the chlorobial strains may also be attributed to the biogeographical distribution of the strains.

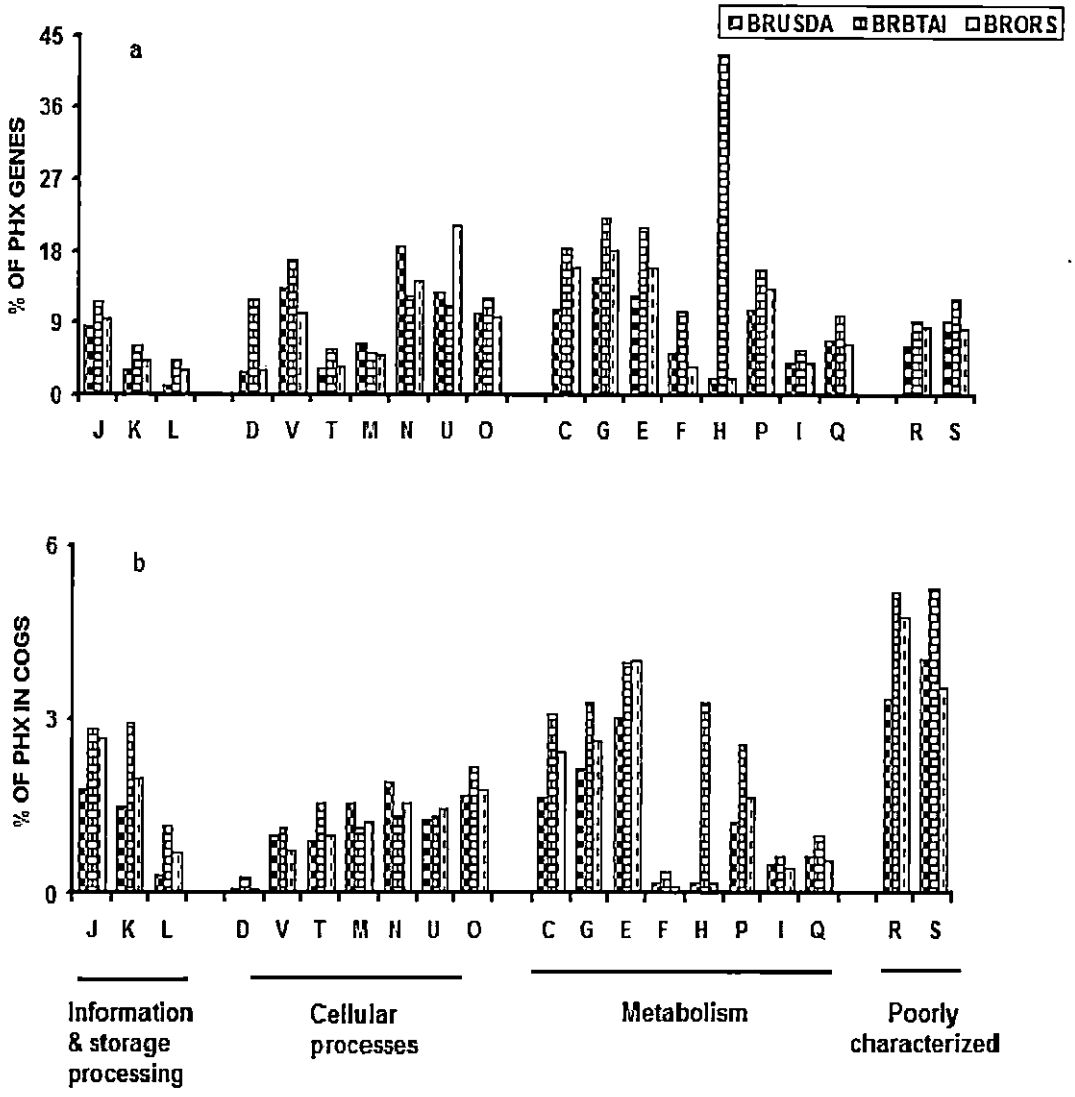

Figure 4.1.5.10: Distribution of potentially highly expressed genes in the proteobacterial members BRUSDA, BRBTAI and BRORS within COG functional groups

Figures 4.1.5.5-4.1.5.7 demonstrated the allocation of PHX genes in the different COG functional groups for the cyanobacterial strains. Some variation was noticed among the strains coming from different habitats and ecological niches. Amongst the fresh water dwelling strains like *Anabaena*, *Lynbya*, *Nostoc* strains 7102 and 73102 the categories, L, V, R and S showed highest percentage. Amongst the strains *Trichodesmium*, *Synechococcus* and *Nodularia* the categories L, V, O, R and S towered over others with respect to the percentages of PHX in COGS. In the *Cyanothece* strains, 7424, 7425, 8801, 51142 and CY0110, it was observed that L, M, F, R and S were the major COG categories. All these provide some information about the nature of lifestyle of the cyanobacterial strains. Most of these strains survive in

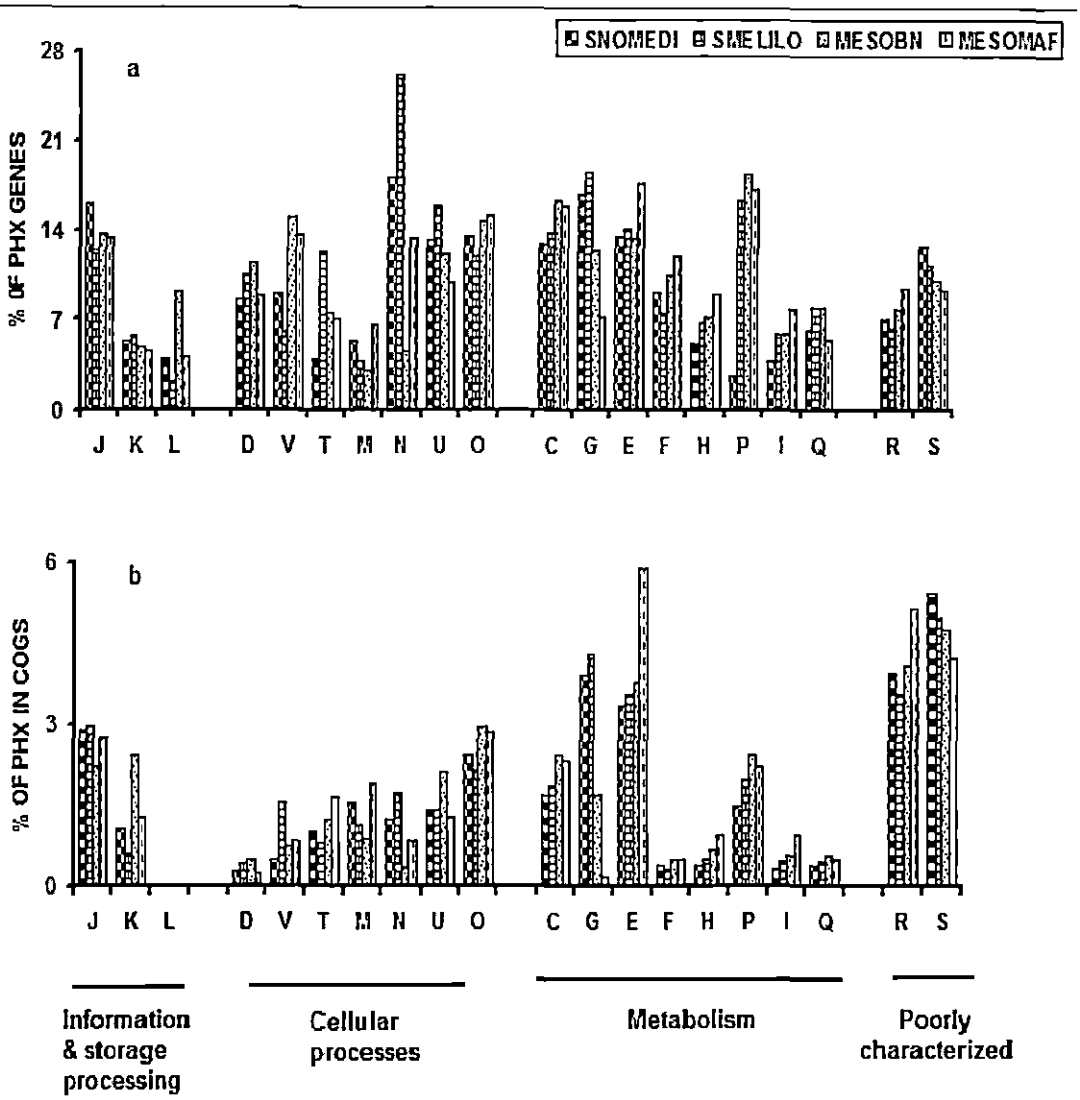

Figure 4.1.5.11: Distribution of potentially highly expressed genes in the proteobacterial members MESOBN, MESOMAF, SNOMEDI and SMELILO within COG functional groups

fresh water and some in marine water where the environment is not that stressed. The high percentage of genes in L (replication recombination repair), R (general function prediction) point out the ability of the genes associated with these functions to be overrepresented and thus influence day to day activities. Interestingly high proportion of genes present in the COGs for the studied cyanobacterial strains point out another aspect that these strains are quite stable and are not under the influence of genome reductions.

The allocation of PHX genes in the different COG functional groups for the studied proteobacterial members are demonstrated in Figures 4.1.5.8-4.1.5.12. In the *Rhizobium* members the COG functional group metabolism appears to have the maximum number of PHX genes. The top 5 COG

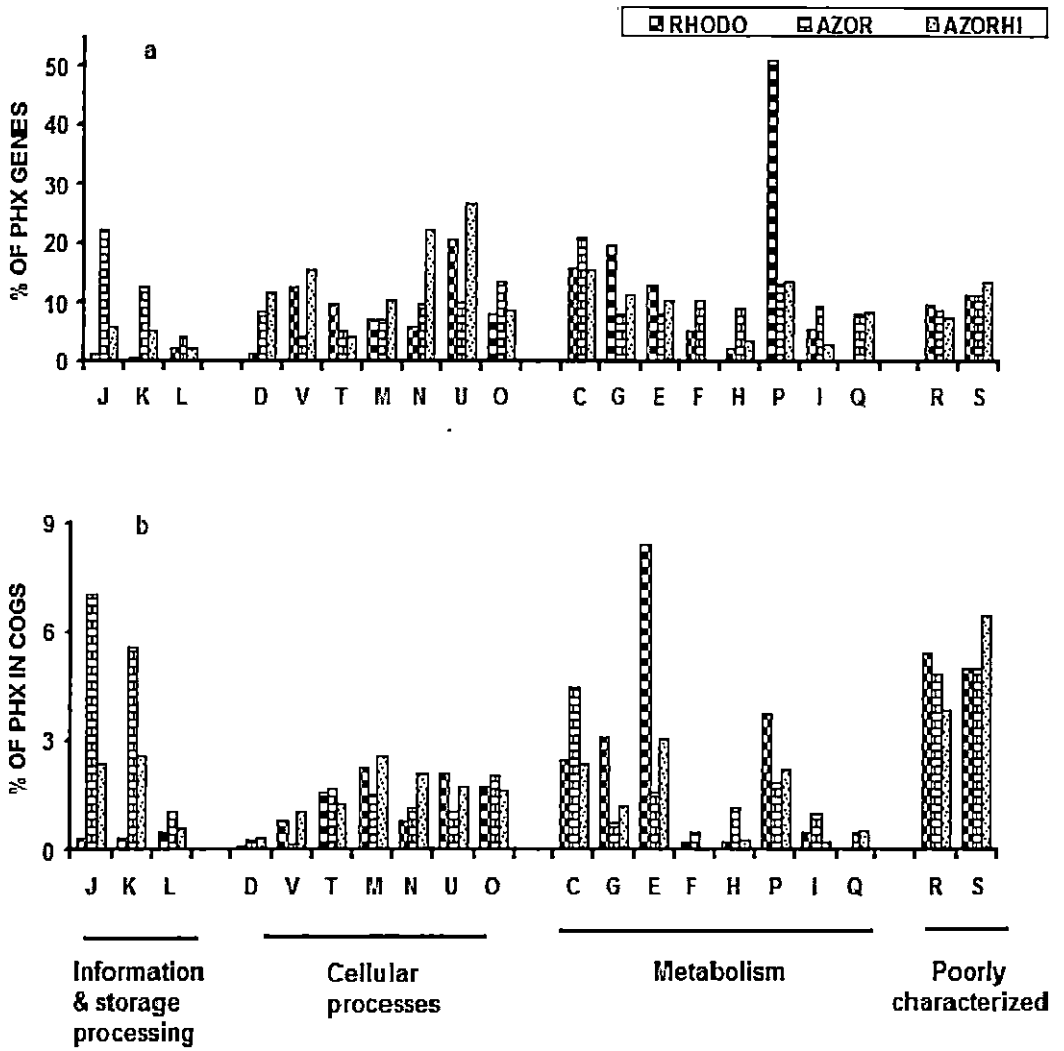

Figure 4.1.5.12: Distribution of potentially highly expressed genes in the proteobacterial members RHODO, AZOR and AZORHI within COG functional groups

categories are G, E, R, S and O. Barring RHICIAT most of the other rhizobial strains showed a more or less similar pattern. Amongst the bradyrhizobial members the metabolism COG group had the largest share of PHX genes. The major COG categories were C, G, E, H, R and S. *Bradyrhizobium* sp. BTAI1 revealed somewhat different pattern compared to the other strains. Sinorhizobial and mesorhizobial members illustrated a similar distribution pattern at the intraspecific level. COG categories, J, G, E R and S were the top ones. In *Azotobacter*, *Beijerinckia*, *Klebsiella* and *Gluconacetobacter* it was observed that the metabolism group had the lion's share and the top 5 categories were J, O, C, R and S. *Klebsiella* showed a different pattern in comparison to, *Beijerinckia* and *Gluconacetobacter*.

Among the other proteobacterial members like *Rhodobacter*, *Azorhizobium* and *Azoarcus* it was noticed that J, K, C, P, R and S were the major categories. The allocation in different COG functional groups throws light upon the lifestyle pattern of the bacteria. High number of COGs in the metabolism category for the rhizobial strains probably assists in carrying out its symbiotic lifestyle while competing with others to establish symbiotic association and fix atmospheric nitrogen. In the bradyrhizobial strains the allocation of PHX in the top COG categories may help in extending the host ranges of the strains and increase their capacity to regulate gene expression levels while in symbiotic association. The somewhat different nature of *Bradyrhizobium* sp. BTAI1 may also be attributed to its host specificity and biogeographical distribution. The larger proportion of PHX genes in specific COG categories for sinorhizobial and mesorhizobial strains may provide them a selective advantage for survival in symbiotic condition. In *Azotobacter* the allocation in COGs 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 different from the conventional one. The high percentage of COGs in the genome of *Klebsiella* is an indication that it has not undergone genome reduction and strongly uses the metabolic machinery to maintain itself in its ecological niche. The ability of *Gluconacetobacter* to survive as an endophyte is probably assisted by its higher share of PHX in metabolism. In *Rhodobacter*, *Azoarcus* and *Azorhizobium* the high percentage of PHX genes in J (translation), K (transcription), (C) energy production and metabolism, (I) inorganic ion transport and general function predictions is advantageous for their existence in their respective ecological niches by increasing their ability to respond to signals and regulation of gene expression levels.

4.2 Synonymous codon usage pattern of nitrogen fixation related genes

4.2.1 The nitrogen fixation related genes

The nitrogen fixation related genes (NFGs) includes the *nif* genes, *nod* genes, *fix* genes, hydrogenases, ferredoxins, nitrogen regulatory proteins etc. for efficiently carrying out the nitrogen fixation mechanism. In case of the clostridial strains the NFGs showed low GC, GC3 content and moderate codon bias. In *Clostridium acetobutylicum* protein coding genes and ribosomal protein genes are less biased compared to NFGs. In the Nc/ GC3 plots depicted in Figure 4.1.1.1 NFGs in *C. beijerinckii* remained clustered together. The Nc values for the NFGs are moderate. The CAI values of the NFGs are quite high indicating they have higher expression levels. The CBI values are negative. The Fop values are also low indicating lower degree of optimal codons. The aromaticity values are lower than that of the protein coding genes whilst in case of *C. acetobutylicum* the GRAVY values of NFGs are higher than other categories of genes.

The NFGs for the *Frankia* strains showed high codon bias. As expected correlation between the GC3 and Nc values were observed. NFGs had lower mean effective number of codon values in comparison to mean values for the protein coding genes. As seen for the representative strain of *Frankia* in Figure 4.1.1.1 NFGs are more or less clustered strongly with the ribosomal protein genes. NFGs had high CAI values however they were comparatively lower than protein coding genes in *Frankia* ACN14a and CcI3. Mutational pressure strongly acts upon the NFGs of *Frankia* owing to its very high GC3 content. The Fop and CBI values showed minor differences amongst the strains indicating that in the NFGs too other than mutational pressure other forces may be acting. The GRAVY values for these genes are lower in comparison to other genes while the aromaticity values are more or less similar.

In case of the studied archaeal members the NFGs showed low codon bias. However, the Nc value for NFGs for *Methanococcus maripaludis* were comparatively lower than that of the other strains indicating moderate codon bias. With the exception of *Methanococcus maripaludis*, NFGs for the other representative strains in the

Nc/GC3 (Figure 4.1.1.1) remain scattered plot implying dissimilar codon bias for these genes. In *Methanococcus maripaludis* it is clustered along with the ribosomal protein genes. The CAI values for these genes were moderate but they were higher in comparison to the protein coding genes signifying that they are better expressed than the protein coding genes. Excluding *Methanosarcina mazei*, the CBI values for NFGs were negative indicating lower degree of optimal codons and low codon bias. Lower Fop values implied lesser percentage of optimal codons in these genes. The GRAVY score is also much lower than the ribosomal protein genes indicating that GRAVY does not play a significant role in codon usage variations in the NFGs. Low aromaticity scores signified less number of aromatic amino acids in these genes. In the *Chlorobium* genomes, like the protein coding genes and ribosomal protein genes the NFGs exhibit low bias and have moderate GC and GC3 content. The expression levels of the NFGs are quite high barring *C. chlorocromatii*. The Fop values of this category of genes are higher in comparison to the total protein coding genes implying greater degree of optimal codons and the role of factors other than mutational bias in influencing codon usage variations.

It is seen from Table A4.1.1 (See appendix) that there lies some difference in the interspecific level with respect to the GC, GC3 and Nc values amongst the cyanobacterial strains. The NFGs in the Nc/GC3 plots for the cyanobacterial strains remained scattered with the exception of *Synechococcus* and *Trichodesmium*. The CAI values for the nitrogen fixation related genes were high in most of the strains. NFGs had low CBI and moderate Fop values signifying lower content of optimal codons in them. Aromaticity and GRAVY scores were also low for these genes in cyanobacterial strains implying lesser role played by them in influencing codon usage bias.

In the proteobacterial members the NFGs show variation with respect to GC, GC3 and Nc values. In the Nc/GC3 plots the NFGs for the strains *Azoarcus*, *Rhodobacter*, *Gluconacetobacter*, *Azorhizobium*, and *Bradyrhizobium japonicum* USDA are clustered at one lower end of the plot and they are highly biased while in the other members they are scattered. Highly biased NFGs are probably under the influence of translational efficiency. The expression levels of NFGs for the studied

proteobacterial members are high and in some members it is higher than the averages for the protein coding genes. This implies that gene expression levels play an important role in effecting codon usage variations in the proteobacterial members. The CBI values for the NFGs in some of the studied members are lower than that of the protein coding and ribosomal protein genes indicating that they have lower fraction of optimal codons in their genes. The Fop values are moderate in most of the NFGs implying lesser quantity of optimal codons in these genes. The mean GRAVY and aromaticity scores for the NFGs are low in majority of the cases cutting across proteobacterial groups.

4.2.2 Correspondence analysis of codon and amino acid usages

In clostridial strains especially *Clostridium acetobutylicum* and the *Frankia* strains the nitrogen fixing genes remained grouped together at the core while in *Clostridium beijerinckii* they remain scattered in the correspondence analysis of CU and AAU. Clustering at the core region pointed out their conserved nature. In methanogens, the NFGs like their ribosomal counterpart in COA of CU for the bacterium *Methanospirillium hungatei* remained scattered suggesting dissimilar codon usage patterns. Moreover, in the other strains they remained confined to the core. But in COA of AAU all the NFGs were located in the core. The NFGs in the chlorobial strains remained in the core region in the COA of CU and AAU. All these implied the conserved nature of the nitrogen fixation related genes in the chlorobial strains. In case of the cyanobacterial strains it is seen in all the cases that NFGs remained confined in the central regions as found for other studied genomes pointing out its conserved nature. In the COA for CU and AAU for *Azoarcus* and *Rhodobacter* the NFGs remain in the core region as observed for other organisms. In *Azotobacter*, *Klebsiella*, *Beijerinckia* and *Gluconacetobacter* NFGs are located in the central regions. In *Mesorhizobium* sp. BNC1, *Sinorhizobium meliloti*, *S. medicae*, *Azorhizobium*, *Bradyrhizobium* sp. ORS278 and *Bradyrhizobium* sp. BTAI, as in other cases NFGs are confined along the core region. In the rhizobial strains the NFGs are scattered in the thick core regions. It implied that these genes had dissimilar codon usage patterns compared to that of the ribosomal protein genes. The presence

of NFGs in the core regions in all the studied genomes across lineages point out an important fact that these genes are not under the influence of outside factors and since they are linked to one of the most important systems in the ecosystem they probably lack the capability to undergo massive changes.

4.2.3 Potentially highly expressed NFGs

The potentially highly expressed nitrogen fixation related genes for the studied organisms are depicted in Table A4.1.3 (See appendix). As per the CAI cut-off in the clostridial strains, *Clostridium acetobutylicum* and *C. beijerinckii* had 2 and 9 NFGs in the PHX category. Majority of the NFGs in CBEI that are PHX play a vital role in the nitrogen regulatory processes and strongly assist in nitrogen fixation as free-living ones in the soil. In the *Frankia* strains, CcI3, ACN14a and EAN1pec had 7, 6 and 13 NFGs in the PHX category. Core NFGs that are PHX in *Frankia* strains provide them with the selective advantage to compete and fix nitrogen in stressed soils worldwide as well as with wider host ranges especially in *Frankia* EAN1pec. The number of PHX genes amongst NFGs for the archaeal members were 6, 7, 15 and 5 for *Methanospirillium hungatei*, *Methanococcus maripaludis*, *Methanosarcina mazei* and *Methanobacterium thermoautotrophicum* were correspondingly. Presence of a number of important NFGs in the PHX category portray their role in fixing nitrogen in environments like sewages and sludges under stressed condition. On the basis of the CAI cut–off values *Chlorobium limicola*, *C. phaeobacteroides* DSM 266, *C. phaeobacteroides* BS1, *Chlorobium ferrooxidans*, *C. tepidium* and *C. chlorocromatii* housed 7, 6, 6, 9, 2 and 5 NFGs in PHX. The details of these genes are depicted in Table A4.1.3 (See appendix). Majority of the NFGs in the PHX category are hydrogenases, nitrogen regulatory proteins and nitrogen iron protein complexes. These are extremely important for carrying out the process efficiently and the anoxic sulfide rich environments in which majority of the studied chlorobial strains survive favors it. Given the fact that the chlorobial strains survive in stratified aquatic environments higher expression levels are indispensable for carrying out nitrogen fixation.

As per the CAI cut-off values, *Nostoc* sp. 7102, *Nostoc* sp. 73102, *Trichodesmium*, *Synechococcus*, *Nodularia* and *Lynbya* had 4, 13, 7, 5, 5

and 3 NFGs in the PHX category respectively. The number of PHX genes for the NFGs in *Cyanothece* strains C7424, C7425, C8801, C51142 and CCY0110 were 4, 0, 4, 1 and 0 respectively. In *Anabaena* the total number of NFGs in the PHX category was 11. The cyanobacterial NFGs that are PHX are represented in Table A4.1.3 (See appendix). Among the proteobacterial members, *Azotobacter*, *Beijerinckia*, *Klebsiella* and *Azoarcus* had 12, 10, 6 and 11 NFGs in the PHX category. In the *Rhizobium* members, RHIZCFN4, RHIVICIA, RHIWSM and RHICIAT the number of NFGs in the PHX category were 9, 1, 11 and 10 correspondingly. Among the *Bradyrhizobium* and *Sinorhizobium* members, the amount of NFGs in the PHX category was 7, 16, 6, 1 and 0 for BRORS, BRBTAI, BRUSDA, SNOMEDI and SMELILO respectively. In case of the *Mesorhizobium* members like MESOBN and MESOMAF the number of NFGs in PHX were 3 and 4 while in other members like *Rhodobacter*, *Azorhizobium* and *Gluconacetobacter* the number of PHX in NFGs were 2, 10 and 2 respectively. It is quite fascinating to see that a number of proteobacterial strains had a high number of NFGs in the PHX category. Most of them are associated with the core nitrogen fixation machinery and assist in different ways for efficient nitrogen fixation either in obligate symbionts or endophytes or in free-living ones.

4.3 Analysis of TTA codons containing genes from GC rich diazotrophs

4.3.1 TTA codons in high G+C organisms

The organisms listed in Table 3.3 were considered for the analysis of TTA codons owing to their high GC content. Like many other GC rich microorganisms, the TTA codons are the rarest ones in the studied GC rich genomes. The general characteristics of the TTA codon containing sequences are depicted in Table 4.3.1. *Bradyrhizobium sp.* BTAI1 had the highest number of TTA containing sequences i.e., 824 followed by *Azotobacter* and *Frankia* EAN1pec.

Codon usage analysis of the TTA codon containing genes in the studied GC rich organisms revealed diversity at the intraspecific as well as interspecific level. The mean values of GC3, GC, Nc, CAI, CBI, Fop, GRAVY and Aromaticity for the TTA codon containing genes (TCA) are illustrated in Table 4.3.2. It is clearly seen that the mean GC3, GC, CAI, CBI, Fop and aromaticity values of the TTA codon containing gene sequences (TCA) are all less than the average values of the protein coding genes whereas the mean Nc and GRAVY values are higher. Variation in the mean Nc values for the different genes groups were observed within the same species

Table 4.3.1 General characteristics of TTA codons containing GC rich organisms

ORGANISM	Codon	Total TTA	TTA+Sequences	Minimum TTA %
Azoarcus sp. BH72	TTA	237	186	0.0177
Azorhizobium caulinodans ORS 571	TTA	402	301	0.0249
Azotobacter vinelandii AvOP	TTA	1220	678	0.752
Bradyrhizobium sp. ORS278	TTA	682	553	0.0319
Bradyrhizobium sp. BTAI1	TTA	1076	824	0.044
Frankia alni ACN14a	TTA	394	350	0.018
Frankia sp. Ccl3	TTA	415	363	0.026
Frankia sp. EAN1pec	TTA	747	625	0.029
Gluconacetobacter diazotrophicus PAL5	TTA	570	394	0.0694
Rhodobacter sphaeroides 2.4.1	TTA	229	152	0.016

Table 4.3.2 Mean values of GC3, GC, Nc, CAI, CBI, Fop, GRAVY and Aromaticity for the TTA codon containing genes (TCA)

Organism	Genes	GC%	GC3%	Nc	CAI	CBI	Fop	GRAVY	Aromaticity
Azoarcus sp. BH72	PCG	68±0.08	86.5±0.05	33.55±5.32	0.80±0.05	0.35±0.09	0.61±0.05	-0.01±0.37	0.07±0.02
	TCG	65.4±0.06	79.1±0.10	39.78±8.80	0.80±0.03	0.27±0.12	0.56±0.07	-0.02±0.38	0.07±0.02
Azotobacter vinelandii AvOP	PCG	65.5±0.05	81.4±0.11	36.87±7.87	0.67±0.06	0.29±0.12	0.58±0.07	-0.12±0.40	0.07±0.02
	TCG	59.8±0.07	68.4±0.15	45.49±9.09	0.64±0.07	0.17±0.14	0.51±0.08	-0.19±0.41	0.07±0.02
Azorhizobium caulinodans ORS 571	PCG	67.5±0.03	85.5±0.05	35.26±4.77	0.69±0.04	0.23±0.07	0.54±0.05	0.03±0.39	0.06±0.02
	TCG	65.2±0.05	78.7±0.10	40.62±7.34	0.68±0.04	0.19±0.08	0.52±0.04	-0.01±0.36	0.07±0.02
Bradyrhizobium sp. BTAI1	PCG	65.1±0.03	81.6±0.07	37.97±6.56	0.76±0.04	0.22±0.08	0.54±0.05	-0.02±0.41	0.07±0.02
	TCG	62.7±0.04	75.2±0.09	43.20±8.02	0.75±0.04	0.18±0.08	0.51±0.05	-0.03±0.42	0.07±0.02
Bradyrhizobium sp. ORS278	PCG	65.9±0.03	84.1±0.07	36.23±5.96	0.71±0.04	0.25±0.08	0.55±0.04	-0.02±0.41	0.07±0.02
	TCG	64±0.04	78.4±0.10	41.08±8.34	0.70±0.04	0.21±0.09	0.53±0.05	-0.004±0.42	0.07±0.02
Frankia alni ACN14a	PCG	72.3±0.04	88.9±0.07	33.88±5.70	0.72±0.11	0.24±0.08	0.54±0.05	-0.07±0.38	0.05±0.02
	TCG	70.7±0.05	84.7±0.10	37.03±8.98	0.70±0.13	0.22±0.09	0.53±0.06	-0.11±0.37	0.05±0.02
Frankia sp. CcI3	PCG	70.1±0.04	84.8±0.06	36.91±4.85	0.70±0.07	0.22±0.07	0.53±0.04	-0.08±0.34	0.05±0.02
	TCG	69.3±0.04	82.3±0.07	39.15±5.85	0.67±0.09	0.20±0.07	0.52±0.04	-0.05±0.34	0.06±0.02
Frankia sp. EAN1pec	PCG	70.7±0.05	91.9±0.03	36.31±5.53	0.82±0.06	0.24±0.07	0.54±0.04	-0.08±0.34	0.05±0.02
	TCG	68.4±0.04	82.1±0.10	39.34±7.19	0.70±0.10	0.21±0.07	0.53±0.04	-0.09±0.35	0.06±0.02
Gluconacetobacter diazotrophicus PAL5	PCG	66.9±0.05	82.1±0.09	36.82±6.70	0.67±0.05	0.27±0.10	0.56±0.05	-0.02±0.40	0.06±0.02
	TCG	64.0±0.06	75.9±0.12	40.99±8.29	0.63±0.06	0.21±0.12	0.53±0.07	-0.07±0.43	0.06±0.02
Rhodobacter sphaeroides 2.4.1	PCG	68.8±0.04	87.3±0.07	34.90±5.38	0.70±0.04	0.20±0.07	0.52±0.04	-0.006±0.41	0.06±0.02
	TCG	61.5±0.08	70.7±0.10	46.51±9.88	0.66±0.06	0.13±0.10	0.48±0.05	-0.13±0.43	0.06±0.02

*PCG- Protein coding genes and TCG- TTA codon containing genes

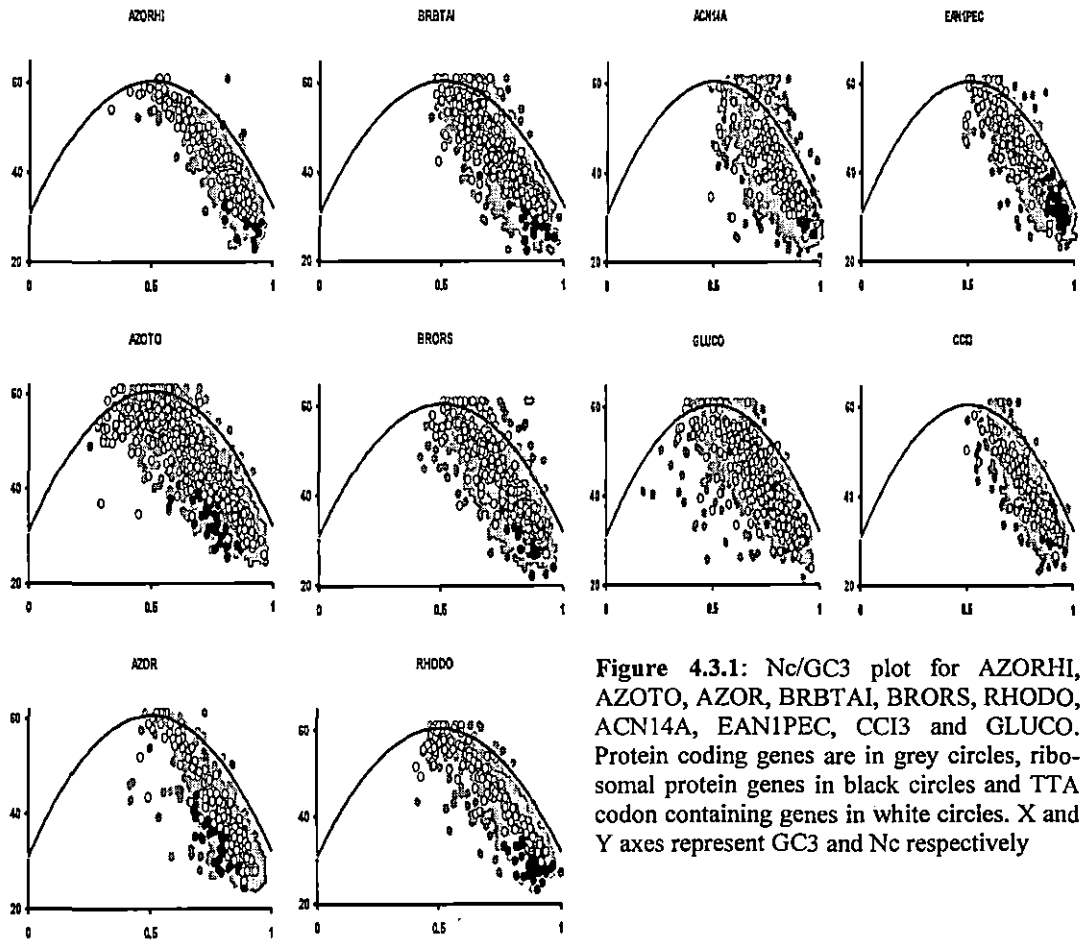

Figure 4.3.1: Nc/GC3 plot for AZORHI, AZOTO, AZOR, BRBTAI, BRORS, RHODO, ACN14A, EAN1PEC, CCI3 and GLUCO. Protein coding genes are in grey circles, ribosomal protein genes in black circles and TTA codon containing genes in white circles. X and Y axes represent GC3 and Nc respectively

as well as other species. As expected a correlation between GC3 and Nc values were clearly observed. Nc decreased with increase in GC3 and vice-versa. The Nc/GC3 plots (Figures 4.3.1) for the studied organisms suggest an effective way to analyze codon usage variations among the genes. It is seen that most of the TTA codon containing genes remain scattered signifying their varying codon usage patterns while the ribosomal protein genes remain at the lower ends of the plot. Except for a few TTA codons containing genes majority of the genes lie well below the expected values. If synonymous codon bias in TTA codon containing genes were absolutely dictated by GC3s, all Nc values might have had fallen on the expected curve of the Nc/GC3 plots.

In a nutshell, all these indices indicate that these TTA codon containing genes are under the influence of mutational pressure and are less biased in their codon usage patterns as exemplified by their high Nc values. Their expression levels are also predicted to be less than the average protein coding genes. The comparatively lower CBI values for the

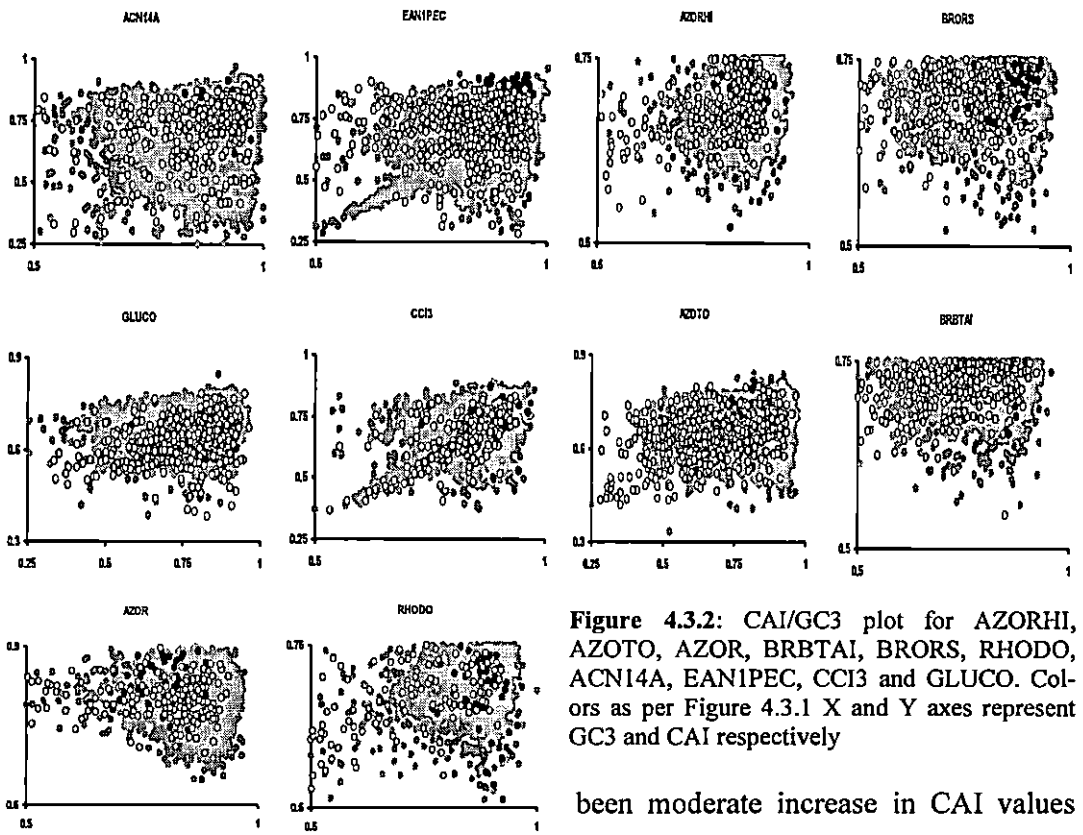

Figure 4.3.2: CAI/GC3 plot for AZORHI, AZOTO, AZOR, BRBTAI, BRORS, RHODO, ACN14A, EAN1PEC, CCI3 and GLUCO. Colors as per Figure 4.3.1 X and Y axes represent GC3 and CAI respectively

TTA codon containing genes substantiates claim that mutational pressure influences codon bias in these genes. However, the moderate Fop values for these genes imply that there are other factors other than mutational bias effecting codon usage variation in these genes. If mutational bias alone influenced codon bias, these genes would have had a low Fop value.

When GC3 values were correlated with the CAI values (Figures 4.3.2) it was found that in the three *Frankia* strains and in *Azorhizobium* CAI values sharply increased with the increase of GC3. In the other studied organisms there has been moderate increase in CAI values with the increase of GC3. It is evident from the studies that codon usage of the TTA containing genes in GC rich diazotrophs are highly skewed towards codon alternatives ending in G or C.

Although the expression levels of the TTA containing genes have been found to be less in comparison to the protein coding genes TTA codon containing genes had a number of potentially highly expressed genes. *Azoarcus*, *Azorhizobium* and *Azotobacter* had 19, 23 and 41 TTA codon containing genes in the PHX category. *Frankia* Ccl3, EAN1pec and ACN14a had 20, 34 and 34 TTA containing PHX genes. *Bradyrhizobium sp.* BTAI1 and *Bradyrhizo-*

*bium sp.*ORS278 had 75 and 48 TTA containing PHX genes respectively. *Gluconacetobacter* and *Rhodobacter* had 23 and 5 genes for the TTA category. Most of the genes are associated with the metabolism of the organisms and seems to play and important role in effecting the nature of the organisms.

When functional analysis of the TTA codon containing genes performed with respect to the COG functional groups it was observed that COG groups: Information and storage processing, Cellular processes, Metabolism and Poorly characterized had 12, 7, 13, 6; 20, 9, 52 and 13 for *Azoarcus* and *Azorhizobium* respectively. So it is seen that TTA codon containing genes in metabolism play a prominent role in the lifestyle of these organisms. In the *Bradyrhizobium* strains, *Bradyrhizobium sp.* BTAI1 and *Bradyrhizobium sp.*ORS278 had 25, 25, 68, 19; 26, 8, 46 and 13 genes in the COG groups: Information and storage processing, Cellular processes, Metabolism and Poorly characterized. TTA containing genes associated with metabolism towered over categories for these strains. In order to survive in the symbiotic condition the metabolism genes play an important role and thus the result. In case of the *Frankia* strains it was seen that COG groups Information and storage processing, Cellular processes, Metabolism and Poorly characterized had 7, 0, 29, 6; 8, 5, 10, 2; 5, 3, 17 and 6 TTA containing genes for *Frankia* EAN1pec, CcI3 and ACN14a respectively. In these cases too COGs associated with metabolism are the highest and play a role in maintaining the lifestyle of these organisms in the soil as well as in symbiotic condition. In *Rhodobacter*, *Gluconacetobacter* and *Azotobacter* TTA containing genes had 3, 3, 6, 2; 29, 25, 36, 15; 44, 37, 85 and 47 numbers for COG groups: Information and storage processing, Cellular processes, Metabolism and Poorly characterized categories. Like the other GC rich studied organisms the COGs linked to metabolism had the highest number of genes implying the importance of TTA containing genes in influencing the metabolism in the respective habitats and ecological niches of the organisms. This particular trend is similar for all the studied GC rich organisms and is an indicator of the fact that genes probably have some adaptive values in specific ecological conditions.

4.4. Analysis of proteomes

4.4.1 Isoelectric point

The proteomes for the studied organisms showed a tri-modal distribution for pI i.e., it has protein distributed in three distinct clusters, acidic (pI<7.4) neutral (pI=7.4 to 8.1) and basic (pI>8.1). These findings are in line with those obtained for other proteomes (Knight *et al.*, 2004; Nandi *et al.*, 2005; Schwartz *et al.*, 2001; Kiraga *et al.*, 2007). Tri-modal distributions of the pI are largely dominated either by the acidic or by basic clusters and are highly influenced by ecological niche of the organism (Knight *et al.*, 2004). The studied proteomes are by and large acidic in nature; among which nitrogen fixation related proteins are more acidic in contrast to whole proteomes (Table 4.4.1) (also see Figures 4.4.1 & 4.4.2). The average energetic costs of proteins are more or less similar with very little difference compared to nitrogen fixation related proteins (data not shown). The wide variation of pI values in the acidic and basic class point out the diversity exists between the proteomes and that pI distribution is influenced by organism's habitat. Inspite of differences; all these nitrogen fixing proteomes are by and large acidic. This may be due to the fact that nitrogen fixing regulon needs supply of electrons for its activity. Since it accepts electrons it is acidic in nature. Moreover, in symbiotic plasmids of *Rhizobium etli* CFN 42, *Bradyrhizobium* sp. BTAI and *Sinorhizobium meliloti* symbiotic plasmids A and B the pI values were 11.35; 11.10; 11.28, 11.32 respectively. These are highly basic in contrast to the total proteome and the nitrogen fixing proteins. This basic nature may be attributed to the fact that their obligatory host associations and intracellular nature.

Significant results were obtained, when distribution of pI values across COGs functional groups were studied. Variations of pI among functional categories of COGs were noted with majority of COGs showing distributions across acidic and basic clusters (Figures.4.4.3-4.4.12). It is observed that the allocation of pI across COGs represent the organism's adaptability to different habitat and ecological niche. In case of clostridial proteomes, *Clostridium beijerinckii* has an acidic set whereas *C. acetobutylicum* has a basic tilt (Figure 4.4.3). The prominent COG cluster appears to be information and storage processing (Isp) in both the

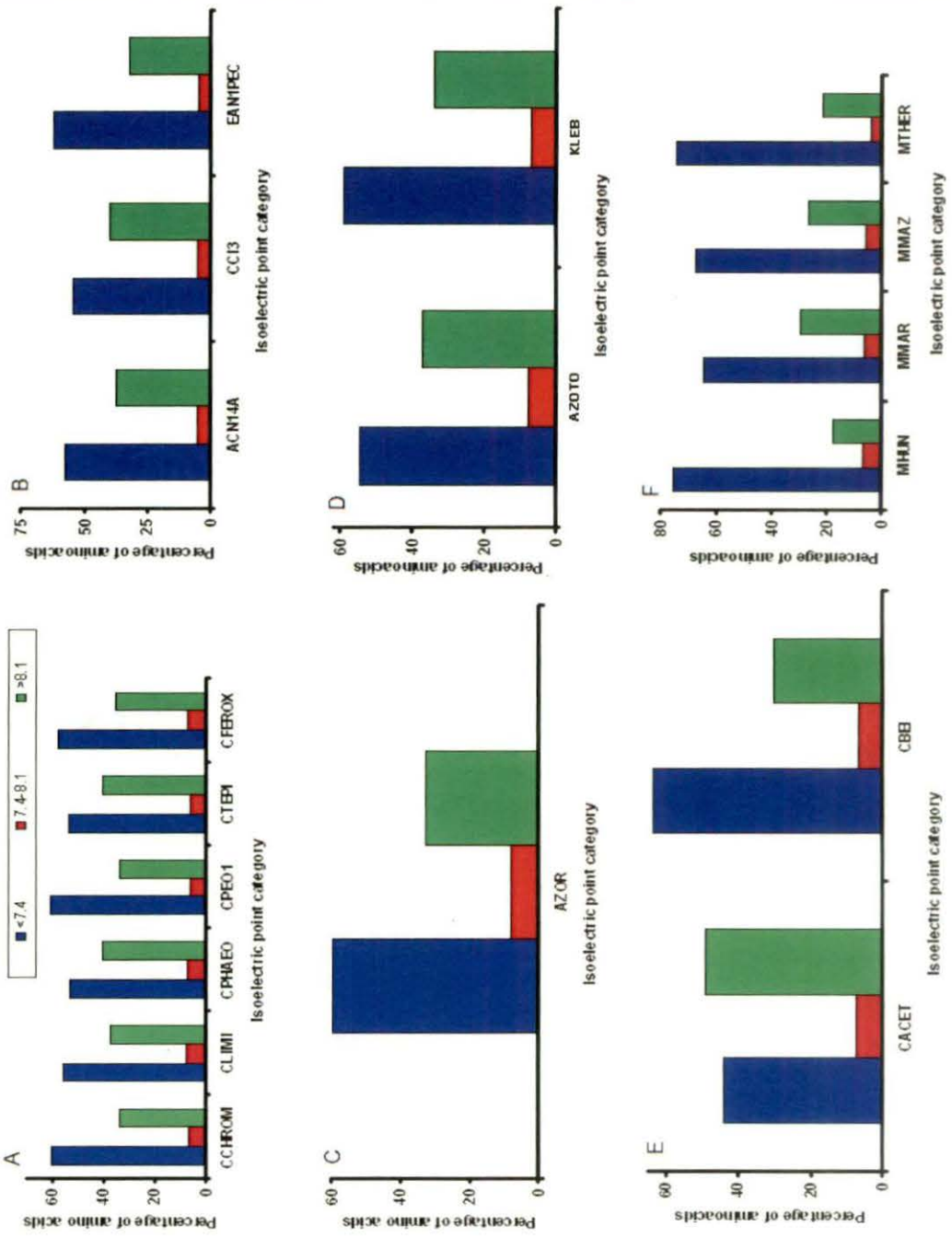

Figure 4.4.1: Tri-modal distribution of pI values in A) *Chlorobium* strains B) Actinobacterial *Frankia* C) *Azoarcus* D) *Azotobacter* and *Klebsiella* E) *Clostridium* strains F) Methanogenic archaea.

genera. Figure 4.4.1a shows the difference in the number of amino acids in the metabolism (Me) group between *C. beijerinckii* and *C. acetobutylicum* when all amino acids in the COGs are taken into consideration. Chi-square test for the COG groups in the acidic dataset as well as basic dataset revealed significant differences. Chi square result obtained for the acidic set and basic set

Table 4.4.1: Mean values of AAAI and pI of proteins in the studied organisms *WFP indicates whole proteomes and NFP nitrogen fixing proteomes

Organism	AAAI		pI	
	WP	NFP	WP	NFP
Anabaena variabilis ATCC 29413	0.69±0.01	0.50±0.04	7.0±2.16	6.17±1.42
Azoarcus sp. BH72	0.85±0.01	0.82±0.02	7.40±2.09	5.88±1.41
Azorhizobium caulinodans ORS 571	0.52±0.002	0.61±0.04	7.60±2.16	6.48±1.53
Azotobacter vinelandii AvOP	0.82±0.004	0.78±0.09	7.64±2.25	6.28±1.99
Beijerinckia indica subsp. indica ATCC 9039	0.86±0.01	0.80±0.05	7.23±2.03	5.97±1.55
Bradyrhizobium sp. ORS278	0.51±0.002	0.58±0.04	7.73±2.09	6.45±1.52
Bradyrhizobium sp. BTAi	0.52±0.01	0.59±0.04	7.84±2.11	6.44±1.54
B. japonicum USDA 110	0.53±0.01	0.62±0.01	8.04±2.10	7.37±1.35
Chlorobium ferroxidans DSM 13031	0.71±0.006	0.59±0.06	7.35±2.07	6.77±1.62
C. chlorocromatii CaD3	0.68±0.01	0.63±0.03	7.23±2.06	6.29±1.53
C. limicola DSM 245	0.70±0.01	0.57±0.06	7.38±2.13	6.41±1.57
C. phaeobacteroides DSM 266	0.70±0.01	0.61±0.04	7.52±2.13	6.32±1.35
C. phaeobacteroides BS1	0.67±0.01	0.54±0.08	7.13±2.17	5.66±1.64
C. tepidium TLS	0.72±0.01	0.59±0.05	7.62±2.28	6.45±1.65
Clostridium acetobutylicum ATCC 824	0.28±0.007	0.34±0.04	7.82±2.14	6.78±1.57
C. beijerinckii NCIMB 8052	0.27±0.01	0.37±0.03	6.88±2.07	5.94±1.21
Cyanothece sp. PCC 7424	0.64±0.01	0.58±0.03	6.87±2.20	5.66±1.30
Cyanothece sp. PCC 7425	0.76±0.01	0.63±0.07	6.98±2.18	5.99±1.46
Cyanothece sp. PCC 8801	0.66±0.01	0.55±0.04	6.77±2.15	5.83±1.60
Cyanothece sp. PCC 51142	0.62±0.01	0.60±0.03	6.87±2.26	5.64±1.32
Cyanothece sp.CY0110	0.65±0.01	0.58±0.02	6.9±2.24	5.79±1.40
Frankia ACN14a	0.54±0.002	0.62±0.02	7.60±2.64	5.98±1.59
Frankia CcI3	0.56±0.003	0.58±0.02	7.73±5.56	5.76±1.34
Frankia EAN1pec	0.56±0.03	0.66±0.02	7.30±2.50	5.64±1.56
Glucanoacetobacter diazotrophicus PAL 5	0.68±0.02	0.64±0.02	7.75±2.24	6.19±0.16
Klebsiella pneumoniae pneumoniae MGH78578	0.85±0.006	0.82±0.02	7.32±2.10	6.38±1.43
Lynbya sp. PCC 8106	0.65±0.01	0.55±0.03	6.71±2.24	5.69±1.34
Mesorhizobium sp. BNC1	0.69±0.01	0.58±0.13	7.25±2.26	5.65±1.11
M. loti MAFF303099	0.65±0.01	0.64±0.03	7.60±2.20	6.64±1.60
Methanospirillium hungatei JF-1	0.52±0.01	0.51±0.03	6.34±1.82	6.13±1.38
Methanococcus maripaludis S2	0.28±0.006	0.31±0.01	6.83±2.10	6.71±1.76
Methanosarcina mazei Goe 1	0.50±0.01	0.47±0.03	6.65±2.08	6.37±1.61
Methanobacterium thermoautotrophicum Delta H	0.36±0.01	0.28±0.02	6.32±2.15	5.30±1.16
Nodularia spumigena CY9414	0.69±0.006	0.57±0.03	6.94±2.24	5.72±1.27
Nostoc sp. 7120	0.70±0.008	0.59±0.03	7.09±2.22	6.07±1.42
Nostoc sp. 73102	0.70±0.01	0.55±0.10	6.96±2.17	5.92±1.50
Rhizobium etli CFN42	0.63±0.005	0.58±0.06	7.30±2.20	6.39±1.72
R etli CIAT 652	0.62±0.005	0.60±0.08	7.25±2.19	6.40±1.93
R leguminosarum bv. viciae 3841	0.62±0.003	0.63±0.01	7.20±2.17	6.81±1.69
R leguminosarum bv. trifolii WSM1325	0.62±0.003	0.58±0.02	7.11±2.13	6.71±1.55
Rhodobacter sphaeroides 2.4.1	0.68±0.005	0.63±0.03	7.18±2.34	6.18±1.57
Sinorhizobium medicae	0.62±0.03	0.59±0.01	7.30±2.23	6.96±1.44
S. meliloti 1021	0.62±0.005	0.61±0.03	7.31±2.23	7.04±1.61
Synechococcus sp. JA-2-3B'a(2-13)	0.80±0.03	0.72±0.01	7.33±2.20	6.64±1.66
Trichodesmium erythraeum IMS101	0.58±0.01	0.48±0.06	7.62±2.33	5.49±1.28

revealed values of 12.61 and 7.55 respectively, which is much higher than the tabulated chi-square values at p=0.05 level. Kiraga *et al.*, (2007) postulated that pI of free-living organisms/extracellular organisms show a tendency towards acidity, while those associated with host plants were basic.

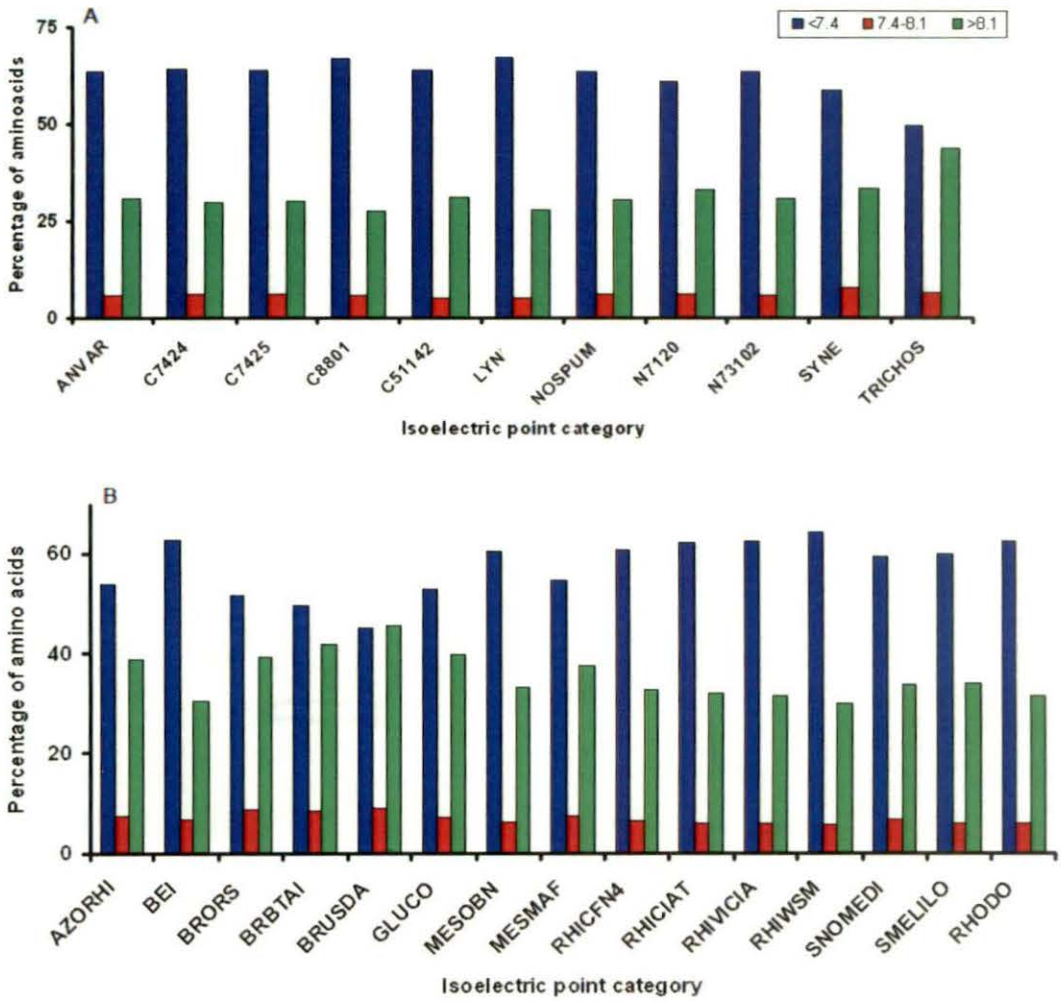

Figure 4.4.2: Tri-modal distribution of pI values in A) Cyanobacterial strains B) Alpha proteobacterial strains.

The diversity among the acidic and basic clusters for the two clostridial strains is clearly evident from the chi-square test. It may be due to the alteration of pI bias en route basicity compared to the pI of the species thriving in other habitats (Kiraga *et al.*, 2007). In general, the Isp group having the functional categories translation, transcription and DNA replication play a significant role in influencing pI in clostridia. Proteome diversity between *C. beijerinckii* and *C. acetobutylicum* may be the outcome of the diverse composition of the amino acids in the proteome. The difference in number of amino acids in the Me group for *C. beijerinckii* substantiates the fact.

The actinobacterial *Frankia* strains showed an inclination towards acidity despite the fact that they are host associated. It is observed from Figure 4.4.4A-B, that Ccl3 is less acidic compared to the other two. Earlier

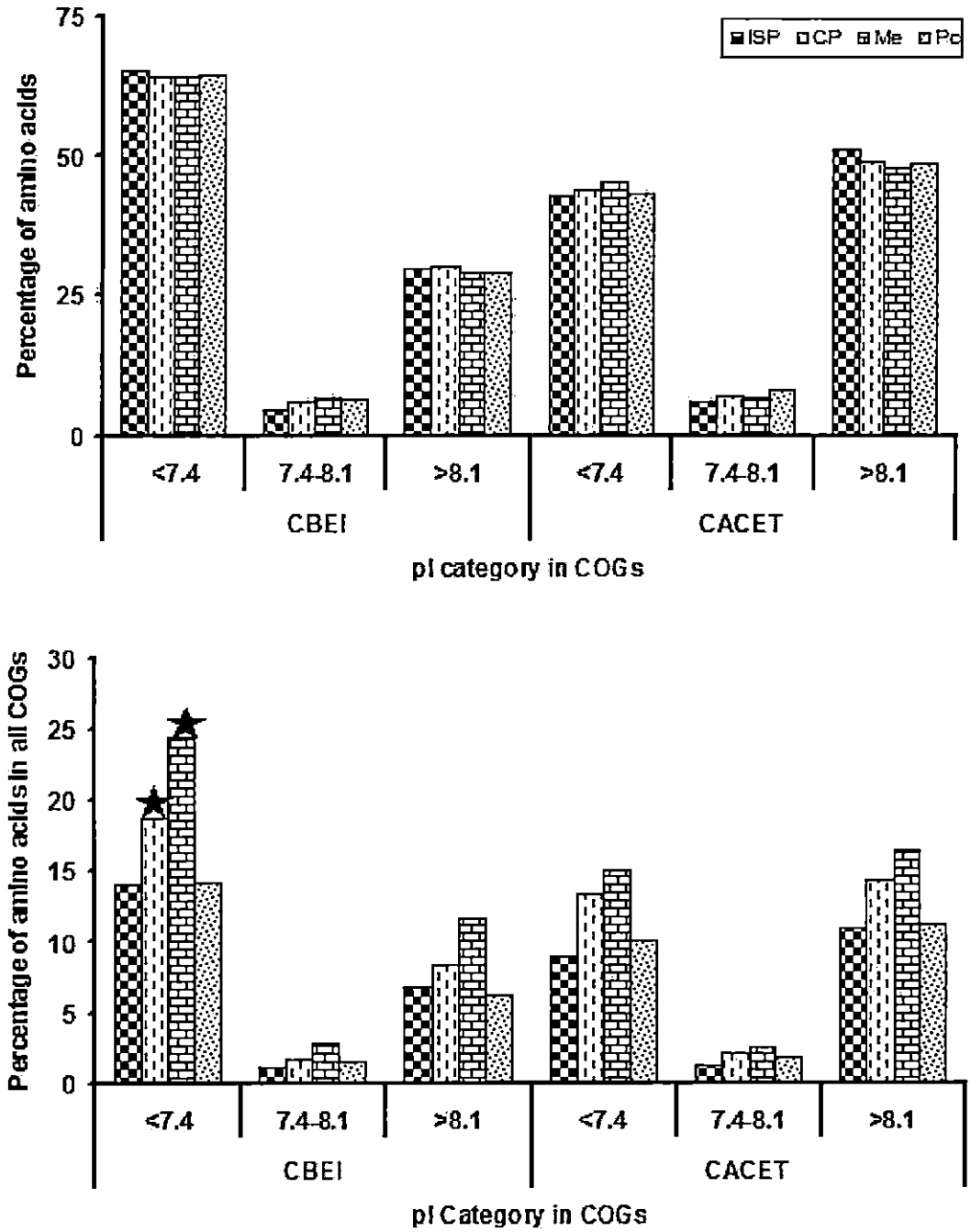

Figure 4.4.3(A-B): Distribution of pI in different COG clusters for *Clostridium aceto-butylicum* and *Clostridium beijerinckii*. Isp group is represented by black checkers, Cp by dashed verticals, Me by horizontal bricks and Pc by small confetti colors.

reports (Sen *et al.*, 2008) suggest that *Frankia* ACN14a and EAN1pec have greater host ranges, and can survive without symbiosis compared to CcI3 having a restricted host range without evidence of survival outside the host. CcI3 is biogeographically restricted to Australia and Pacific islands (Sen *et al.*,

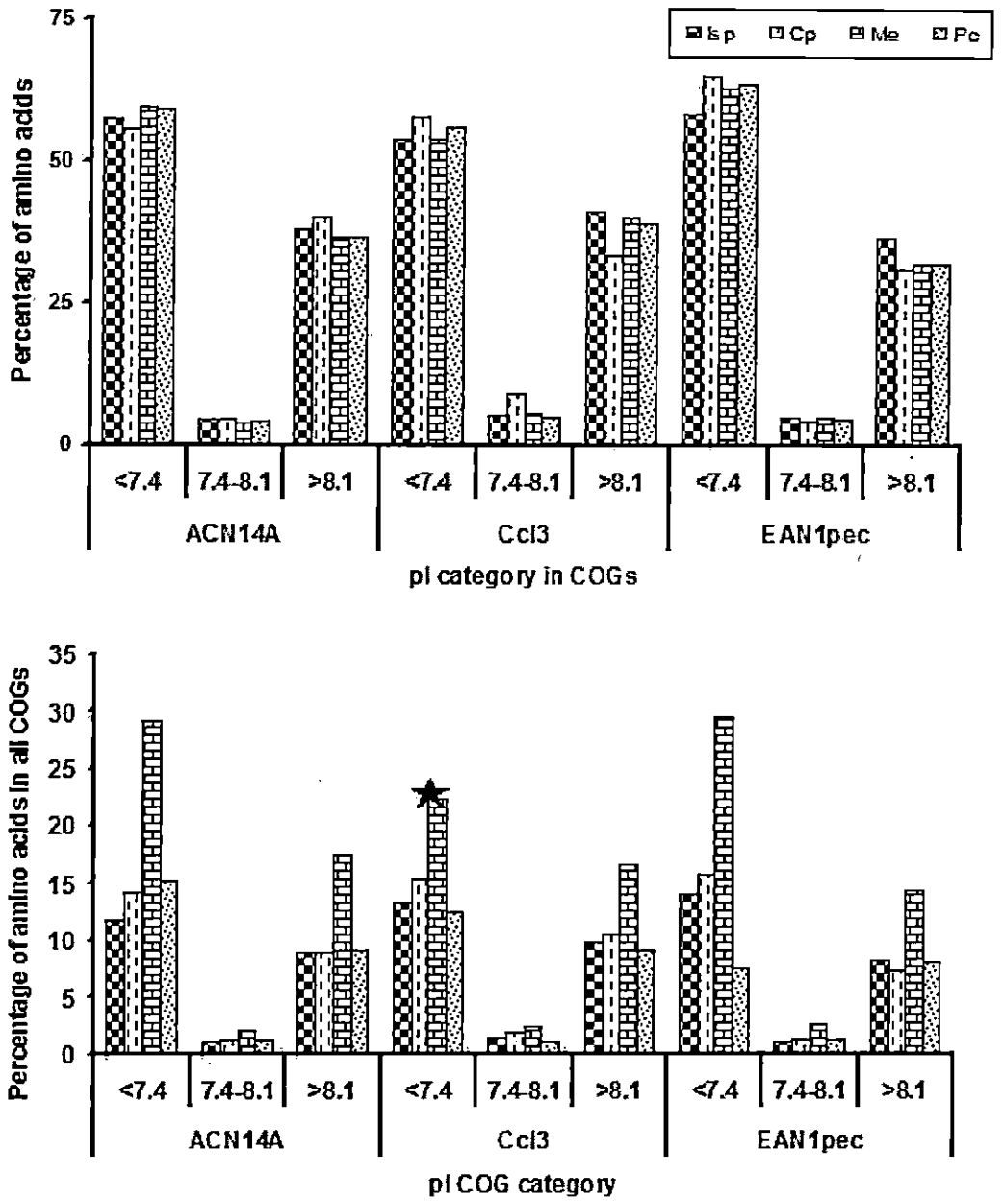

Figure 4.4.4(A-B): Distribution of pI in different COG clusters for actinobacterial *Frankia*. Colors as per Figure 4.4.3 (A-B).

2008) compared to *Frankia* ACN14a and EAN1pec which has worldwide distributions. The ability of *Frankia* ACN14a and EAN1pec to survive outside the symbiotic association might have shifted their pI towards acidity. COGs belonging to metabolism (Me) groups in ACN14a revealed maximum number of amino acids in the acidic category, while in CcI3 and EAN1pec it is observed that cellular processes (Cp) peaks in the acidic class while Isp peaks

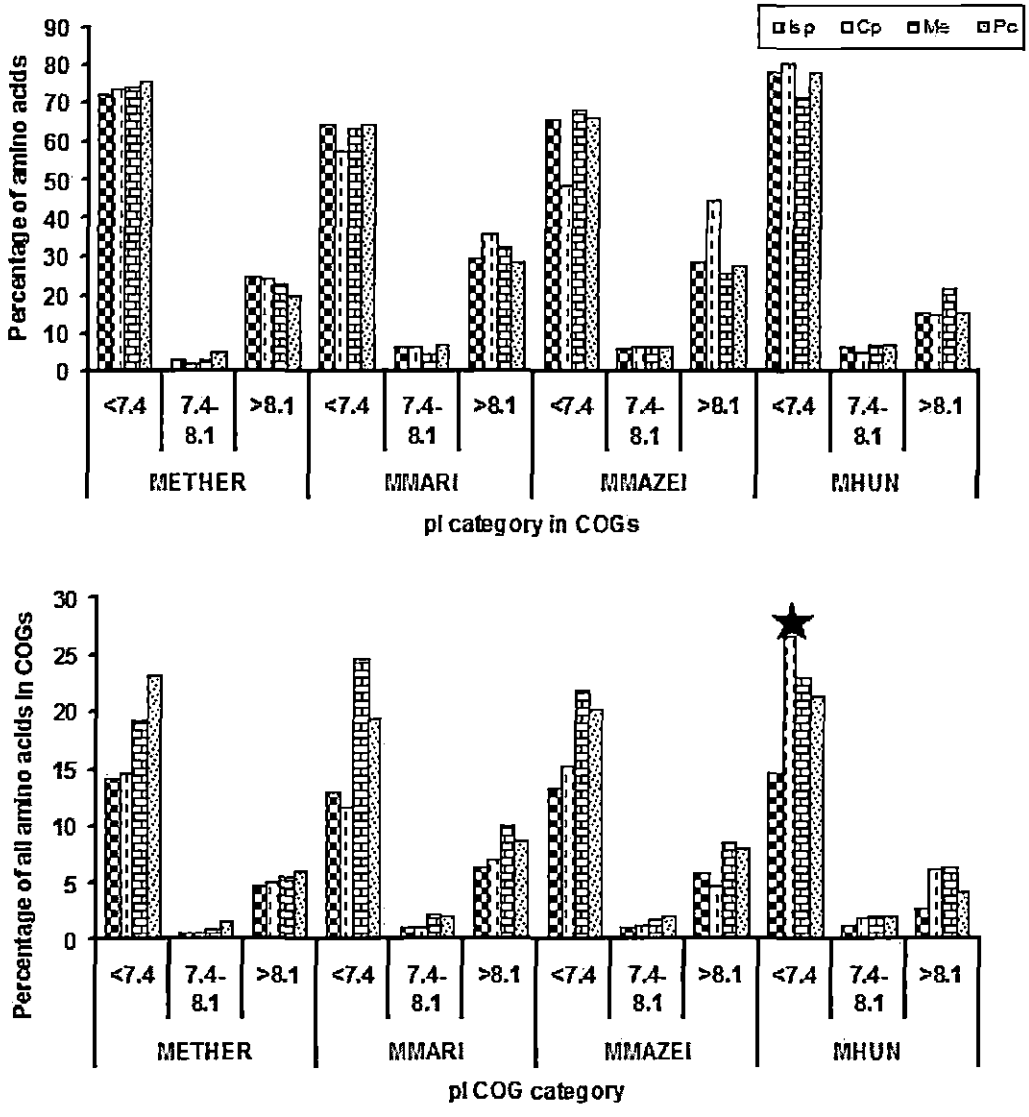

Figure 4.4.5(A-B): Distribution of pI in different COG clusters for methanogenic archaea. Colors as per Figure 4.4.3 (A-B).

in the basic class. The elevated levels of amino acids in the Me and Pc group for ACN14a in acidic category implied the role played by metabolism in influencing the lifestyle of ACN14a. Similarly the abundance of amino acids in the Cp group towards acidity for *Frankia* CcI3 and EAN1pec indicates that COGs associated with cellular processes increases the ability of these strains to respond to signals in the soil and the environment of the cells modifies the shift towards acidity. Figure 4.4.4B shows the diversity in amino acids number for the metabolism (Me) group between *Frankia* ACN14a, CcI3 and EAN1pec. Data obtained for the metabolism (Me) group of *Frankia*

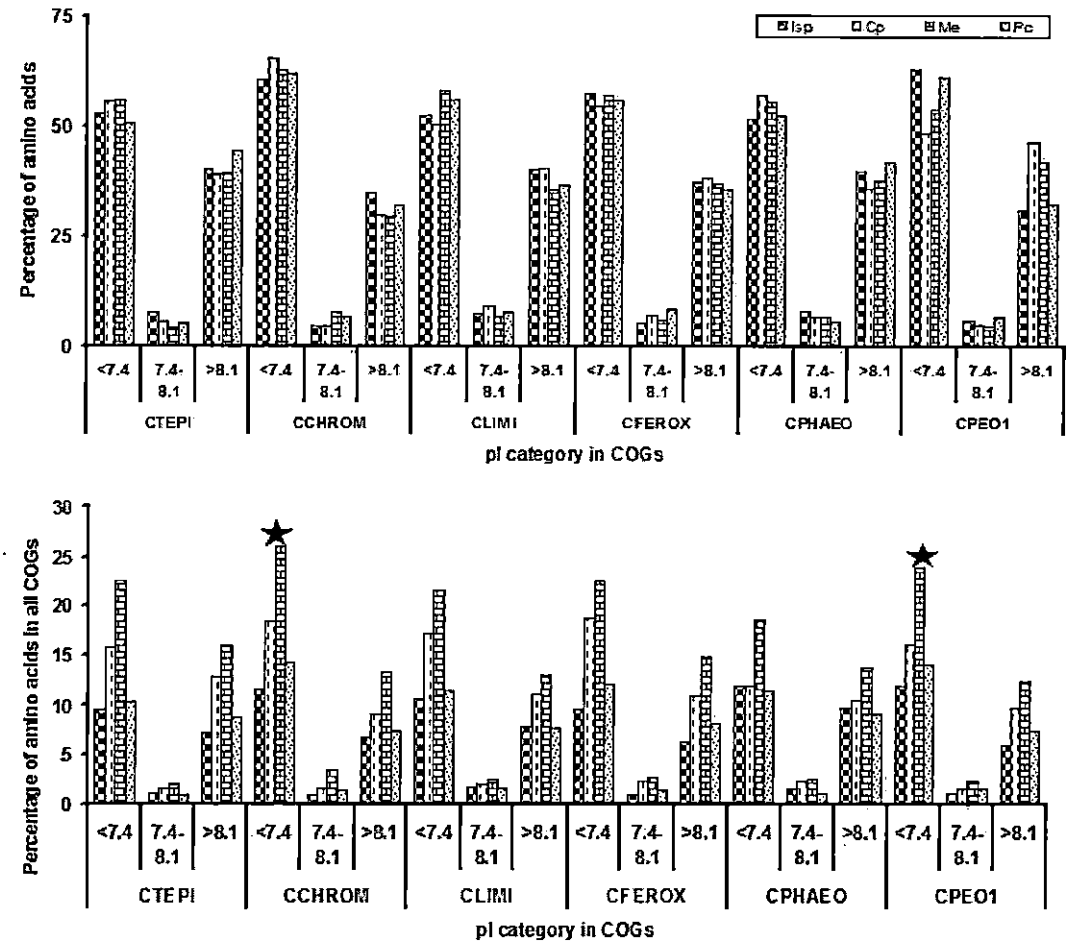

Figure 4.4.6(A-B): Distribution of pI in different COG clusters for *Chlorobium* strains. Colors as per Figure 4.4.3 (A-B)

CcI3 with that of ACN14a and EAN1pec revealed significant differences at p=0.05 level. On the other hand as observed from Figure 4.4.4B, lower number of amino acids in the cellular processes (Cp) category for CcI3 may be an indicator of its symbiotic nature.

In case of methanogenic archaea the pI values are highly biased towards the acidic class as expected. The outcome obtained for methanogenic archaea used in the study point out that they are more acidic than any of the studied organisms. These organisms are free-living and survive in acidic environments like sewage and sludge consequently having a high acidic pI. In *Methanobacterium thermoautotrophicum*, Pc group had the maximum amino acids in the acidic class while in *Methanococcus maripaludis* and *Methanosarcina mazei* Isp and Me dominates the same. Cp and Isp contained huge number of amino acids in the acidic class for

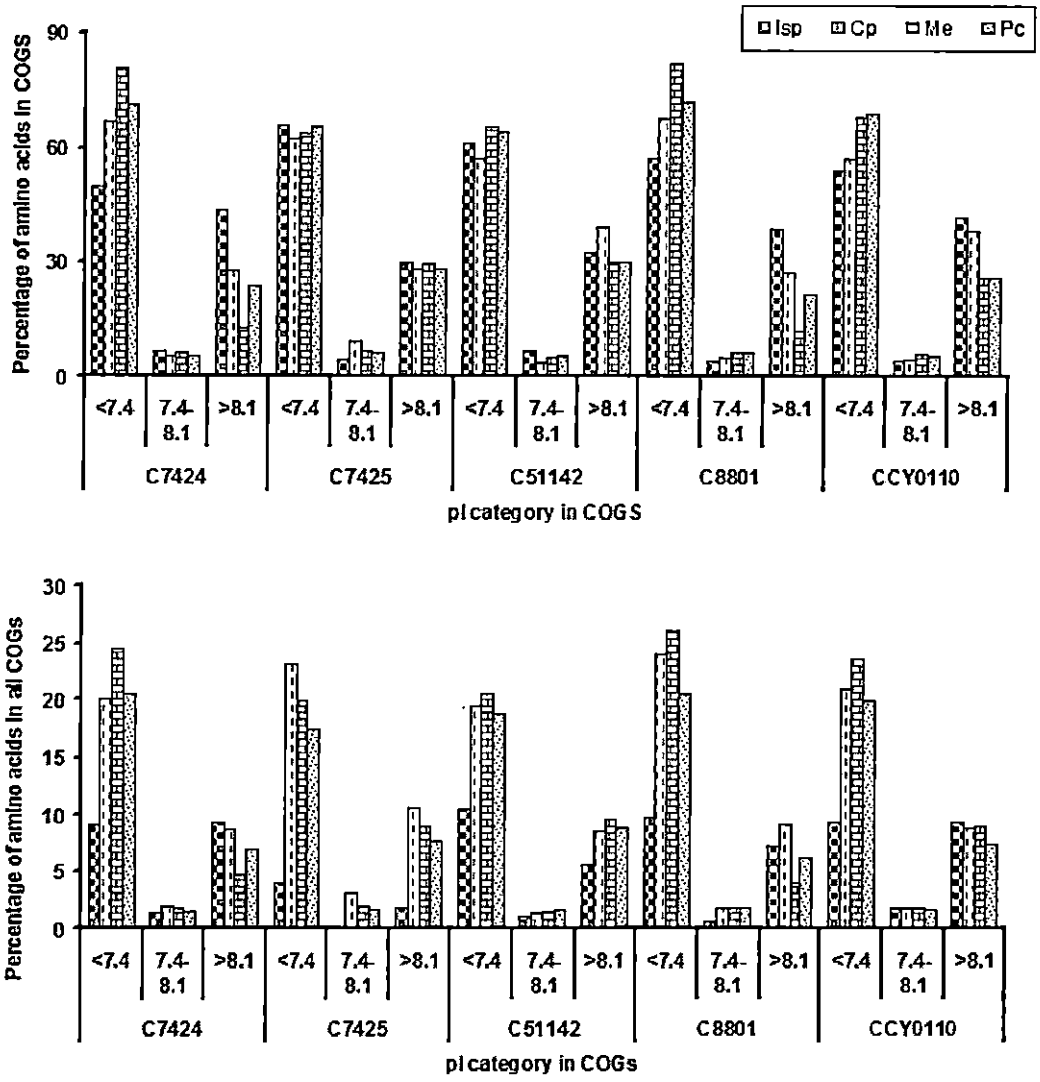

Figure 4.4.7(A-B): Distribution of pI in different COG clusters for *Cyanothece* strains. Colors as per Figure 4.4.3 (A-B).

Methanospirillium hungatei. When all the COGs for amino acids were considered it was found that the *Methanospirillium hungatei* Cp group towers above other methanogenic strains which is significant even at p=0.05 level. Figure 4.4.5A-B shows the results for methanogenic archaea. The overrepresentation of amino acids as seen in Figure 4.4.5 in the COGs group Isp and Cp for *Methanospirillium hungatei* indicated that amino acids associated with translation, transcription, cellular processes are vital for survival in their habitats. Similarly the dominance of Isp and Me groups of COGs for *Methanosarcina mazei* and *Methanococcus maripaludis* depicts the role of metabolic amino acids and those associated with translation, transcription

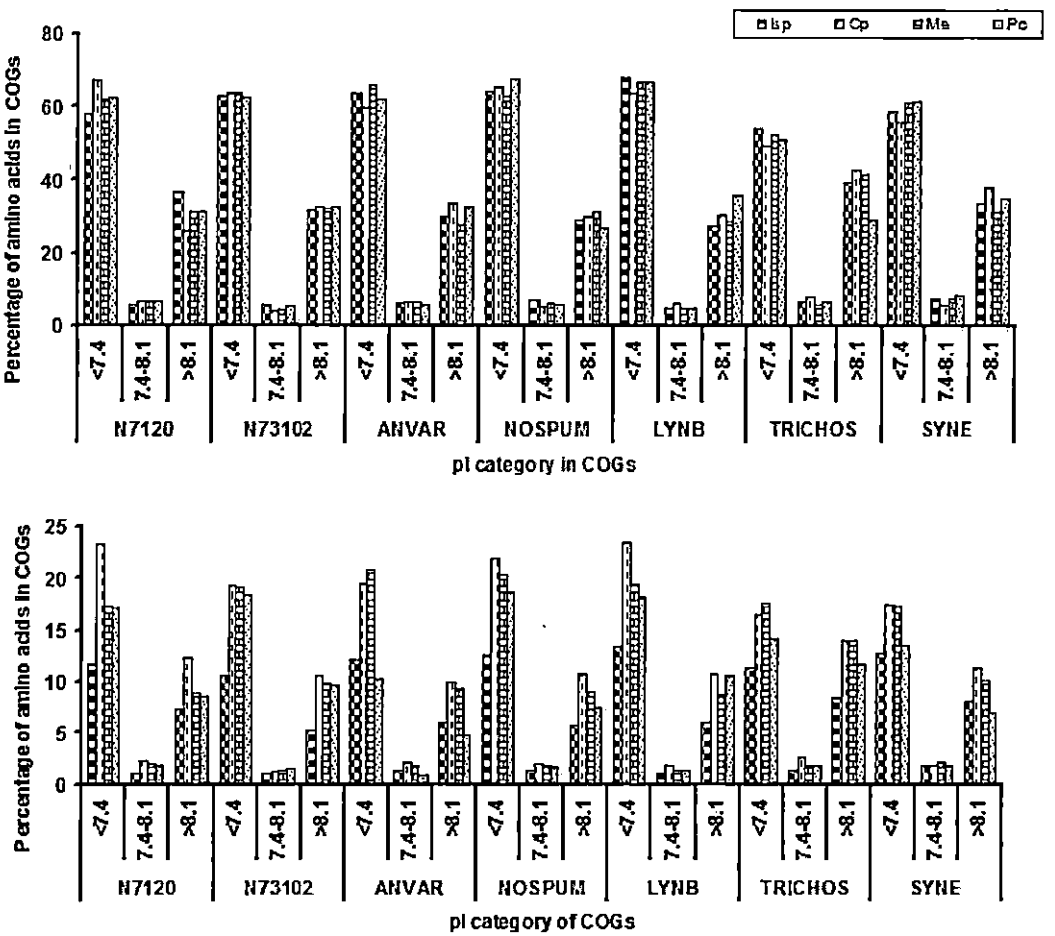

Figure 4.4.8(A-B): Distribution of pI in different COG clusters for cyanobacterial strains. Colors as per Figure 4.4.3 (A-B).

and DNA recombination for their existence in acidic sludges, sewages etc. They probably assist in stability and solubility of acid rich proteins. Figure 4.4.5B indicated the diversity amongst the studied archaeal strains with respect to the overrepresentation of the amino acid residues connected to cellular processes highlighting their role in their survival in sludges and sewages. These strains come from different locations and environmental adaptations leading to proteome diversity.

Free-living *Chlorobium* strains on the other hand showed an acidic tilt. The findings for pI categorization in COGs of the *Chlorobium* strains complemented earlier reports (Kiraga *et al.*, 2007) that acidic proteomes are characteristic of free-living aquatic bacteria. There is some degree of variation with respect to the dominance of the COGs functional groups in *Chlorobium* (Figure 4.4.6). The *Chlorobium* strains undertaken in this study come from different habitats as

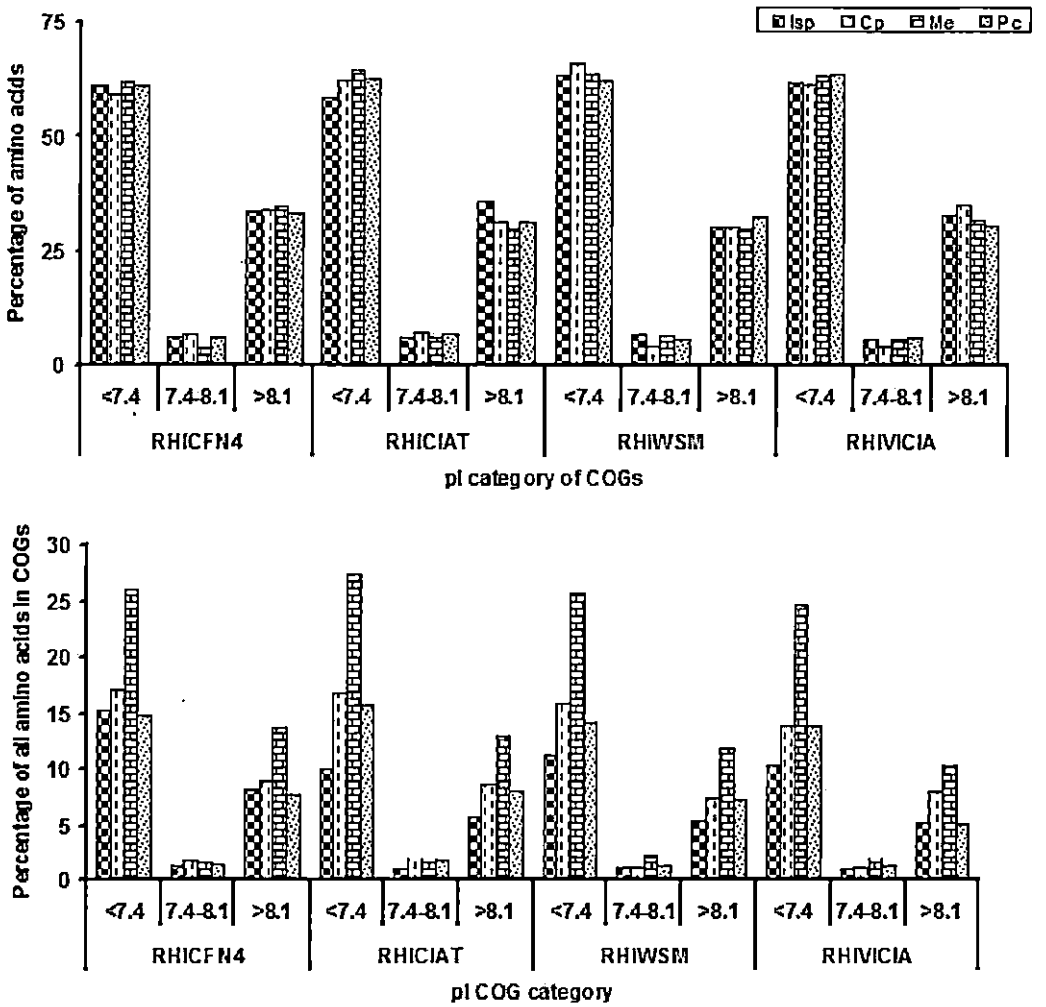

Figure 4.4.9(A-B): Distribution of pI in different COG clusters for *Rhizobium* strains. Colors as per Figure 4.4.3 (A-B).

well as biogeographic locations and habitat specificity (Sur *et al.*, 2008) along with ecological adaptations play an important role in diversity at the intraspecific level. Among the *Chlorobium* strains, *C. chlorocromatii* and *C. phaeobacteroides* BS1 had a huge percentage of amino acids from Cp and Isp respectively. Evaluation of all the COGs for amino acids illustrated that metabolism (Me) group for *C. chlorocromatii* and *C. phaeobacteroides* BS1 had a statistically significant (at p=0.05 level) higher representation compared to other chlorobial strains. However, the other strains bore a more or less similar pattern. The dominance of the COG functional groups Me in *C. chlorocromatii* and *C. phaeobacteroides* BS1; Isp and Me for others reveals the role of amino acids associated with gene regulation, metabolic pathways, inorganic ion transport, cell wall membrane envelope biogenesis, cellular

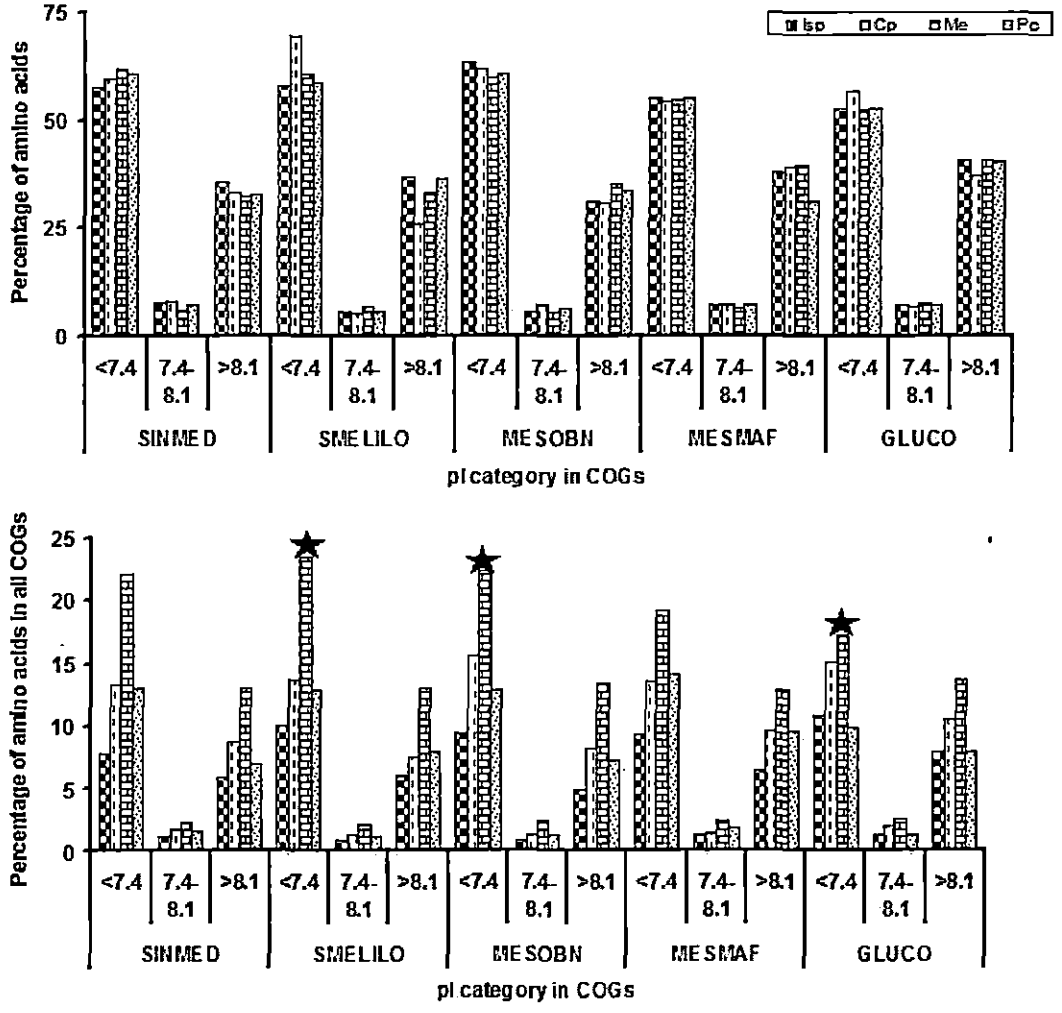

Figure 4.4.10(A-B): Distribution of pI in different COG clusters for sinorhizobial, mesorhizobial species and *Gluconacetobacter*. Colors as per Figure 4.4.3 (A-B).

mechanisms, signal pathways etc. to harmonize the cellular processes for their survival under varied environments.

In the cyanobacterial members especially *Cyanothece* strains, the pattern revealed in the acidic and basic class is quite similar with the domination of Me. C7424 and C8801 had a comparatively similar pattern showing less diversity as exemplified in Figure 4.4.7. The results obtained for cyanobacterial organisms are in line with the observations of Knight *et al.*, (2004), with respect to acidic proteome set in aquatic condition. The proteomes of *Cyanothece* are more or less similar. Amino acids belonging to the COGs functional group Me play a vital role for the *Cyanothece* strains in channeling different metabolic pathways for effective survival and habitual activities

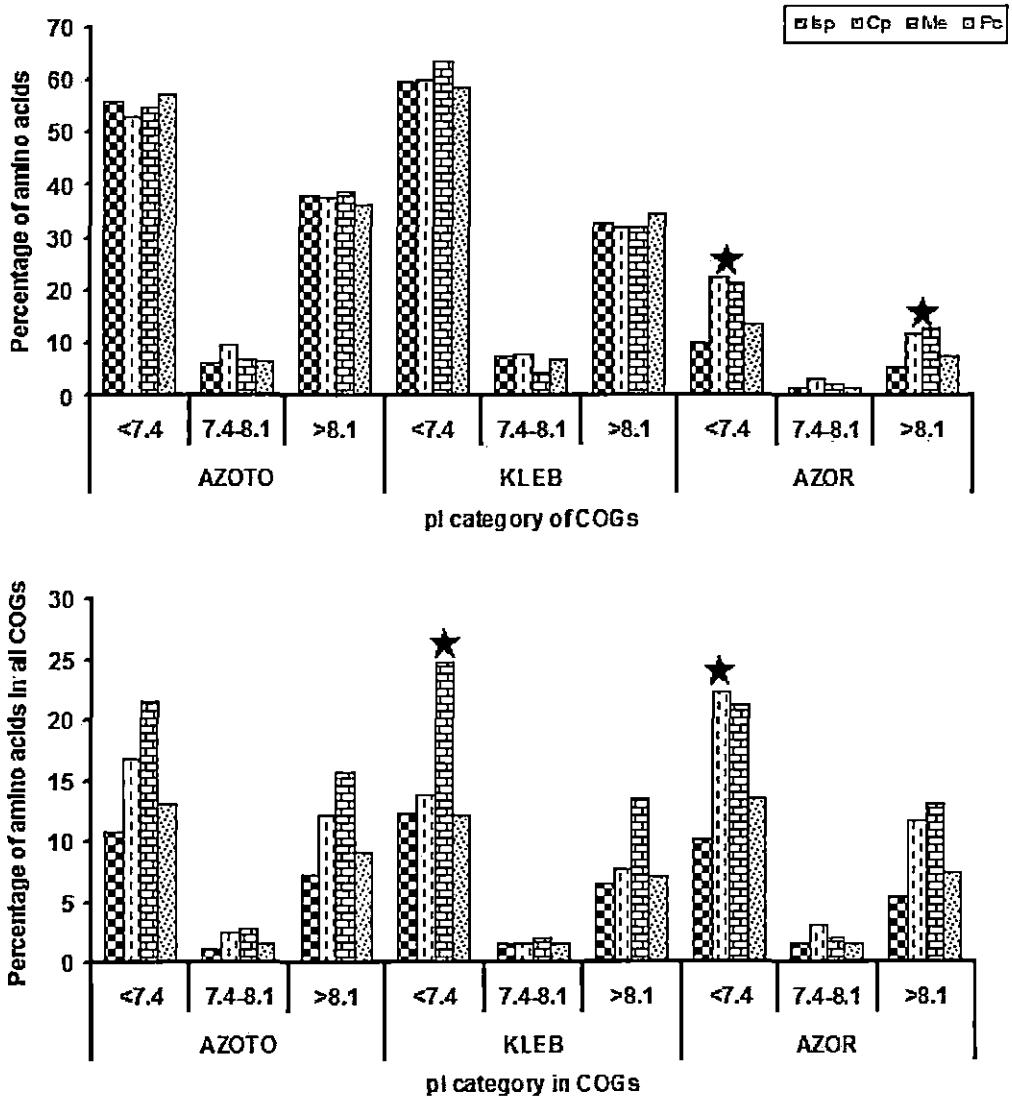

Figure 4.4.11(A-B): Distribution of pI in different COG clusters for *Azotobacter, Klebsiella* and *Azoarcus.* Colors as per Figure 4.4.3 (A-B).

in the aquatic environment. However, there is some variation between the cyanobacterial members other than *Cyanothece*. In *Nostoc* and *Anabaena* strains Cp and Isp predominate the acidic and basic classes. A good deal of diversity is noted amongst other cyanobacterial strains with respect to their ecological niche and habitats compared to *Cyanothece* strains Figure 4.4.8. The diversity is especially evident amongst the marine and fresh water strains with respect to distribution of different COG groups. While Me group dominated in the *Cyanothece* strains, Cp was more prevalent for *Nostoc, Nodularia, Lyngbya* and *Synechococcus.* While amino acids in

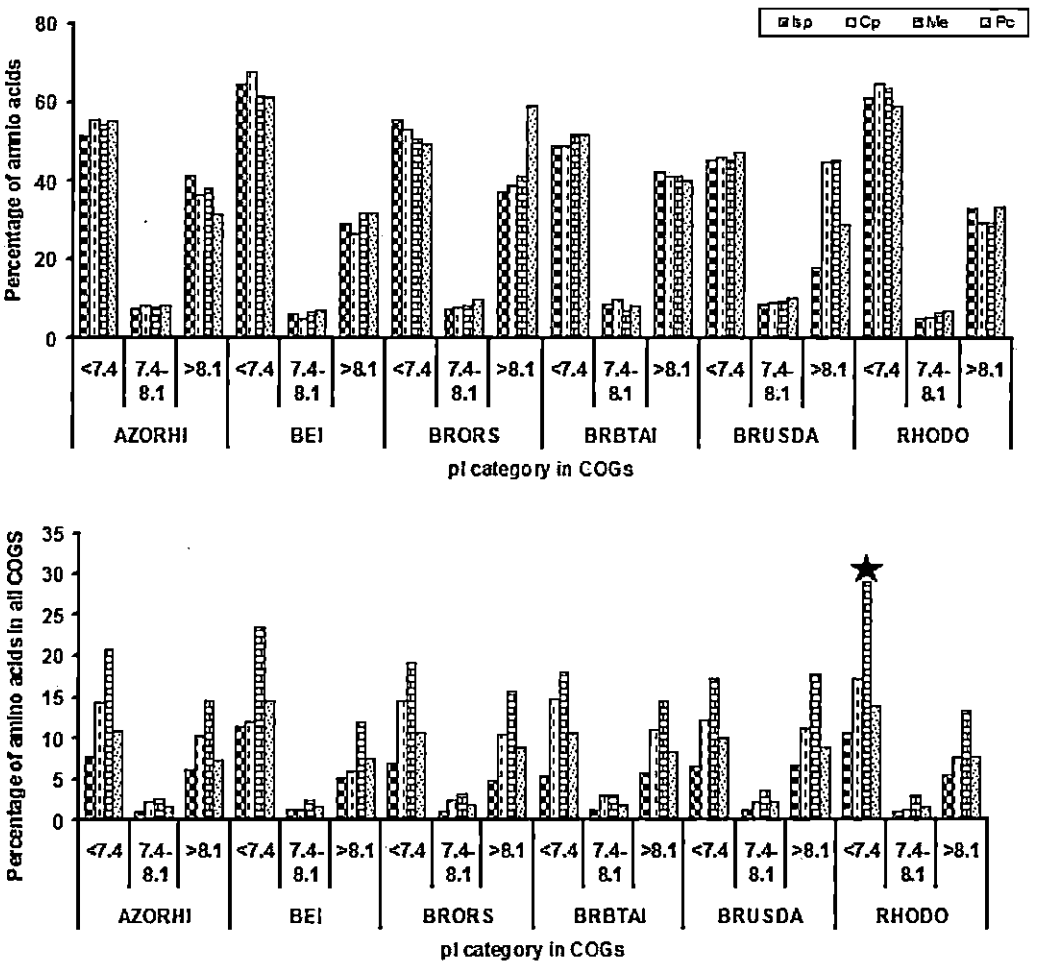

Figure 4.4.12(A-B): Distribution of pI in different COG clusters for *Azorhizobium*, *Beijerinckia*, bradyrhizobial strains and *Rhodobacter*. Colors as per Figure 4.4.3 (A-B).

COGs associated to cellular processes (Cp) show considerable representation in *Nostoc* 7120 and *Lyngbya* information and storage processing (Isp) and metabolism (Me) COGs demonstrate higher prescence for *Nostoc* sp. 73102, *Anabaena variabilis*, *Nodularia spumigena*, *Trichodesmium* and *Synechococcus* implying the role played by different COGs in effecting the survival of these organisms in their respective ecological niches. This variation stems from habitat specificity and biogeographical locations of the microorganisms. The *Nostoc* strains are symbiotic with *Nostoc* 73102 isolated from Australia; *Lyngbya* has been found in marine salt marshes and fresh water of Germany; *Nodularia spumigena* in surface waters of Baltic Sea; *Trichodesmium* in coastal waters of North Carolina and *Synechococcus* from microbial mats of Yellow stone National Park (http://img.jgi.doe.gov/

cgi-bin/pub/main.cgi).

In the α-proteobacterial members, rhizobial strains showed more or less similar pattern of distribution of different COGs clusters as depicted in Figure 4.4.9. Me and Isp peaks in the acidic and basic class respectively. The allocations for mesorhizobial strains too are quite alike. Isp leads the acidic class while Me the basic one. Among the *Rhizobium* strains and mesorhizobia, lack of diversity with respect to pI categorization in COGs may be attributed to sharing similar environmental conditions and habitats. Interestingly, inspite of being associated with hosts they have an acidic tilt. This acidic nature may be due to the presence of acidic residues which play a part in steadiness of the proteins associated with the organism functionality. Acidic amino acid residues related to metabolic pathways, gene regulation, translation and expression play a vital task in maintaining the symbiotic environment for the rhizobial and mesorhizobial members. Kiraga *et al.,* (2007) reported that the host associated species may have an acidic tilt although they shift towards basicity in general. This may be attributed to the fact that loss of non essential genes coding for basic proteins might have directed the shift of the proteomes towards acidity. The classic examples are the sinorhizobial strains; especially *S. meliloti* which is more acidic compared to other symbiotic nitrogen fixers. While *S. medicae* contains huge number of amino acids in the acidic class, Me towers for same in SMELILO (Figure 4.4.10). Acidic residues related to metabolism play a crucial role in sinorhizobial members. GLUCO too has a large share of Me in acidic pI as well as basic pI. The endophytic bacteria *Gluconacetobacter diazotrophicus* (Cavalcante and Dobreiner, 1988) show some difference with respect to sinorhizobial and mesorhizobial strains. The organization of genes associated with nitrogen fixation, host association, sugar metabolism and plant growth promotion may be attributed to the higher representation of metabolism genes. Association of *Gluconacetobacter diazotrophicus* with a different host plant compared to legumes for the rhizobia may be a factor for diversity. In γ-proteobacterial members like *Azotobacter* and *Klebsiella* the findings are relatively similar with the exception of Me group in *Klebsiella* overlooking others in acidic class of isoelectric point (Figure 4.4.11). In case of γ-proteobacterial members the acidic

proteome set is quite expected for free-living ones and they are influenced by their respective ecological niches. The predominance of Me substantiates the action of acidic residues associated with those functions. Most of them thrive in terrestrial habitats competing with other organisms and in order to carry out essential life processes, proper functioning of cellular machinery and gene regulation is indispensable in a manner commensurate with the habitat requirements. However, they differ strongly with that of *Azoarcus*. *Azoarcus* has high number of amino acid residues in cellular processes group (Cp). In *Azoarcus* it is seen that COGs linked to cellular processes play a significant role. *Azoarcus* is an interesting endophyte of Kallar grass and rice. It colonizes their roots in a unique manner and resides within the aerenchyma (Hurek *et al.*, 1994). The amino acid residues linked to cellular processes help in root colonization as well as survival within the host without causing disease. The diversity of pI in COGs for *Azorhizobium*, *Beijerinckia*, bradyrhizobial members and *Rhodobacter* is shown in Figure 4.4.12. While *Azorhizobium* had lesser percentage of representation of COGs in the acidic class, *Beijerinckia* and *Rhodobacter* had huge numbers in the same, which is also statistically significant. However, Cp and Isp predominates the acidic and basic classes in these organisms. The biogeographical location of *Rhodobacter, Beijerinckia* as also their habitat specificity in their respective ecological niches makes the difference. The high percentage of amino acid residues related to metabolism in the acidic set is the outcome of its free-living nature and ability to detoxify metal oxides in the soil. This requires high metabolic efficiency. Intraspecific diversity is well defined for the bradyrhizobial members. The acidic and basic clusters are well balanced amongst the bradyrhizobial strains. However, estimation of all the COGs for amino acids demonstrated that metabolism (Me) group had higher representation compared to others. The findings obtained for pI categorization in COGs for bradyrhizobial members clearly depicts that host associated organisms have less acidic proteomes. The apparent shift of pI bias substantiates the result. The somewhat different result obtained for *Bradyrhizobium sp. ORS278* may be attributed to its existence as an endophyte as also its association with different hosts. The

categorization of pI in COGs allocation pattern does reveal a good degree of disparity amongst different members of Bacteria and Archaea, with respect to their habitat, environmental adaptations and host specificity.

4.4.2 Influence of amino acid adaptation

Amino acid adaptation index (AAAI) values showed wide variation among members of Bacteria and Archaea. The degree of variation in the AAAI values across the studied proteomes point out the level diversity at interspecific and intraspecific level. The nitrogen fixation related proteins had lower AAAI values in contrast to complete proteomes in most of the studied organisms barring bradyrhizobia, *Frankia* strains, clostridia and *Methanococcus maripaludis*. In these cases the AAAI values are higher in nitrogen fixing proteins compared to the whole proteome (Table 4.4.1). It has further observed that in case of symbiotic plasmids of *Rhizobium etli* CFN 42, *Bradyrhizobium* sp. BTAI1 and *Sinorhizobium meliloti* symbiotic plasmids A and B the AAAI values were 0.302; 0.205; 0.3217, 0.261 respectively. These values are much lower than obtained for the total proteome and the nitrogen fixing proteins. The pattern of variation in AAAI is quite prominent amongst members of cyanobacteria, α-

Table 4.4.2: Correlation values for Axis 1 of correspondence analysis of amino acid usage with hydropathicity and aromaticity

Organism	Correlation1	Correlation2
ANVAR	-0.708	IC
AZOR	0.359	IC
AZORHI	-0.309	-0.373
AZOTO	IC	IC
BEI	-0.551	0.586
BRORS	-0.80	-0.32
BRBTAI	0.773	0.291
BRUSDA	-0.477	IC
CFEROX	0.671	IC
CCHROM	-0.414	IC
CLIMI	-0.615	IC
CPHAEO	-0.715	IC
CPEO1	0.906	0.420
CTEPI	0.842	0.515
CACET	IC	IC
CBEI	0.668	IC
C7424	0.723	IC
C7425	-0.766	-0.173
C8801	0.73	IC
C51142	-0.738	IC
ACN14A	IC	0.411
CcI3	IC	-0.371
EAN1PEC	IC	-0.463
GLUCO	IC	IC
KLEB	0.885	0.302
LYN	0.436	IC
MESOBN	0.923	0.440
MESMAF	-0.86	-0.345
MHUN	0.735	IC
MMAR	0.754	IC
MMAZ	-0.821	-0.354
MTHER	0.722	0.449
NOSPUM	0.716	IC
N7120	0.6130	IC
N73102	-0.59	IC
RHICFN4	-0.875	0.445
RHICIAT	-0.886	-0.440
RHIVICIA	0.917	0.409
RHIWSM	IC	IC
RHODO	0.370	IC
SNOMEDI	-0.905	-0.411
SMELILO	-0.930	-0.40
SYNE	-0.649	-0.263
TRICHOS	-0.32	0.388

IC=Inconsequential;
Correlation1=Axis 1 with hydropathicity;
Correlation2=Axis 1 with aromaticity

proteobacteria, β-proteobacteria, γ-proteobacteria, green sulphur bacteria, clostridia, actinobacteria and the methanogenic archaea. The variation is low at intraspecific level compared to that at the interspecific level. Interestingly, at the intraspecific level even organisms surviving in different ecological niche showed less diversity in AAAI values. *Azotobacter*, *Azoarcus*, *Klebsiella* and *Synechococcus* and a number of free living organisms had greater AAAI values compared to the symbiotic ones. The pattern obtained for free-living organisms and symbiotic plasmids implied that higher AAAI containing proteomes are better adapted for shaping frequency of amino acids in proteins and free living ones score over symbiotic ones in this regard. Similarly nitrogen fixing related proteins had reduced level of flexibility in influencing incidence of amino acids in proteins compared to whole proteomes. When AAAI, pI and energetic costs of the proteins were correlated with GC3 and GC content no significant results were obtained. pI values also did not reveal noteworthy correlation with aromaticity and hydropathicity. All these pointed out that codon and amino acid usage is not influenced by AAAI for these organisms. Energetic costs of proteins did not reveal anything significant and low degree of variation in energetic costs of proteins revealed that some level of homogeneity subsists at the interspecific and intraspecific echelon, with regards to metabolic efficiency of these proteomes in carrying out reactions.

4.4.3 Hydropathicity and aromaticity influences amino acid usage

The principal axis of correspondence analysis for amino acid usages showed biologically important correlations with hydropathicity and aromaticity. The findings obtained for these correlations are tabulated in Table 4.4.2. Hydropathicity levels in more or less all the organisms showed strong positive as well as negative correlations compared to aromaticity levels. The strong positive and negative correlations of principal axis of variation of amino acid usage with hydropathicity indicate that genes associated with hydrophobic (positive) and hydrophilic (negative) proteins are preferred by translationally optimal codons. Hydropathicity appears to be strong force in controlling amino acid usage in these organisms. Likewise the correlations of Axis 1 with aromaticity levels revealed that amino acid composition can be relevant for evaluating amino acid usage in these

microorganisms. However, lack of correlation of pI with aromaticity and hydropathicity meant that isoelectric point seldom influenced aromaticity and hydropathicity of amino acids. GC/GC3 richness is not controlled by isoelectric point in nitrogen fixing proteomes. Again, results obtained for correlation of GC and GC3 content with aromaticity and hydropathicity pointed out that GC compositional content do not play a crucial role in amino acid usages of nitrogen fixing organisms.

4.5. Phylogenetic analysis of *nif* genes: An alternative approach

4.5.1 Condensed matrix based phylogeny

The Laa values depicted in Table 4.5.1, signifying the length of the amino acids for the studied *nif* genes are fairly constant. Figures 4.5.1-4.5.5 represent the phylograms constructed for 16S rRNA, *nif*H, D & K genes and whole genomes using distance matrices. The distance matrices are given in A4.5.1-A4.5.5 (See appendix).

In case of 16S rRNA genes, the phylogram revealed a topology that is somewhat different from the existing phylogenies based on sequence-based approaches. From Figure 4.5.1 it is seen that some members of one lineage has been placed among the members of another lineage. Archaeal members belonging to the methanogens remain scattered between members of cyanobacteria, and proteobacteria. However, the distribution patterns of the eubacterial members are not that speckled. Although, majority of the proteobacterial members are clubbed together in a major clade, yet *Azoarcus* BH72 and *Klebsiella pneumoniae* are placed in a different manner. The former is placed alongside *Cyanothece* sp. PCC7424 while the latter is placed along with *Cyanothece* sp. PCC 8801. Similar condition is noticed in case of actinobacteria, *Frankia* ACN14a in between two species of *Cyanothece.* The two clostridial species are not placed together, rather *Clostridium beijerinckii* occurs on a branch with *Trichodesmium erythraeum* while *C. acetobutylicum* occurs in a subclade with *Chlorobium phaeobacteroides.* Majority of the chlorobial members belonging to the green-sulfur bacterial group are sister to the members of actinobacterial *Frankia.* Most of the cyanobacterial species undertaken in this study are placed together, barring *Cyanothece* PCC 7424, *Cyanothece* PCC 7425, *Nostoc* sp. and *Anabaena variabilis.* The latter is placed in a different clade altogether.

The pattern obtained in the phylogram for 16S rRNA genes based on leading eigen values seem interesting at the very first glance. Especially the placements of *Anabaena variabilis* as isolated, *Azoarcus* BH72 with *Cyanothece* PCC7424, *Frankia* ACN14 with cyanobacteria, *Clostridium beijerinckii* with *Methanococcus maripaludis*,

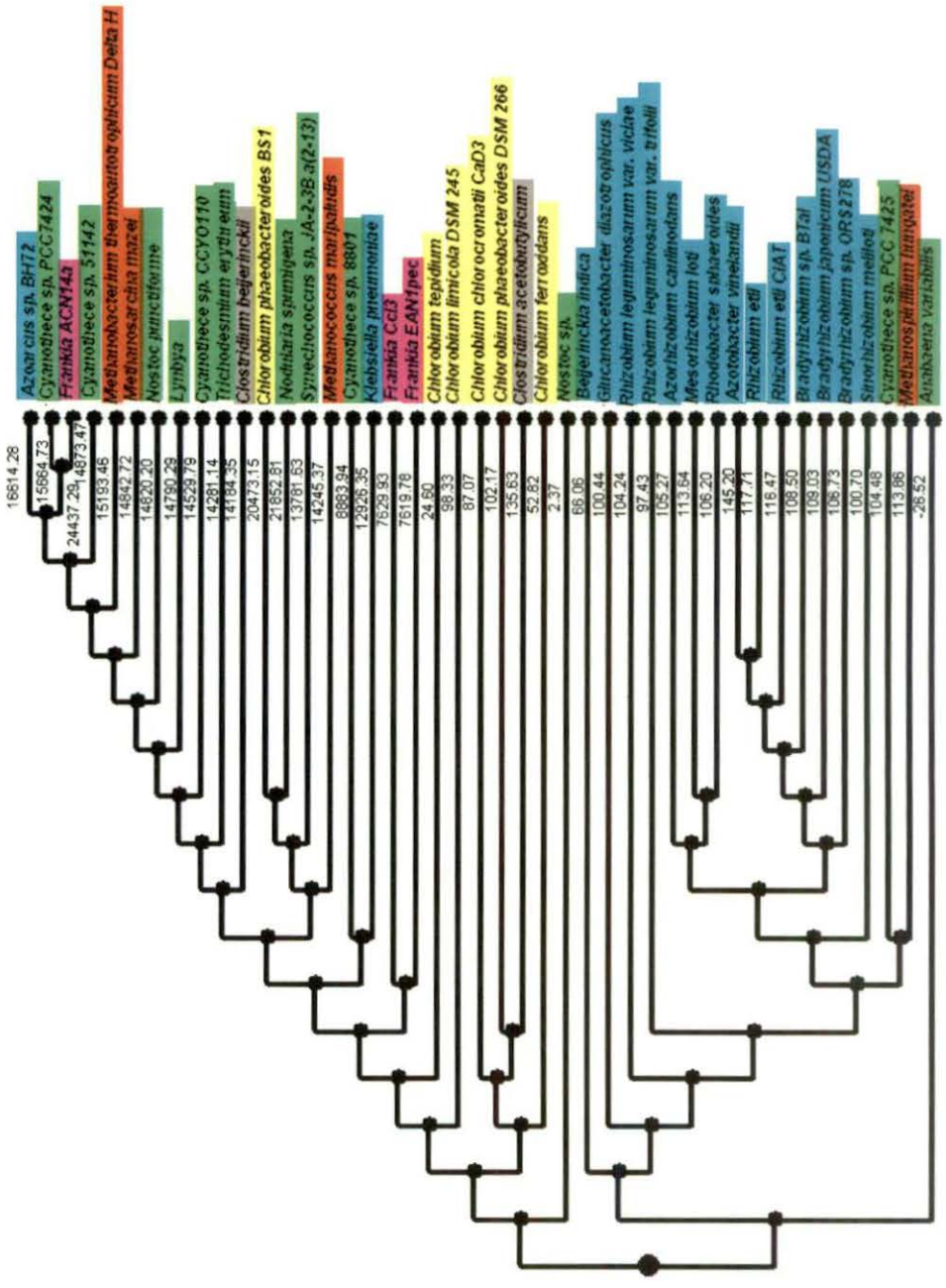

Figure 4.5.1: Phylogram of the 16S rRNA genes. Numbers represent root distances. Cyanobacterial species are marked by light green background, proteobacterial species by light blue background, actinobacteria in pink, firmicutes in grey, green sulphur bacteria in light yellow and methanogenic archaea in light brown.

Klebsiella pneumoniae near to *Cyanothece* PCC 8801, clostridia with members of other lineages. These interesting findings of similarities and dissimilarities due to leading eigen value vectors points out very clearly that dissimilar descriptors and invariants (Randic *et al.*, 2001) decide upon diverse structural features of DNA sequences. Some workers (Randic *et al.*, 2001) pointed out that species similarity or dissimilarity ought not to be depending upon a sole descriptor or sole gene. Taking this fact into consideration

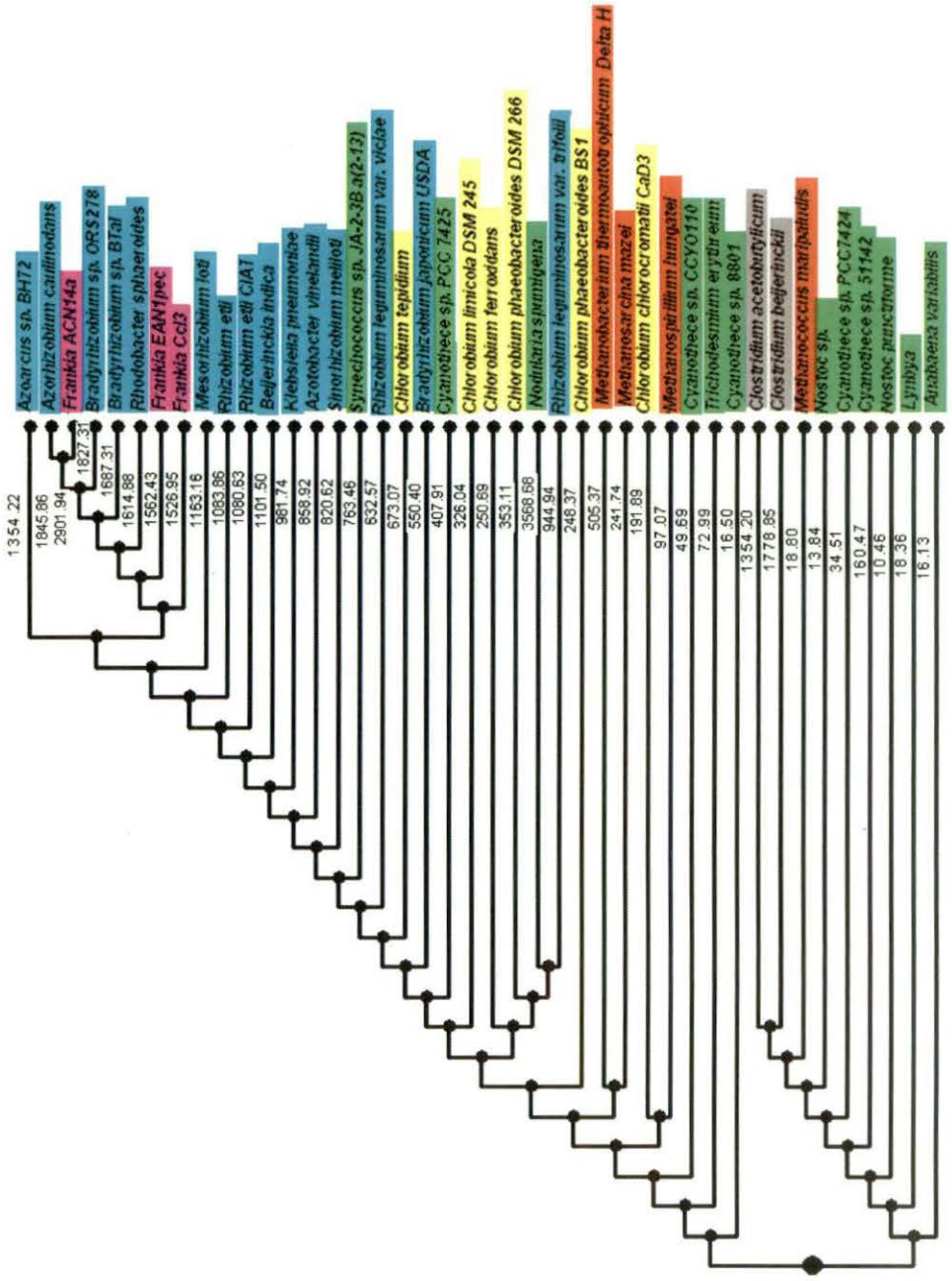

Figure 4.5.2: Phylogram of *nif* H genes. Numbers indicate root distances. Background colors as per Figure 4.5.1.

the results are quite novel and significant since, this technique is relatively unique in portraying the features of DNA nucleotide sequences. The interesting fact is that members of one lineage are placed among members of other lineage. The findings may be the consequence of ancient paralogous copies that might have separated during the course of time. The apparent structural similarity between the aforesaid members may be an indication that they have been derived from some common ancestral gene and have

diverged due to gene duplication events. The placement of *Methanococcus, Methanospirillium* and *Methanobacterium thermoautotrophicum* close to cyanobacterial associates merely indicate that these sequences might have arose preceding the separation of Eubacteria and Archaebacteria. These results thus seem to break the monophyly rule applicable for 16S rRNA genes. Again these genes have close lengths and their occurrence in the same clade may be attributed to the length, which may be related to their molecular evolution. The results obtained for 16S rRNA strengthen the view of Randic *et al.,* (2001) that a single gene cannot be a pointer of evolutionary similarity or dissimilarity. Each molecule reveals its history and traces back to the last common ancestral state (Koonin and Martin, 2005). It is possible that the molecular ancestors were present in diverse organisms during the course of evolutionary time period. Even though 16S rRNA's are used as phylogenetic markers, a number of studies (Syvanen, 1987; Yap *et al.,* 1999; Wang and Zhang, 2000; van Berkum *et al.,* 2003; Boucher *et al.,* 2004; Eardly *et al.,* 2005) suggest that they can be horizontally transferred too. It is therefore highly likely that these 16S rRNA genes reflect only a part of their evolutionary history, not addressed by the earlier studies based on sequence alignment.

Figures 4.5.2-4.5.4 shows the phylograms of *nif*H, D & K genes in the studied organisms. The overall topology of the genes showed a pattern quite comparable to that obtained for 16S rRNA genes with respect to the presence of members of one lineage among other lineages. The placements of the archaeal members vary in each of the genes. In most cases they remain scattered indicating that they are not monophyletic. In *nif*H genes, two clostridial members lie in the same branch and occur as sister to some cyanobacterial members while in *nif*D clostridial members lie in different clades. *C. beijerinckii* is clustered with *Nodularia spumigena* while *C. acetobutylicum* is present in between two bradyrhizobial species. In *nif*K they remain scattered. In case of *nif*H genes the actinobacterial members, *Frankia* CcI3 and *Frankia* EAN1pec are clustered together and remain in between proteobacterial members while; *Frankia* ACN14a is associated with *Azorhizobium caulinodans*. In *nif*D genes, *Frankia* CcI3 and EAN1pec are

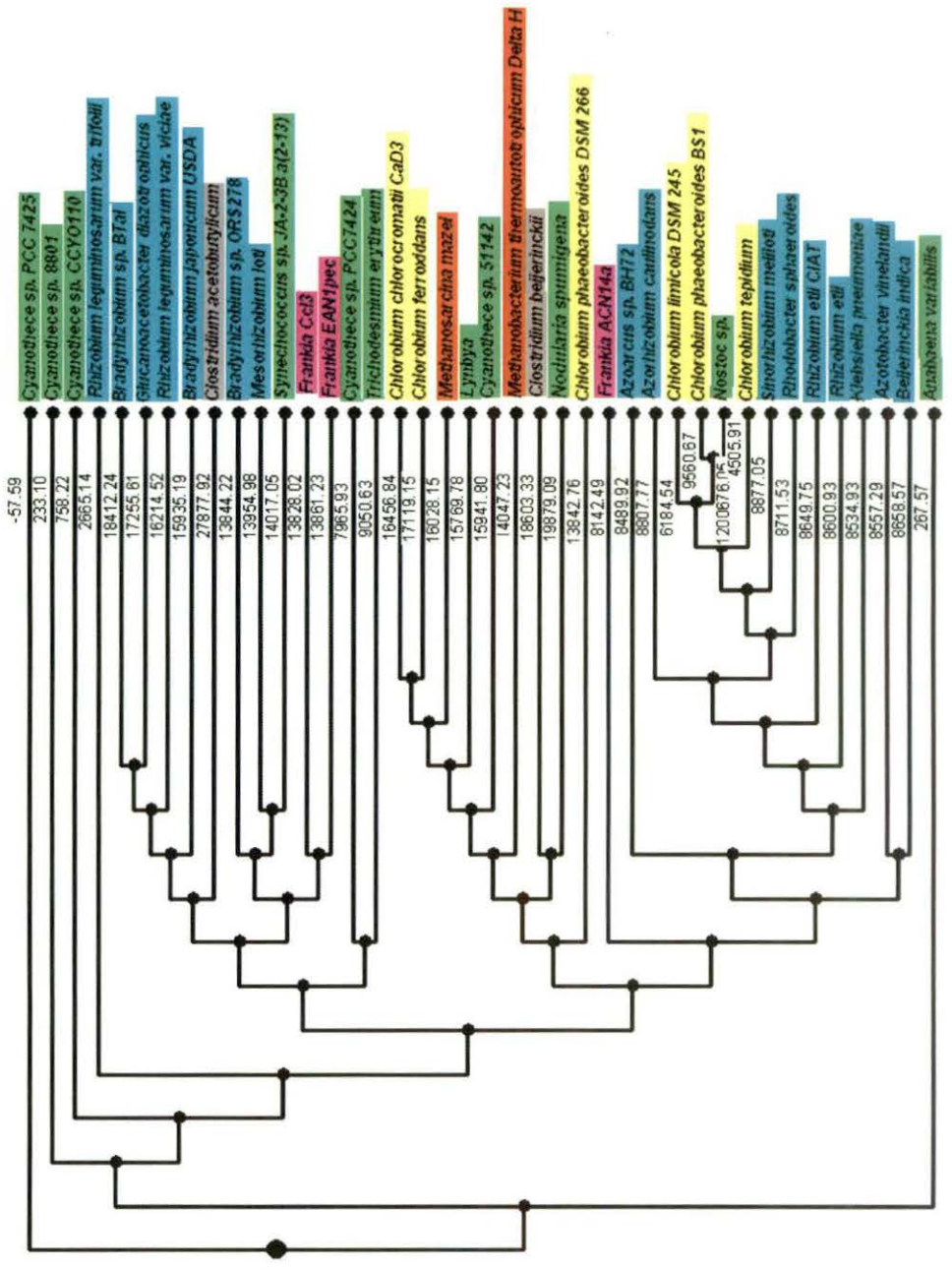

Figure 4.5.3: Phylogram of *nif*D genes. Root distances represent numbers. Numbers indicate root distances. Background colors as per Figure 4.5.1.

clustered together and lay sister to cyanobacterial and proteobacterial members while *Frankia* ACN14a is sandwiched between *Chlorobium phaeobacteroides* and *Azoarcus* BH72. *Frankia* Ccl3 and *Frankia* ACN14a remain close to each other and *Frankia* EAN1pec lies between *Rhodobacter sphaeroides* and *Chlorobium tepidium* for the *nif*K genes. The green sulphur bacterial members of *Chlorobium* are interspersed between organisms of other lineages. The pattern of placements of the cyanobacterial members vary in the

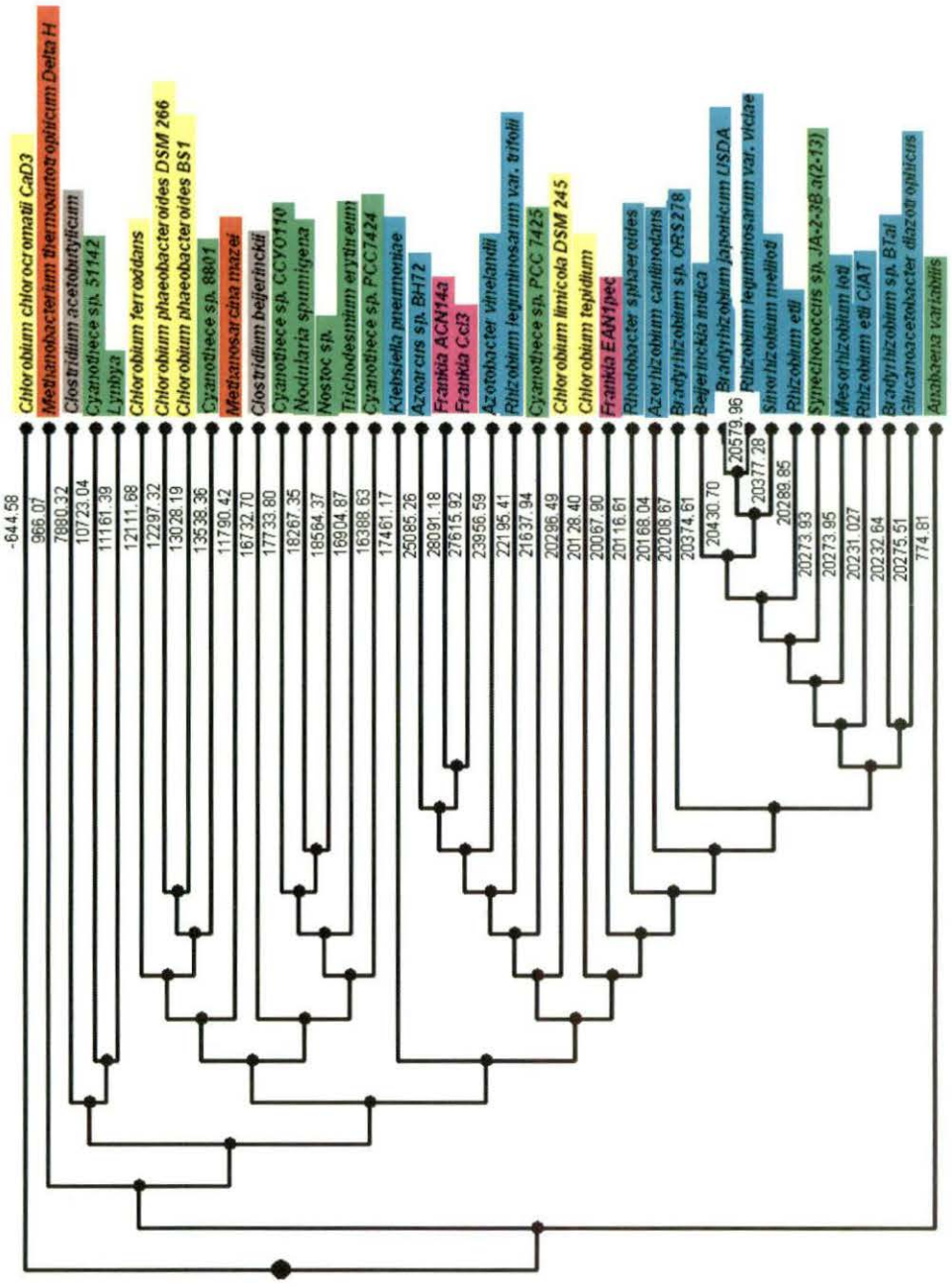

Figure 4.5.4: Phylogram of *nif* K genes. Numbers above represent root distances. Numbers indicate root distances. Background colors as per Figure 4.5.1.

three genes. However, some members of cyanobacteria lie amongst affiliates of other lineages indicating polyphyly. While a number of proteobacterial associates are clustered together in branches in the *nif*H genes other members like *Rhizobium leguminosarum bv. viciae*, *R. leguminosarum bv. trifolii* and *Bradyrhizobium japonicum* USDA 110 lie strewn between other bacterial lineages. The proteobacterial bacterial lineages for *nif*D genes specify the polyphyletic origin. Similar nature is

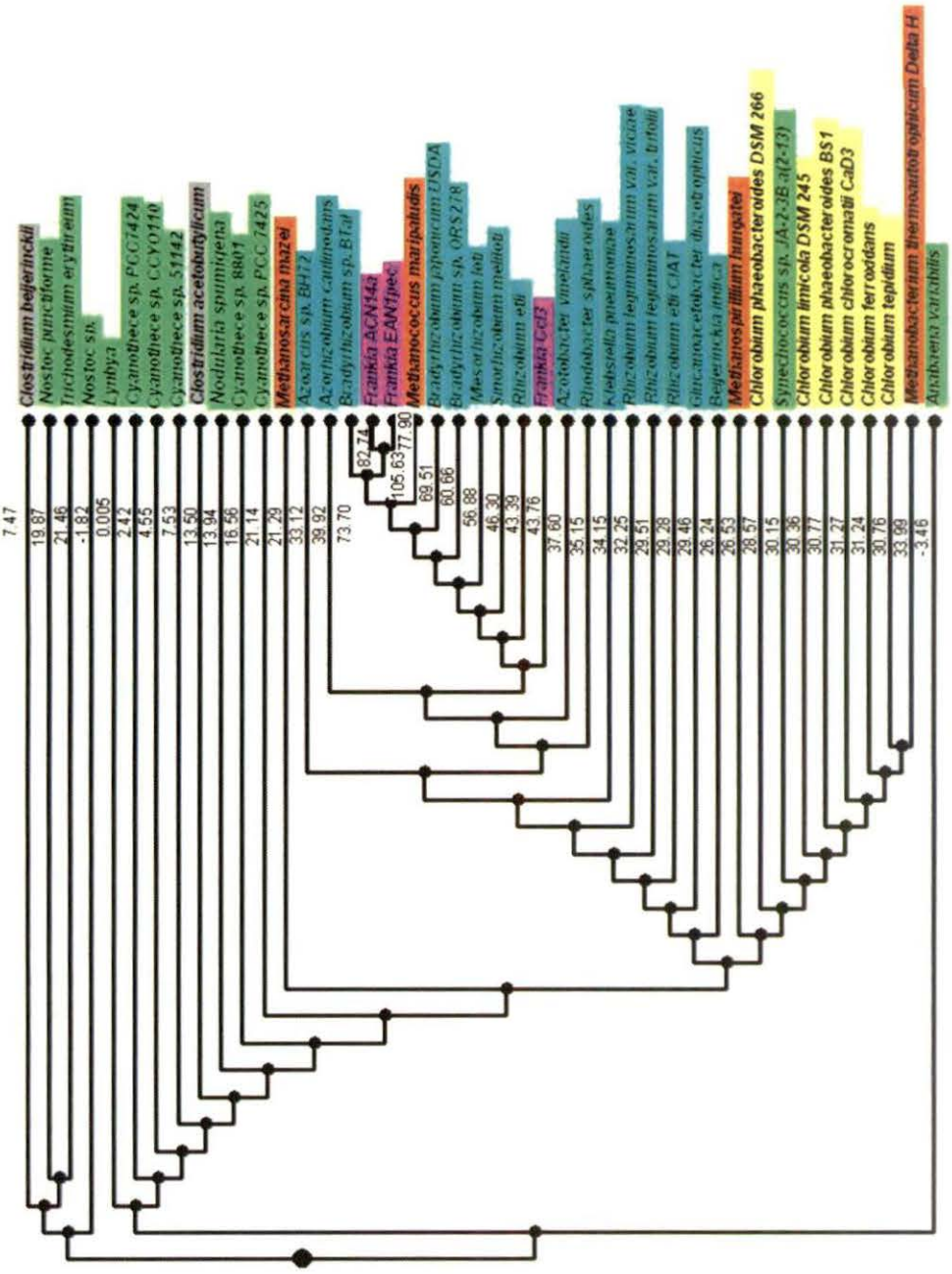

Figure 4.5.5: Phylogram of whole genomes. Root distances represent numbers. Numbers indicate root distances. Background colors as per Figure 4.5.1.

observed for the *nif*K genes as evident from Figure 4.5.4. None of the proteobacterial subgroups, be it alpha, beta or gamma proteobacteria are supported to be monophyletic.

It has been observed by previous workers (Normand and Bousquet, 1989; Fani *et al.*, 2000; Kechris *et al.*, 2006) that the core nitrogen fixation genes are horizontally transferred. The findings for *nif*H, D & K genes strongly support this owing to the mosaic pattern exemplified in the phylograms. Had lateral gene transfer not occurred major

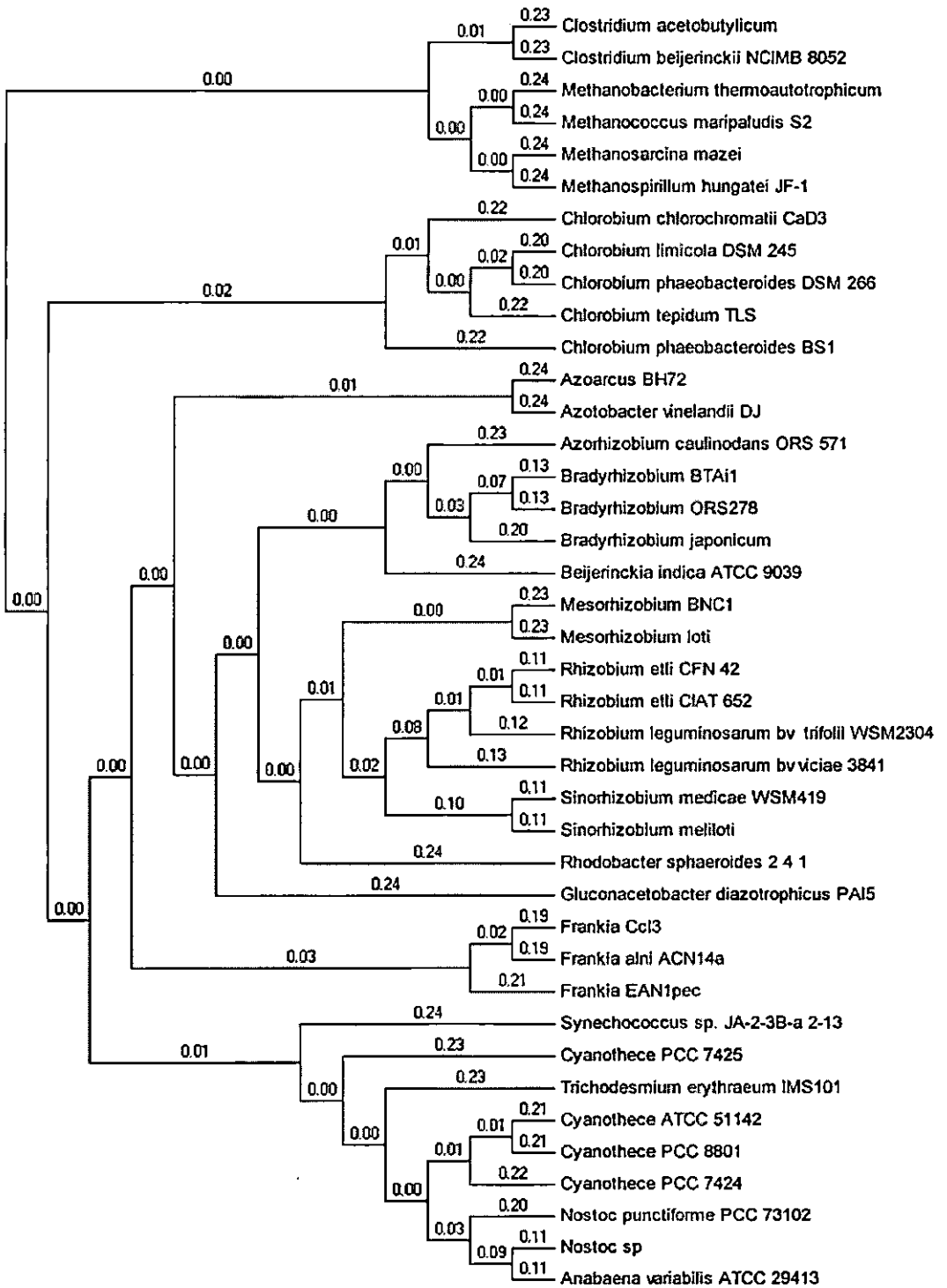

Figure 4.5.6: Topology of whole genomes for nitrogen fixing organisms obtained from CVTree. Numbers represent branch length.

lineages like actinobacteria, proteobacteria, firmicutes, green sulfur archaebacteria, cyanobacteria, bacteria and low G+C containing gram-

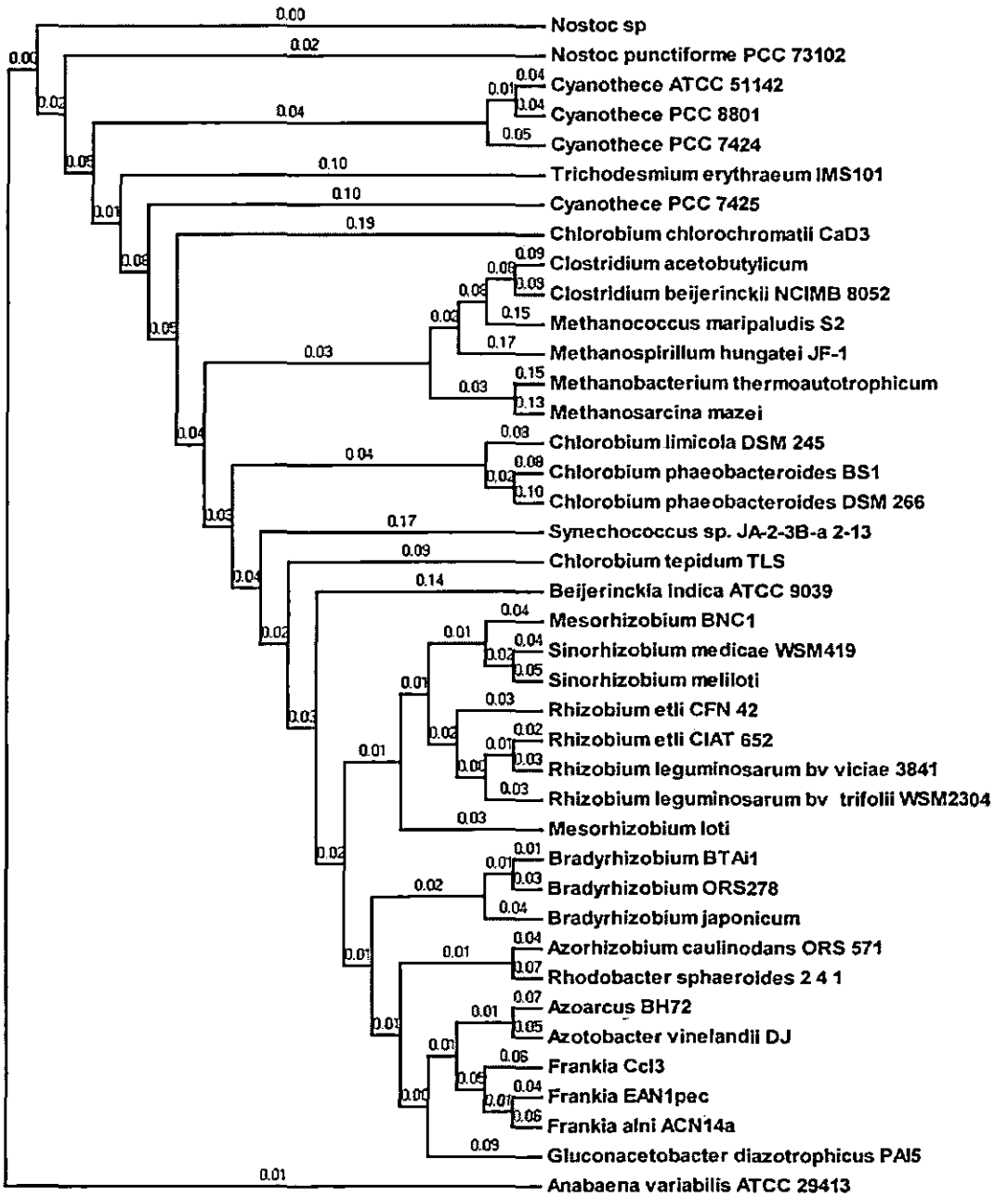

Figure 4.5.7: Topology of whole proteomes for nitrogen fixing organisms obtained from CVTree. Numbers represent branch length.

positive bacteria would have been monophyletic. Yet, it was found that members of one lineage are placed among members of other lineage suggesting that *nif*H, D & K phylogenies are polyphyletic. These results for the said genes by condensed matrix method reveal that even some members of cyanobacteria, α, β, γ proteobacteria, actinobacteria, chlorobia, firmicutes and methanobacteria are not monophyletic.

Interestingly it has been noticed that archaebacterial members are present along with eubacterial members. Their placements are attributed to gene duplications in the ancestral state that might have resulted before the division of archaebacterial and eubacterial members. Again, the mosaic pattern exhibited in the placements of some chlorobial, actinobacterial, cyanobacterial, proteobacterial and clostridial taxa may be due to the sharing of universal evolutionary history. Moreover, it is reasonable that they were paralogous copies. It is quite possible that gene duplication events occurring during the evolutionary history has given rise to the present *nif* genes. The origin of the *nif* genes by paralogous duplication events has been previously reported by Fani *et al.*, (2000). In this context the outcomes are quite significant since, all three codon positions in nucleotide triplets were considered, giving them equal weight, thus incorporating their influence during evolution. Accordingly, these findings support the view of Randic *et al.*, (2001) that the various structural properties of a DNA sequence are guided by different descriptors and invariants (Randic *et al.*, 2001).

The phylogram of whole genomes for studied bacterial members are depicted in Figure 4.5.5. The pattern illustrated by the phylogram is somewhat different from that obtained for *nif*H, D & K genes. The archaeal members are scattered amongst members of other lineages. The clostridial affiliates remain scattered. *C. acetobutylicum* occur in between cyanobacterial members while *C. beijerinckii* occur as part of a subclade. *Frankia* sp. ACN14a and *Frankia* sp. EAN1pec are clubbed together and occur as sister to some proteobacteria while *Frankia* sp. CcI3 lie in between proteobacterial members. Most of the cyanobacterial affiliates are present very close to each other with the exception of *Synechococcus* sp. JA-2-3B and *Anabaena variabilis* ATCC 29413. *Synechococcus* JA-2-3B is placed between the chlorobial species while *Anabaena variabilis* lie in a different clade altogether. The majority of the chlorobial species are placed together. They occur as sister to the proteobacteria. Most of the proteobacterial members are placed together. However, *Frankia* and *Methaonococcus maripaludis* are placed in between the proteobacterial affiliates. The whole genome phylogram however does not reveal the vibrant mosaic pattern noticed for *nif*H, *nif*K and *nif*D

genes. Studies of non-nitrogen fixing microorganisms spanning Bacteria and Archaea, using 16S rRNA and whole genomes (data not shown) applying this methodology revealed that distant organisms remain separated from each other.

The findings for whole genomes of the studied microorganisms have something to offer. Studies on comparative genomics have revealed the origin of all life forms from the last common universal ancestral state i.e. LUCAS (Koonin, 2009). Owing to gene duplication events (Ohno, 1970), which are regarded as a major evolutionary force, genomes have numerous copies of paralogous genes grouped into different families (Huynen and Nimwegen, 1998). Here, the occurrence of methanogenic archaeal members scattered amongst other lineages; placements of clostridial members amongst the cyanobacterial ones; placement of *Frankia* CcI3 and *Synechococcus* JA-2-3B between the chlorobial members indicate very clearly that these organisms share common evolutionary history. It also point out that these organisms may contain huge number of ancient paralogous genes and a very significant part has been played during primitive stages of molecular evolution by in tandem gene duplication (Fani *et al.*, 2000) events within the DNA. It is possible that evolution of the duplicated genes might have been by subfunctionalization in the genomic complement (Lynch and Force, 2000; Lynch and Katju, 2004) by which they preserved the discrete subsets of the original functions of genes found in LUCAS. Findings for the whole genomes agree with the fact that single gene duplications, whole genome duplications, lateral gene transfer, genome rearrangements are vital events in molecular evolution (Koonin, 2009).

4.5.2 Composition vector based phylogeny

The Neighbour-joining (NJ) topologies obtained for the whole genomes and whole proteomes of nitrogen fixing microorganisms are depicted in Figures 4.5.6-4.5.7. All the archaeal members are placed together and lay sister to clostridia. The clostridial members are clubbed together. The actinobacterial *Frankia* strains lay collectively and sister to proteobacteria. Most of the chlorobial members remain together with the exception of *C. chlorocromatii* CaD3 and *C. tepidium* TLS. The former lie in between cyanobacterial and clostridia while the latter is placed in

between *Synechococcus* sp. JA-2-3B and *Beijerinckia indica*. Although majority of cyanobacterial members are grouped together, *Synechococcus* sp. JA -2-3B lie between lineages of other bacteria and *Anabaena variabilis* lie differently. All proteobacterial members remained close together. Whole proteome analysis illustrates the grouping of cyanobacteria, proteobacteria, clostridia, methanogenic archaea, and green-sulphur bacteria lineages.

The results obtained for the whole genomes and proteomes of the nitrogen fixing microorganisms using the CV method is different from that obtained for the condensed matrix method. While genome phylogeny reported the placement of *Chlorobium chlorocromatii* CaD3, *Chlorobium tepidium* TLS among clostridial ones and placement of *Synechococcus* sp. JA-2-3B between lineages of other bacteria and *Anabaena variabilis* as outgroup, whole proteome phylogeny did not report any mixing. This technique offers very little to throw light on the events like lateral gene transfer, genome duplications, genome rearrangements etc. that has been emphasized by the nucleotide triplet based molecular phylogeny. It has been reported (Qi *et al.*, 2004) that lateral gene transfer events do not influence the methodology to a large extent thus exposing its weakness.

4.5.3 The probable scenario

Nitrogen fixation by microorganisms is an ancient property arising at the start of cellular evolution (Cloud, 1976; Broda and Peschek, 1983; Postgate, 1987). It has been reported that primitive microorganisms were heterotrophic anaerobes and utilised abiotic organic matter for metabolism (Fani *et al.*, 2000). It is well known that chemical evolution and organic evolution took place simultaneously. There are two schools of thought regarding the early atmosphere and the nature of primitive nitrogenase (Fani *et al.*, 2000; Normand and Bousquet, 1989). However, the hypothesis regarding the reducing nature of the early atmosphere containing huge quantity of ammonia (Fani *et al.*, 2000) is more accepted. Some researchers (Normand and Bousquet, 1989) on the other hand addressed that primitive nitrogenase was an iron dependent molybdenum free compound. With the change in Earth's atmosphere i.e. from reducing to and oxidising one, the primitive iron dependent and molybdenum free nitrogenase (Normand and Bousquet,

Table 4.5.1: GC content, GC3, Nc and Laa values for the, *nif* H, *nif* K and *nif* D genes in the studied organisms

Organism	Gene	GC%	GC3%	Nc	Laa
Anabaena variabilis ATCC 29413	*nif* H	48.3	47	44.86	296
	nif K	46.4	44.9	50.28	512
	nif D	47.9	47.5	47.89	491
Azoarcus sp. BH72	*nif* H	62.9	83.7	28.11	297
	nif K	60.8	89	27.96	522
	nif D	60.6	86.2	29.03	491
Azorhizobium caulinodans ORS 571	*nif* H	64.3	91.9	29.97	296
	nif K	63.3	90.5	28.86	519
	nif D	62.3	90.1	29.51	500
Azotobacter vinelandii AvOP	*nif* H	59.1	77.8	27.67	290
	nif K	59.4	84.2	29.82	523
	nif D	59.5	82.4	28.09	492
Beijerinckia indica subsp. indica ATCC 9039	*nif* H	60	79.5	31.36	290
	nif K	56.8	74.1	35.36	520
	nif D	56.2	73.8	35.28	486
Bradyrhizobium sp. BTAI1	*nif* H	63.5	87.9	31.11	295
	nif K	62	87.2	30.85	519
	nif D	61.7	89.3	30.93	500
B. japonicum USDA 110	*nif* H	55.2	66	52.12	294
	nif K	55.9	68	47.21	518
	nif D	54.6	66.6	48.98	500
Bradyrhizobium sp. ORS278	*nif* H	64.6	90.8	29.88	295
	nif K	62.7	92.2	28.79	500
	nif D	62.4	89	28.58	519
Chlorobium chlorochromatii CaD3	*nif* H	45.5	38.4	39.96	274
	nif K	48.2	47.3	45.84	485
	nif D	46	43.5	49.59	544
C. ferrooxidans DSM 13031	*nif* H	49.4	50.2	47.55	274
	nif K	52.9	58.7	48.88	460
	nif D	49.8	57.5	49.23	544
C. limicola DSM 245	*nif* H	52.1	56.7	44.68	274
	nif K	54.5	67.1	44.89	459
	nif D	52.7	64.3	43.14	546
C. phaeobacteroides DSM 266	*nif* H	52.3	57.3	42.81	274
	nif K	51.9	57.8	49	459
	nif D	49.8	56.5	49.3	546
C. phaeobacteroides BS1	*nif* H	51.1	54.2	50.6	274
	nif K	53	58.3	48.48	460
	nif D	49	57.4	49.05	540
C.tepidum TLS	*nif* H	57.3	72	35.55	274
	nif K	59.8	80.9	34.11	459
	nif D	56.3	77.3	35.53	543
Clostridium acetobutylicum ATCC 824	*nif* H	35.9	14.9	35.65	272
	nif K	36.3	22.5	41.6	531
	nif D	35.5	19.7	38.79	454
C.beijerinckii NCIMB 8052	*nif* H	33.7	10.7	32.92	273
	nif K	33.1	10.6	33.2	528
	nif D	32.5	14.1	34.51	454
Cyanothece sp. PCC 8801	*nif* H	45.6	40.2	34.86	296
	nif K	45.3	40.9	40.21	511
	nif D	44.9	40.7	39.35	476
Cyanothece sp. ATCC 51142	*nif* H	45.2	36.1	35.9	327
	nif K	42.7	34.8	35.67	511
	nif D	45.8	41.2	36.92	480
Cyanothece sp. CCY0110	*nif* H	44.8	35	33.25	290
	nif K	42.9	35	35.47	480
	nif D	45.9	41.2	37.35	511
Cyanothece sp. PCC 7424	*nif* H	45	37.2	47.23	299
	nif K	45.8	40.5	45.31	437
Cyanothece sp. PCC 7425	*nif* H	54.1	62.6	41.41	298
	nif K	54.9	65.9	42.29	511
	nif D	51.4	56.7	42.5	494

Continued to the next page

Table 4.5.1: Continued from previous page

Organism	Gene	GC%	GC3%	Nc	Laa
Frankia alni ACN14a	*nif*H	67	92.7	30.67	331
	*nif*K	68	98.6	28.38	520
	*nif*D	65.4	94.8	29.46	486
Frankia sp. CcI3	*nif*H	64.2	93.4	29.25	287
	*nif*K	67.9	94.6	30.84	518
	*nif*D	64.8	93.1	31.35	486
Frankia sp. EAN1pec	*nif*H	64.5	93.4	28.41	289
	*nif*K	67.6	94.9	30.65	518
	*nif*D	65.9	95	29.46	486
Gluconacetobacter diazotrophicus	*nif*K	61	82.8	33	511
PAL5	*nif*D	59.1	79	31.94	499
Klebsiella pneumoniae pneumoniae	*nif*H	58.5	76.2	34.93	293
	*nif*K	58.5	76.4	36.18	520
	*nif*D	58.4	75.5	34.94	483
Lyngbya sp. PCC 8106	*nif*H	46.1	40.9	48.86	300
	*nif*K	46	43.6	45.89	511
	*nif*D	48.1	47.6	48.64	480
Mesorhizobium loti MAFF303099	*nif*H	60.7	82.6	36.14	297
	*nif*K	59.5	76.7	40.58	513
	*nif*D	59.6	80.6	37.31	501
Methanobacterium thermoautotro-	*nif*H	50.6	52.7	45.58	265
phicum Delta H	*nif*K	44.4	37.5	46.57	459
	*nif*D	45	42.7	43.99	469
Methanococcus maripaludis	*nif*H	40.2	29.8	33.13	276
Methanosarcina mazei strain Goe1	*nif*H	49.6	50.2	47.53	273
	*nif*K	51	52.4	49.59	456
	*nif*D	47.3	52.2	46.75	532
Methanospirillum hungatei JF-1	*nif*H	44.8	34.9	50.43	280
	*nif*K	47.2	37	49.81	455
	*nif*D	45.9	36.6	47.22	493
Nodularia spumigena CCY9414	*nif*H	45.3	30.9	36.88	145
	*nif*K	44.2	40	42.31	511
	*nif*D	43.5	38.5	42.95	480
Nostoc sp. PCC 7120	*nif*H	44.7	40.5	41.42	297
	*nif*K	46.4	47.5	38.9	512
	*nif*D	39.9	36.6	56	4049
N. punctiforme PCC 73102	*nif*H	46.9	42.3	38.9	297
Rhizobium etli CFN 42	*nif*H	60	78.4	39.72	297
	*nif*K	58.9	76.9	39.01	531
	*nif*D	59.1	78.8	37.82	496
R. etli CIAT 652	*nif*H	60	78.4	39.72	297
	*nif*K	59.8	78.6	37.36	513
	*nif*D	59	78.7	37.86	500
R.leguminosarum bv. trifolii	*nif*H	56.5	71.2	46.86	219
	*nif*K	55.3	64.7	51.87	510
	*nif*D	54.4	64.9	49.23	411
R. leguminosarum bv. viciae 3841	*nif*H	55.4	64.8	48.32	297
	*nif*K	54.3	62.6	52.17	513
	*nif*D	53.8	62.4	51.4	501
Rhodobacter sphaeroides	*nif*H	64	89.3	29.2	291
	*nif*K	64.4	91.9	27.01	506
	*nif*D	62.7	90.1	30.08	493
Sinorhizobium meliloti 1021	*nif*H	58.2	71.8	45.87	297
	*nif*K	56.4	67.3	48.62	513
	*nif*D	55.4	69.4	46.77	500
Synechococcus sp. JA-2-3B'a(2-13)	*nif*H	57.8	74.4	37.71	292
	*nif*K	58.8	73.7	42.27	511
	*nif*D	57.6	75.6	35.93	487
Trichodesmium erythraeum IMS101	*nif*H	43.2	27.6	37.52	296
	*nif*K	40.7	29.6	39.34	485
	*nif*D	41.4	29.9	36.62	512

1989) gradually modified itself to adapt to the new conditions. It has been stated (Fani *et al.,* 2000) that changing environment led to alteration in enzyme specificity, modification and adaptation that were retained by some eubacterial and archaebacterial members for survival. Changes associated with nitrogenase during the course of evolution might have resulted in alteration within the genes. The relevance of gene duplications in molecular evolution has been very well highlighted (Fisher, 1928; Ohno, 1970; Lynch and Katju, 2004). Results from whole genomes of bacteria and archaea point out that they enclose huge number of paralogous copies of genes (de Rosa and Labedan, 1998). The results for *nif* H, D & K genes are in line with the concept of gene duplications. They support the view of Fani *et al.,* (2000) that duplications in portions of genes, or operons or stretches of nucleotide sequences (Fani *et al.,* 2000) transformed the primitive nitrogenase to the present form and played a momentous role in producing genes with alike properties. The outcome of the studies also support occurrence of horizontal gene transfer across microbial lineages. It has been pointed out (Figge *et al.,* 1999; Lawrence, 1999) that inter as well as intra kingdom horizontal gene transfer play a vital role in shaping the genes for survival in the changing environment and thus molecular evolution. The findings based on condensed matrix method strongly supports it and agree with the views of Fay (1992) and Zehr *et al.,* (1997) that, during the course of evolution since ancient times, selective pressure and adaptation to constantly changing environment resulted in loss of genetic information in many microorganisms, implying that nitrogenase genes evolved alongside the cells carrying them.

4.5.4 *nif* genes did not evolve as a unit

The results for GC, GC3, Nc and Laa specify that there is a high degree of heterogeneity present amongst them suggesting that mutation and selection pressures are acting with unlike intensities within *nif.* If mutational processes and selection influenced all genes consistently there would have been higher GC3 content and consequently low Nc values in all the cases. These results further stipulate that the genes have been subjected to different recombination events and dissimilar selection pressures resulting in heterogeneity and never evolved as a unit.

4.6 Three dimensional homology models of *Frankia* NifH proteins

4.6.1 Template identification

The PSI-BLAST search found the crystal structure of the *Azotobacter vinelandii* NifH protein (1G5P) [PDB ID] (protein databank ID) to be the best template amongst others. This template match had an E-value of 2e-130 and an X-ray crystallographic resolution of 2.2 angstroms. It is a soil bacteria like that of *Frankia* and the A and B chains of this protein revealed 79%, 76% and 77% sequence identities with *Frankia* CcI3, *Frankia* ACN14a and *Frankia* EAN1pec nitrogenase iron proteins. The NifH proteins from the three *Frankia* strains had greater than 90% identities amongst them and divulged a high degree of conservation. Hydrophilic amino acids occupied the surface whereas the hydrophobic amino acids were confined within the core as expected. The modelled structures of the nitrogenase iron proteins from *Frankia* CcI3, EANIpec and ACN14a revealed their existence as homodimer and consisted of 287, 289 and 287 residues in each chain. The comparable results obtained for these proteins are based on the conservation of protein contacts (Normand *et al.,* 1988). Table 4.6.1 shows some of the important characteristics of the modelled proteins.

4.6.2 Features of the modeled proteins

Figure 4.6.1 shows the modelled structures of the NifH proteins from *Frankia*. The tertiary structure arrangement of the nitrogenase iron protein consists of two subunits that are connected at one surface to the cubane structure of 4Fe-4S cluster, which is the core functional region of the protein. Ligating cysteines are Cys 96 and Cys 130. The helix and sheets remain interspersed throughout the protein structure (Sen *et al.,* 2010). Figures 4.6.2 (a-c) show the regions of identity between the protein and template and the metal binding sites. These sequence conservations have been previously reported in Fe-protein family (Georgiadis *et al.,* 1992). The 4Fe-4S cubane cluster is coordinated by thiol ligands of Cys 96 and Cys 130 from each of the NifH protein subunits. These residues control conformational reaction of the iron proteins. The structures are chiefly supported by the main-chain atoms of residues 96 to 100 and 130-133, and the hydrophobic side-chains of the residues Ala 97, Val 129 and Phe

Table 4.6.1: Summary of the characteristics of the nitrogenase iron proteins from three *Frankia* strains

Feature	*Frankia sp. CcI3*	*Frankia sp. EANIpec*	*Frankia sp. ACN14a*
General characteristics			
Molecular weight	63126.9	63729	62992.7
Net Partial Charge	-29.265	-28.885	-26.562
Number of atoms	4394	4436	4382
Characteristics of secondary structure			
Alpha helix	37.80%	34.08%	36.93%
Extended strand	20.21%	21.97%	20.73%
Random coil	41.99%	43.94%	42.33%
Characteristics of the functional region of the proteins			
Interacting residues in Chain A	25	27	23
Interacting residues in Chain B	30	34	32
Interface area of Chain A	1693 angstroms	1871 angstroms	1669 angstroms
Interface area of Chain B	1644 angstroms	1772 angstroms	1536 angstroms
Total number of non-bonded contacts	128	162	112
Number of hydrogen bonds in the interface area	9	12	9

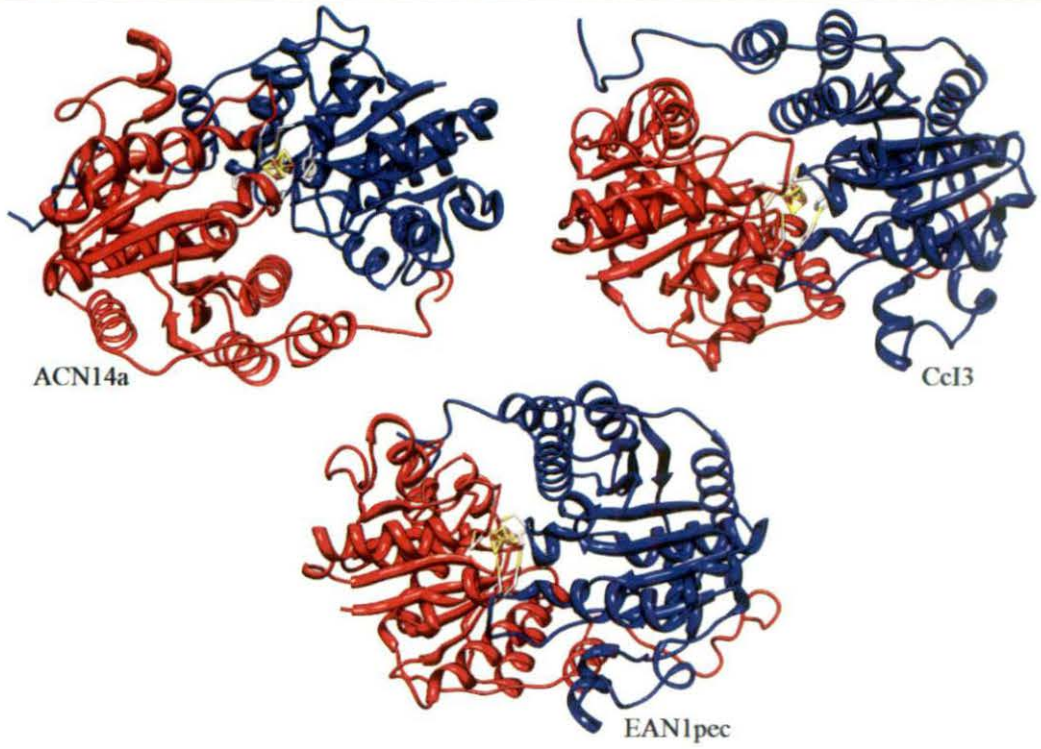

Figure 4.6.1: Ribbon diagram of the 3D structures of the nitrogenase iron proteins from *Frankia* strains obtained by molecular modeling.

133. These residues are vital for sustaining the suitable cluster setting. The cluster ligands in the structures are located at the amino-terminal end of the helices so that amide groups are adjusted in the direction of the cluster and ligands. Electrostatic interactions are favored. The nucleotide-binding section of the proteins consists of the residues in the Walker A motif (at the 8-15 residues), and the Switch 1 (at the 35 -41 residues) and Switch 2 regions (at the 122-126 residues). These three regions are connected by polar contacts, and are essential for maintaining the firmness of the proteins. They help in transmission of messages to other effector molecules. The pattern of residue conservation shown in Figure 4.6.2 revealed that a number of core residues, in the cluster, dimer interface, and nucleotide-binding regions are located in these regions. Figures 4.6.3 (a -c,) (please see page 126a) shows a plot of interaction of the ligands. CASTp analysis revealed the presence of 93, 105 and 90 pockets ligands interaction in each of the three NifH proteins from *Frankia* Ccl3, EAN1pec and ACN14a with varying area and volume. The varying area and volume of each pocket provided information on the protein

Figure 4.6.2a: Alignment of target protein (1G5P) and nitrogenase iron proteins from *Frankia* ACN14a. Residues highlighted in red show the conserved regions around the metal binding sites. Residues marked with * reveal the conserved portions.

surface features. Pockets on the protein surface, over and above definite amino acid arrangement inside it, crafts the inherent physico-chemical features indispensable for NifH protein functioning and residues in the active site pockets are indeed conserved compared to template.

4.6.3 The interface regions

There were two chains for the interacting interface in the functional region of the *Frankia* NifH proteins, and disulphide bonds and salt bridges were not recognized in the interface region. However, salt bridges were found in other regions. Table 4.6.1 shows the number of interacting residues in A and B chains, the total number of non-bonded contacts, number of hydrogen bonds, and the interface areas. Interactions across this interacting interface occurred between the polar

Figure 4.6.2b: Alignment of target protein (1G5P) and nitrogenase iron proteins from *Frankia* Ccl3. Color and * indications are as per Figure 4.6.2a

residues. However, no potential helix-turn-helix DNA binding motifs were identified using ProFunc server (Laskowski *et al.*, 2005). Functional analysis of the Fe-proteins revealed the presence of nests (Watson and Milner-White, 2002) in each chain. These nests are structurally important motifs found in functionally important regions of the protein structures. These nests appear like concave depressions serving as binding sites for atoms. In case of ACN14a, chains A and B of the NifH protein revealed the presence of 10 and 11 nests in each chain. In case of EAN1pec and Ccl3 the numbers of nests in each chain were 10.

The chains A and B of the NifH proteins of *Frankia* ACN14a, *Frankia* EAN1pec and *Frankia* Ccl3 found 36, 34 and 35 matching sequences in the PDB entries. Each of the chains had 17, 16 and 16

Figure 4.6.2c: Alignment of target protein (1G5P) and nitrogenase iron proteins from *Frankia* EAN1pec. Color and * indications are as per Figure 4.6.2a

homologues with genome location data for *Frankia* ACN14a, *Frankia* EAN1pec and *Frankia* CcI3. BLAST search for these proteins revealed 50 matching sequences in UniProt and 34 matching sequences in the PDB entries that had more or less similar type of function in other organisms with varying structural properties. Analysis of binding sites showed the presence of surface clefts and cavities in the proteins. Binding site analysis exposed the volume and percentage of the accessible and buried vertices. The proteins illustrated accessible surface with clefts and vertices for binding. The accessible surface is comparable to other Fe-proteins (Strop *et al.*, 2001; Howard and Rees, 1994; Howard and Rees, 1996; Schlessman, 1998) in microorganisms and reveals solvent accessibility of the proteins. Although

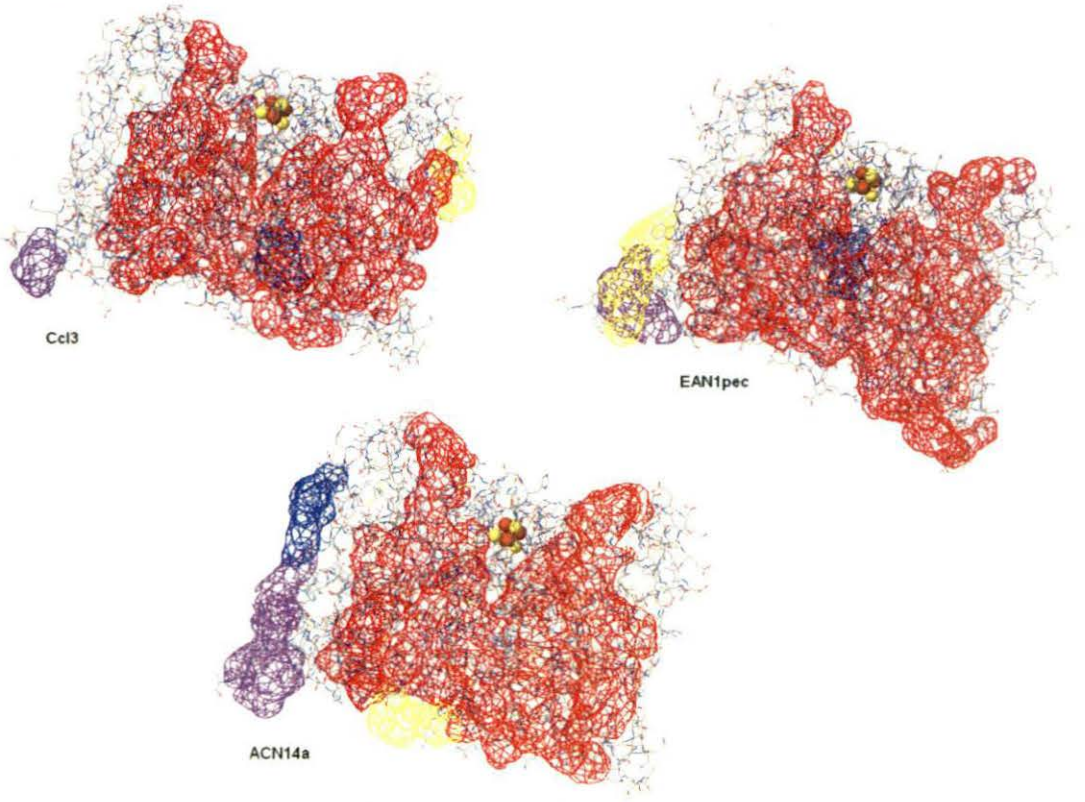

Figure 4.6.4: Clefts and cavities present on the surface of the nitrogenase iron proteins from *Frankia*

clefts and cavities are not directly related with the entrance or exit of substrates, yet they highlight the key residues in the NifH protein especially its functional site. The size of clefts in a protein's surface is one of the key factors determining the interaction of protein with other molecules. The clefts and cavities are biologically important, and reveal the active sites characterized by large and deep clefts and conserved residues in the protein. A cleft and several cavities were identified in the *Frankia* NifH proteins (Figure 4.6.4). The active sites in the cavities play a vital role in protein functioning in nitrogen fixation.

4.6.4 RMSD values

The backbone root mean square deviation (RMSD) calculations demonstrated that the NifH proteins from *Frankia* ACN14a, *Frankia* Ccl3 and *Frankia* EAN1pec had a deviation of 1.66, 1.72 and 1.56 angstroms in the C^{α} residues and 1.64, 1.68 and 1.54 angstroms in the backbone residues, from the template protein thus substantiating the sequence similarity result obtained from BLAST. RMS deviations were less evident in conserved regions indicating structural conservations. RMSD calculations among the three NifH proteins of

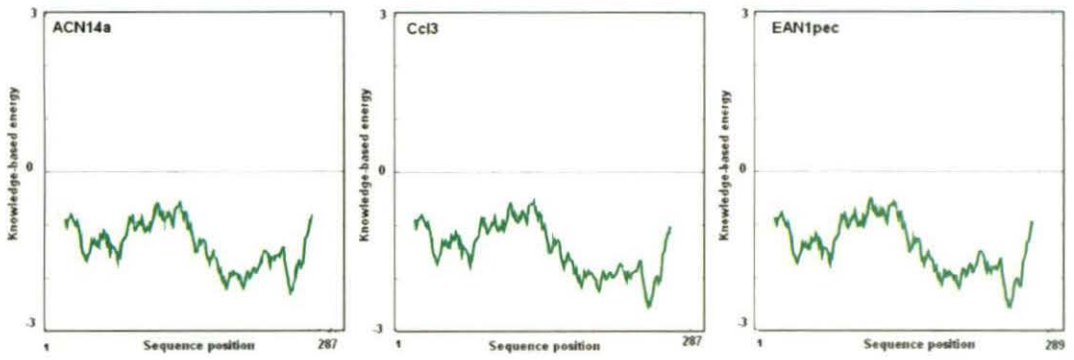

Figure 4.6.5: Energy plots of the nitrogenase iron proteins from *Frankia* strains. Residue energies averaged over a sliding window are plotted as a function of the central residue in the window

Frankia showed minor differences owing to sharing of common structural homology.

4.6.5 *In silico* site-directed mutagenesis

Bursey and Burgess (1998) reported that the alanine residue in the 157 position is important for maintaining conformation changes of *A. vinelandii* NifH protein and mutation of this residue to serine (Bursey and Burgess, 1998) caused complete inactivation of the protein. In the structural models, this alanine residue resides at 158 position and would be predicted to have a similar crucial role. Ser substitutions were observed in homologous proteins. Replacement of Cys 96 and Cys 130 residues with methionine caused influence in reaction potential and 4Fe -4S cluster activity in *A. vinelandii* (Iismaa *et al.*, 1991) that substitution of cysteine residues indeed. *In silico* site directed mutagenesis results indicated that topographical changes are not associated with ACN14a and Ccl3 however, the number of pockets changed in case of EAN1pec. Stability changes were found to be associated with the mutations. The predicted free energy change values (DDG) were found to be negative in positions 158, 96 and 130 suggesting decrease in stability of the proteins during mutations. When the same process was carried out with residues not directly associated with the core functionality or conformations no mutations were predicted. Site-directed mutagenesis results confirm that Ala 158, Cys 96 and Cys 130 are functionally important residues. These studies strongly imply that residues associated with the protein conformations and core functionality are highly conserved structurally and functionally and rendered inactive

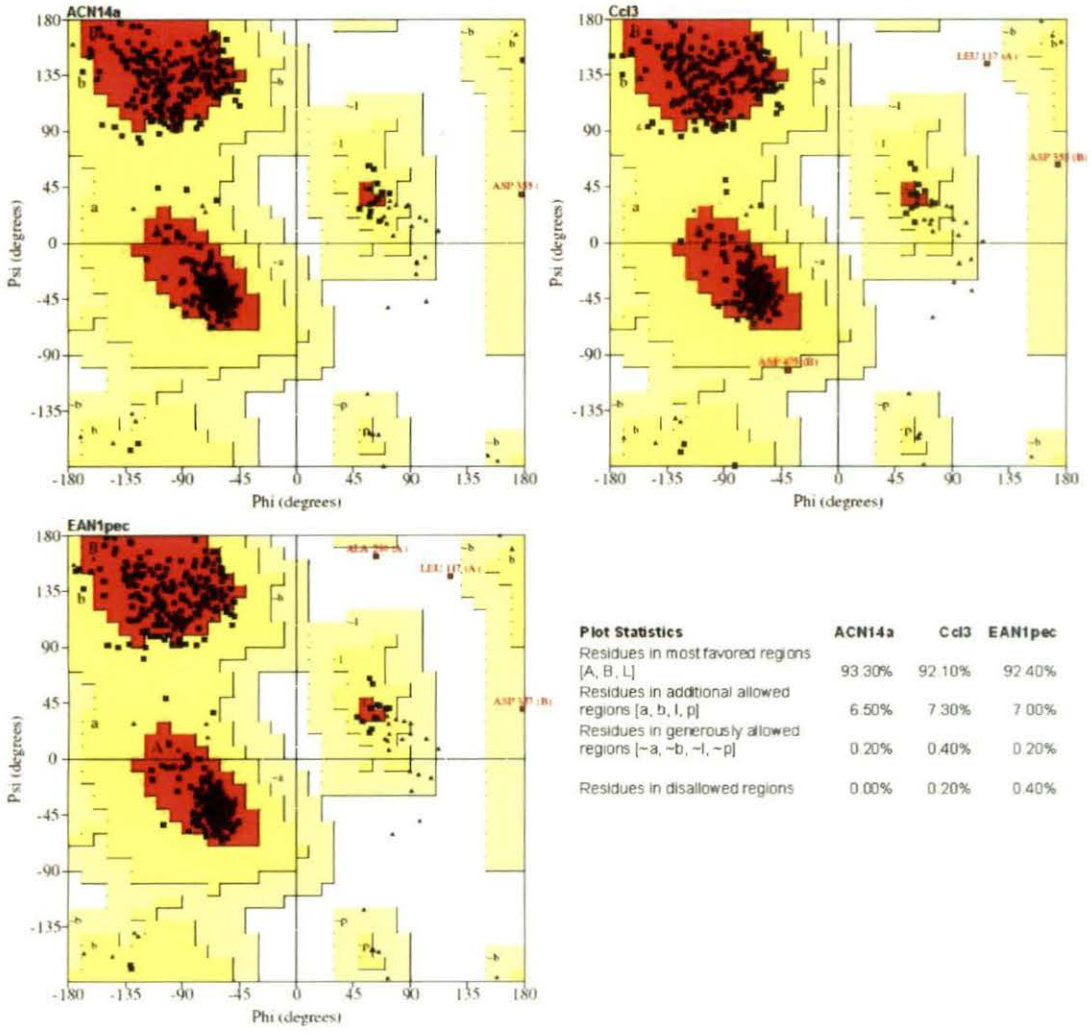

Figure 4.6.6: Ramachandran plot of the nitrogenase iron proteins from *Frankia* showing the residue backbone conformations for the modelled proteins

during mutations, as a result fixation of nitrogen is hampered.

4.6.6 Structure validation

Figure 4.6.5 shows the result for the monomers of the three modelled structures of NifH proteins from *Frankia*. The z-scores obtained from ProSA analysis specifies the overall quality of the models and gauges the extent to which the total energy of the modelled structure strays from the energy distributions of the random conformations (Wiederstein and Sippl, 2007) The NifH proteins of *Frankia* Ccl3, *Frankia* EAN1pec and *Frankia* ACN14a, had z scores of -8.40, -8.27 and -8.28 respectively. Results of z scores from ProSA analysis specify that the z-score of these models are very much within the range of scores normally found for proteins of comparable size. The energy plot (using a window size of 40 as default) of the modelled structures demonstrated

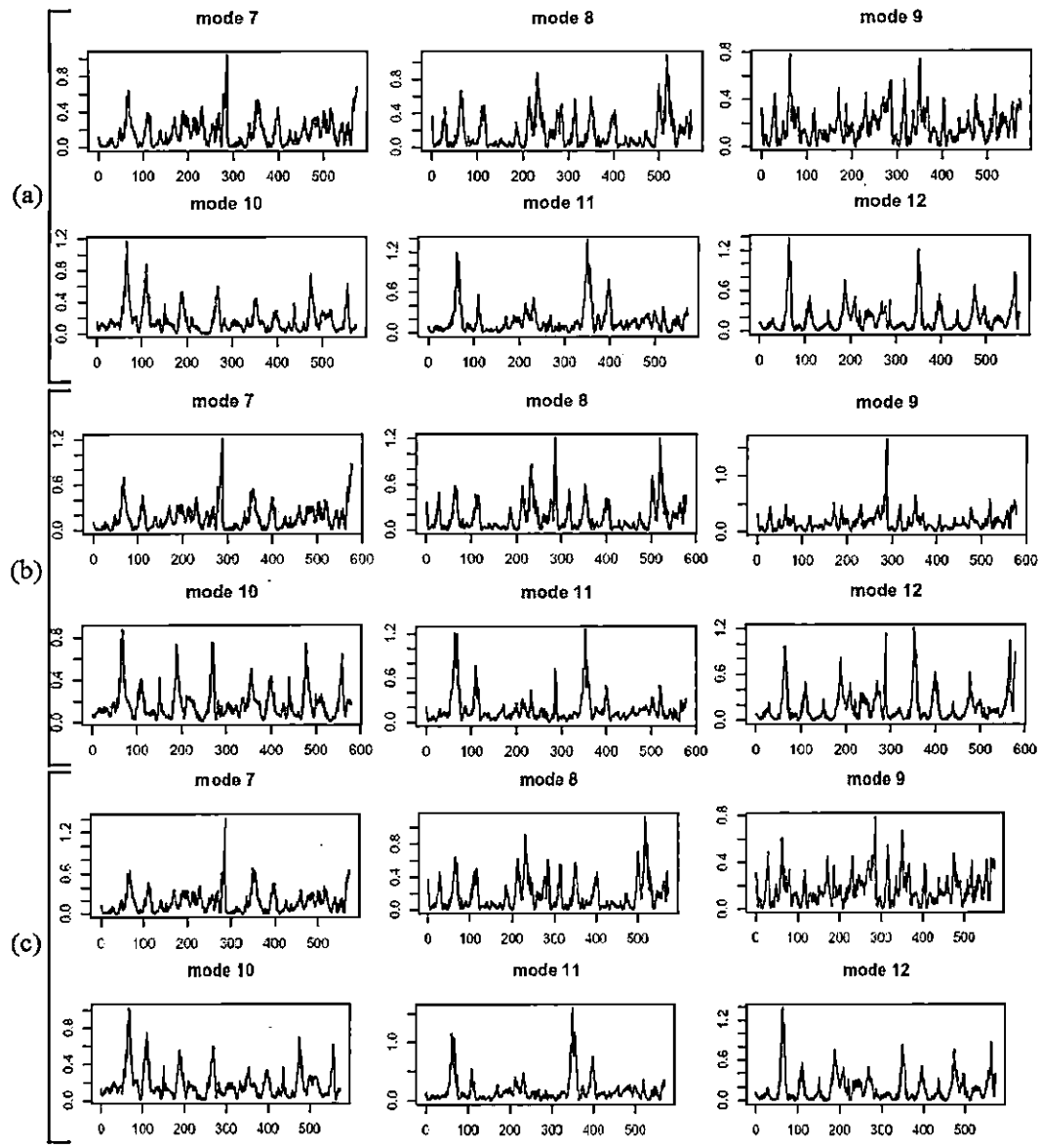

Figure 4.6.7: Normalized atomic displacement plots calculated for modes 7 to 12 in NifH proteins of a) CcI3 b) EANIpec and c) ACN14a. X and Y axis represent residue index in sequence and normal square atomic displacement respectively

and an energy distribution that is entirely below the zero base line, indicating that there are no problematic regions in the models. The value of the z-scores and outcome of the energy plots signify that the 3D models of the NifH proteins are quite reliable and precise.

The constructed models were corroborated by VERIFY 3D to estimate correctness. VERIFY 3D revealed that 91.13%, 90.33% and 90.43% of the residues had an average 3D-1D score > 0.2. Figure 4.6.6 shows the Ramachandran plot illustrating the backbone conformations for the

modelled proteins. On the basis of the analysis of 118 structures having resolution of at least 2.0 angstroms and R factor no greater than 20%, a good quality model is expected to have more than 90% in the most favored regions of the Ramachandran plot (Rajesh *et al.*, 2007). The modelled structures of the NifH protein from *Frankia* strongly satisfy this fact. Ramachandran plot of the NifH proteins from *Frankia* CcI3, EAN1pec and ACN14a revealed that the numbers of non-glycine and non-proline residues in each of the modelled proteins were 495, 498 and 495 respectively. Out of these, 456 (92.1%), 460 (92.4%) and 462 residues (93.3%) were in the most favored regions for *Frankia* CcI3, EAN1pec and ACN14a. The allotment of main chain torsion angles phi and psi evidently illustrated that bulk of the amino-acids are in a phi-psi distribution reliable with right handed alpha helices. ERRAT analysis revealed that the overall quality factor of the NifH proteins from CcI3, EANIpec and ACN14a were 95.38, 91.54 and 93.07 respectively. PROVE, VERIFY 3D and ERRAT results for all three proteins illustrated that the overall quality of the models are good. These results imply that the stereochemical properties and quality of modelled structures are quite suitable and consistent.

4.6.7 Normal mode analysis

In normal mode analysis (NMA) first six modes matching with global rotation and translation of the system are generally ignored (Hollup *et al.*, 2005) and hence lowest frequency mode of concern is the seventh one. Normal mode analysis of the NifH proteins demonstrated that low deformation energies were associated with relatively rigid regions in the protein. NMA indicated the vibrational and thermal properties of the protein at the atomic level. NifH proteins from CcI3, EANIpec and ACN14a had the lowest deformation energies of 1283.40, 1181.58 and 1295.03 in the seventh mode. It implied that the seventh mode with large rigid regions had a superior probability of describing domain motions. B factors calculated from ElNemo analysis were based on the first 100 normal modes and were scaled to match the overall B factors. Very low negative correlations for the C-alpha atoms of the NifH proteins were obtained for the computed and observed B factors. It signifies that the models contain enough rigid regions and are less flexible. Figures 4.6.7 (a-c) shows the plot of the normalized square atomic

displacements calculated for modes 7 to 12. It represents the square of the displacement of each C alpha atom, normalized as a result the sum over all residues is equal to 100. The highest values corresponded to the most displaced regions. However, the results offer little to shed light on the residues with maximum displacements associated with functional sites. Analysis of the solvent accessibility of the modelled proteins using ASAVIEW pointed out that the accessible residues were present on the outermost ring of the spiral (data not shown). Majority of negatively charged residues and polar uncharged residues were present on the outermost surface whereas hydrophobic residues were confined to the inner rings of the spiral.

It is well recognized that the NifH proteins are an indispensable constituent of the nitrogenase system of *Frankia*. The *Frankia* NifH structures are of lower resolution compared to the X-ray crystallographic ones for a number of *Azotobacter vinelandii* and *Clostridium pasteurianum* proteins. Nevertheless in absence of an X-ray crystallographic structure of NifH in *Frankia* the theoretical models projected by us provide a foundation for investigating the functional importance of nitrogenase system in this bacterium. Unlike *Frankia*, in *Azotobacter vinelandii* NifH, multiple nucleotide dependent docking sites have been reported. Studies have also revealed the presence of the all ferrous form of NifH protein (Strop *et al.*, 2001) not yet detected in *Frankia*. It has also been reported by X-ray crystallographic studies that nitrogenases in *Azotobacter vinelandii* also exist as complexes (Tezcan *et al.*, 2005) besides the uncomplexed 1NIP, 2NIP, 1G5P and Av2 structures. The RMSD deviations in the C^{α} regions between these modelled structures and the existing ones in the public domain reveal the differences in the conformations. On the basis of the results obtained from the studies in *Frankia* and those published for some diazotrophs it can be said that the Fe-protein is indeed a molecule of multiple functions and conformations. In *Frankia*, it emphasizes the awareness that nitrogenase is indeed a dynamic enzyme system. Other functions and conformations anticipate innovations as facts surface regarding the interactions of Fe-protein and its associates in *Frankia*.