Chapter 4
Results and Discussion

4.1 Codon and amino acid usage analysis:

Genes and genomes exhibit wide variations and deviant trends across the usage of synonymous codons (Sau and Deb, 2009; Guo and Yuan, 2009). The classical amino acids, except methionine (Met) and tryptophan (Trp) (encoded by unique codons), are encoded by two to six codons that are called synonymous codons. Synonymous codons are often employed in different frequencies to generate the protein sequences of a given organism and also among different organisms. These intragenomic and interspecific variations in codon usage are primarily the resultant of a combination of factors that shape the observed patterns. Considerable number of microorganisms with a balanced AT/ GC composition display significant codon heterogeneity (Sen *et al*., 2007). Interplay of several factors like genomic compositional pressure (Hou and Yang, 2003; Karlin and Mrazek, 1996), natural selection for efficient translation (Romero *et al*., 2003), gene expression level (Duret and Mouchiroud, 1999; Romero *et al*., 2003; Sharp and Li, 1986), abundance of transfer RNA (tRNA) (Duret, 2000; Ohkubo *et al*., 1987), replicational-transcriptional selection (Das *et al*., 2005; Guo and Yu, 2007; Guo and Yuan, 2009; McInerney, 1998; Romero *et al*., 2000), etc., have been reported to contribute to the observed discrepancies. The present work has been aimed at exploring the potential factors that influence the fashion of codon and amino acid usage in selected bacterial genera like *Bifidobacterium* and *Ruminococcus* and some specific strains of *Helicobacter pylori* that reside in human intestine. Proper codon and amino usage analysis promises to unravel an extensive body of information towards the designing of an accurate codon optimization methodology and construction of

Table 4.1 Overall codon usage data of *B. longum infantis* 157F-NC

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	2764	0.24	**Ser**	UCU	2462	0.36
	UUC	20504	1.76		**UCC**	16853	2.45
Leu	UUA	661	0.07		UCA	2258	0.33
	UUG	8446	0.85		**UCG**	9002	1.31
	CUU	5094	0.51		AGU	1883	0.27
	CUC	14918	1.49		**AGC**	8856	1.29
	CUA	1264	0.13	**Cys**	UGU	951	0.30
	CUG	29555	2.96		**UGC**	5292	1.70
Tyr	UAU	5731	0.61	**Pro**	CCU	2894	0.36
	UAC	13045	1.39		**CCC**	8559	1.07
His	CAU	6006	0.82		CCA	2358	0.29
	CAC	8700	1.18		**CCG**	18206	2.27
Gln	CAA	4780	0.41	**Arg**	**CGU**	9754	1.47
	CAG	18641	1.59		**CGC**	19909	3.00
Ile	AUU	8880	0.73		CGA	3203	0.48
	AUC	26317	2.15		CGG	4314	0.65
	AUA	1517	0.12		AGA	865	0.13
Asn	AAU	5964	0.51		AGG	1830	0.28
	AAC	17584	1.49	**Thr**	ACU	3885	0.37
Lys	AAA	6049	0.43		**ACC**	25432	2.39
	AAG	22387	1.57		ACA	2662	0.25
Val	GUU	4561	0.35		ACG	10543	0.99
	GUC	17393	1.32	**Ala**	GCU	8136	0.43
	GUA	3851	0.29		**GCC**	41390	2.16
	GUG	26973	2.04		GCA	8122	0.42
Asp	GAU	16568	0.73		GCG	18824	0.98
	GAC	28627	1.27	**Gly**	GGU	10610	0.77
Glu	GAA	18207	0.91		**GGC**	35412	2.58
	GAG	21702	1.09		GGA	5067	0.37
					GGG	3869	0.28

Amino acids preferentially coded by GC rich codons are marked in bold; RSCU: Cumulative Relative Synonymous Codon Usage values (RSCU) by summation of RSCU of all genes; N: Number of codons; AA: Amino acids

synthetic genes and corresponding gene products, with higher translational efficacy. Our attempt has been intended to address and explore the shifts in mutation-selection balance and unravel the enigma of codon and amino acid usage patterns in concerned bacterial members.

4.1.1 Codon and amino acid usage patterns of Bifidobacterium:

4.1.1.1 Overall codon usage patterns:

Members of the genus *Bifidobacterium* display a GC rich genomic architecture (Bottacini *et al*., 2010; Milani *et al*., 2014). The average genome size of the genus was found to be around 2.21±0.25 millions of base pairs (Mbp) and the average genomic GC composition was found to be around 60.20±1%. Genomic GC composition of *Bifidobacterium longum infantis* 157F-NC has been reported to be around 60.08%. Relative synonymous codon usage (RSCU) analysis (Table

Table 4.2 Comparative display of genomic and codon usage data for different strains of *Bifidobacterium*

Organism	Genomic Data			Number of optimal codons	Number of translationally optimal codons	Number of 'G/C' ending optimal codons	Number of 'C' ending optimal codons	RSCU (Inertia %)	
	Size (Mbp)	G+C (mol %)	tRNAs					Axis 1	Axis 2
B. longum longum KACC 91563	2.39	59.82	56	21	20	20	14	18.85	6.89
B. longum longum F8	2.38	59.62	57	21	20	20	14	17.00	6.88
B. longum longum JCM 1217	2.38	60.33	73	21	20	20	14	17.43	8.56
B. longum longum JDM301	2.48	59.81	55	21	20	20	14	19.25	7.36
B. longum infantis 157F-NC	2.41	60.08	59	21	20	20	14	17.00	7.50
B. longum infantis ATCC 15697	2.83	59.86	79	23	22	22	16	21.90	8.17
B. longum NCC2705	2.26	60.13	57	21	20	20	14	24.84	6.56
B. longum DJO10A	2.39	60.16	58	21	20	20	14	16.75	7.62
B. breve ACS-071-V-Sch8b	2.33	58.73	53	20	19	19	14	18.19	6.52
B. breve UCC2003	2.42	58.73	54	21	20	20	14	21.38	8.14
B. adolescentis ATCC 15703	2.09	59.18	54	21	20	20	14	25.11	7.80
B. dentium Bd1	2.64	58.54	55	21	20	20	14	15.07	7.19
B. asteroides PRL2011	2.17	60.05	44	22	20	22	16	19.84	5.82
B. bifidum PRL2010	2.21	62.67	52	20	19	20	15	15.27	7.40
B. bifidum S17	2.19	62.76	53	21	20	21	15	14.07	6.31
B. animalis animalis ATCC 25527	1.93	60.47	52	21	20	20	14	13.69	6.38
B. animalis lactis AD011	1.93	60.49	52	21	20	20	14	13.43	6.72
B. animalis lactis BB-12	1.94	60.48	52	23	22	22	14	14.30	5.55
B. animalis *lactis* BI-04	1.94	60.48	52	22	21	21	14	14.24	6.45
B. animalis lactis CNCM I-2494	1.94	60.49	52	24	23	23	15	14.57	5.41
B. animalis lactis DSM 10140	1.94	60.48	51	22	21	21	14	14.19	6.41
B. animalis lactis V9	1.94	60.49	52	22	21	21	14	14.20	6.43
B. animalis lactis B420	1.94	60.48	52	22	21	21	14	14.02	6.60
B. animalis lactis BLC1	1.94	60.49	52	22	21	21	14	14.38	6.82

Mbp: millions of base pairs; RSCU: Relative Synonymous Codon Usage

4.1) of *B. longum infantis* 157F-NC (taken as a representative of the bifidobacterial members) revealed that the genome of the bacterium was biased towards the usage of GC rich codons. It was well evident from Table 4.1 that thirteen amino acids were coded by GC rich codons (marked in bold in Table 4.1). Furthermore, nineteen out of twenty four codons,

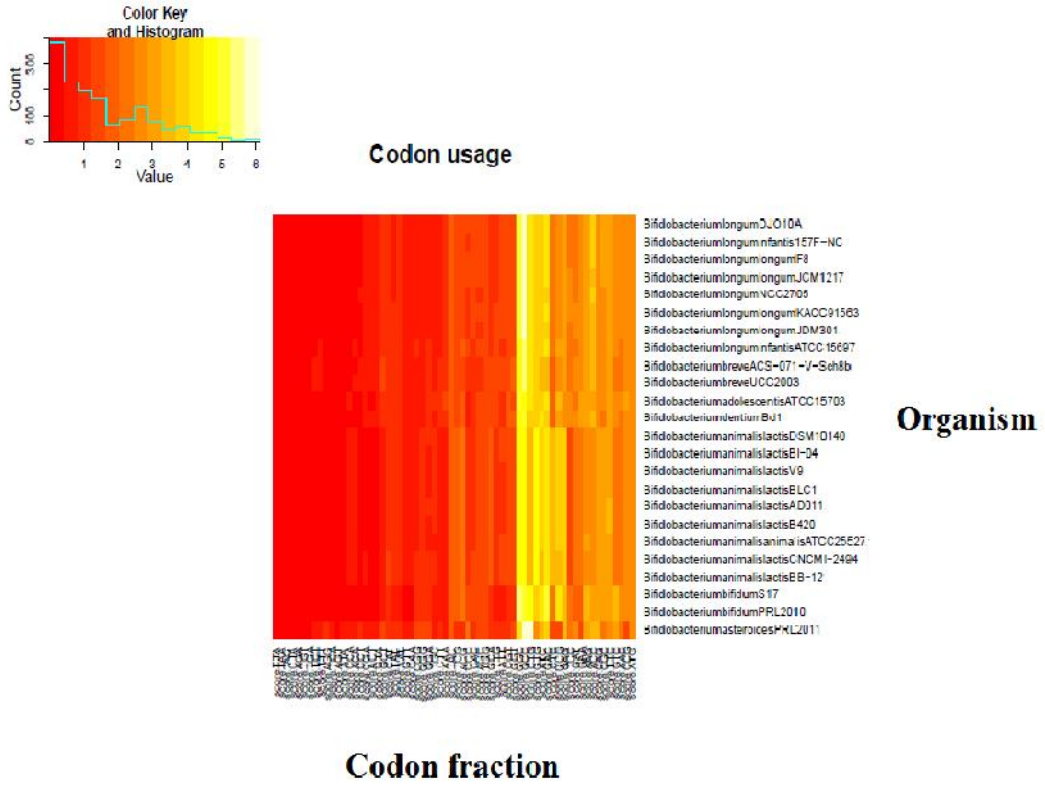

Figure 4.1 (a): Codon usage based two dimensional cluster heatmap. Heatmap based on the usage (frequency) of codons for all the bifidobacterial genomes. Color intensity changes from bright red to yellow with increase in frequency of a particular codon

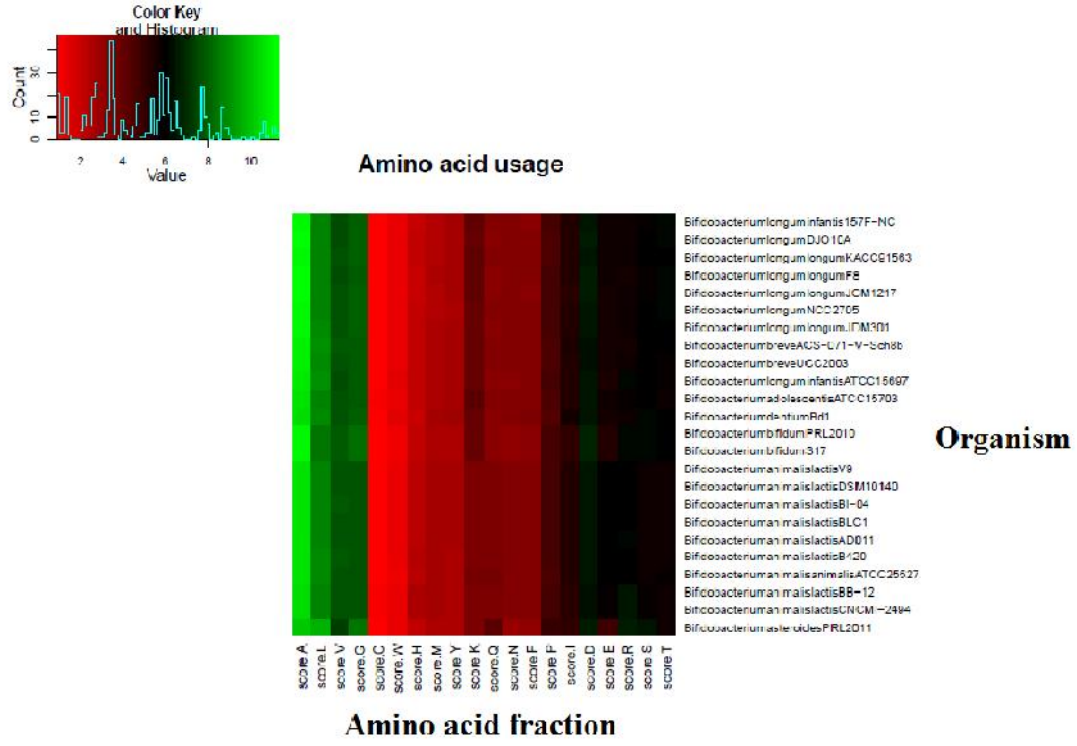

Figure 4.1 (a): Amino acid usage based two dimensional cluster heatmap**.** Heatmap based on the usage of amino acids by all the bifidobacterial genomes. Color intensity changes from bright red to green with increase in frequency of a particular amino acid

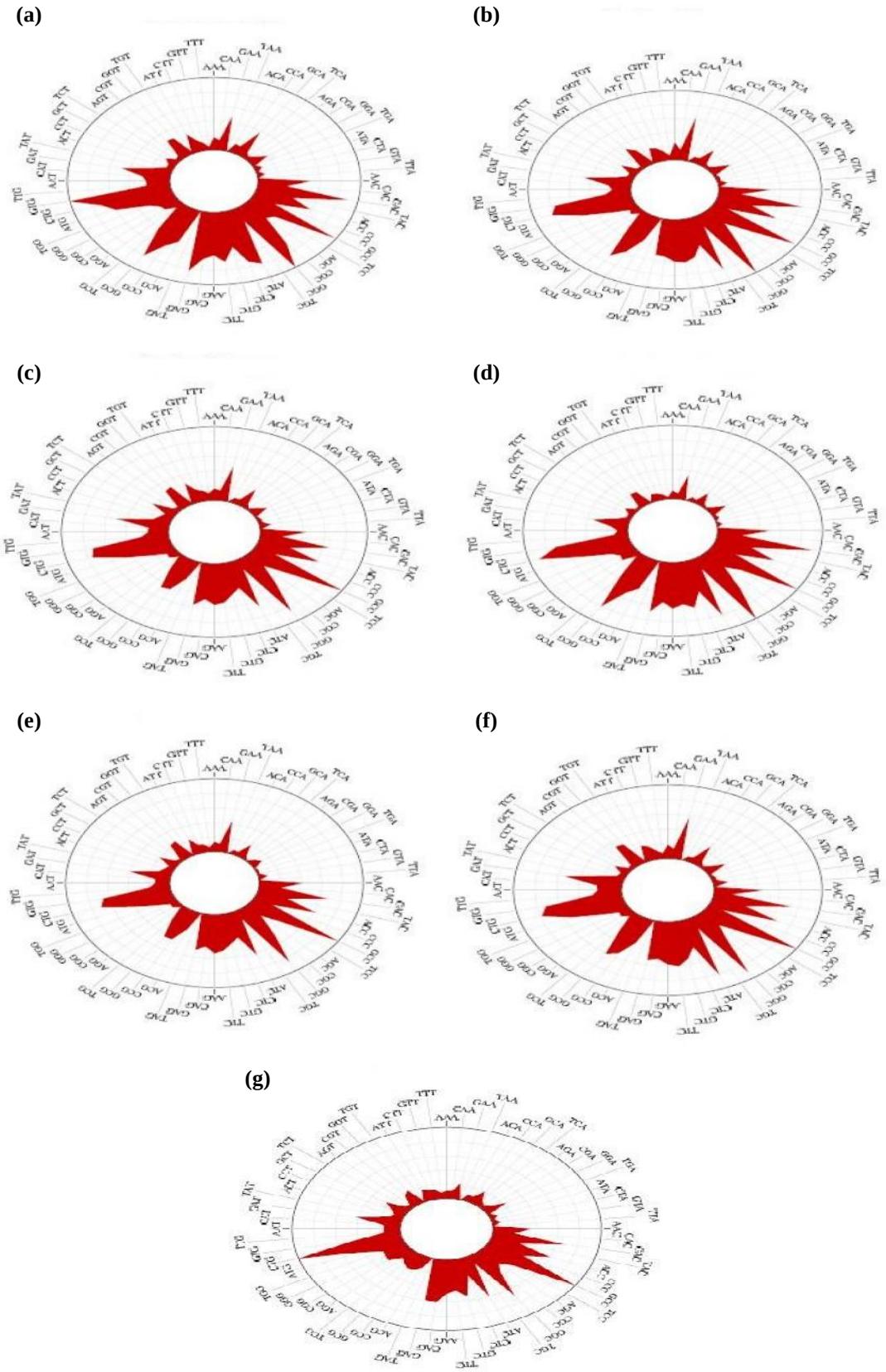

Figure 4.2: Codon usage roseplots for **(a)** *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011. Sharp red peaks represent codons used in higher frequencies

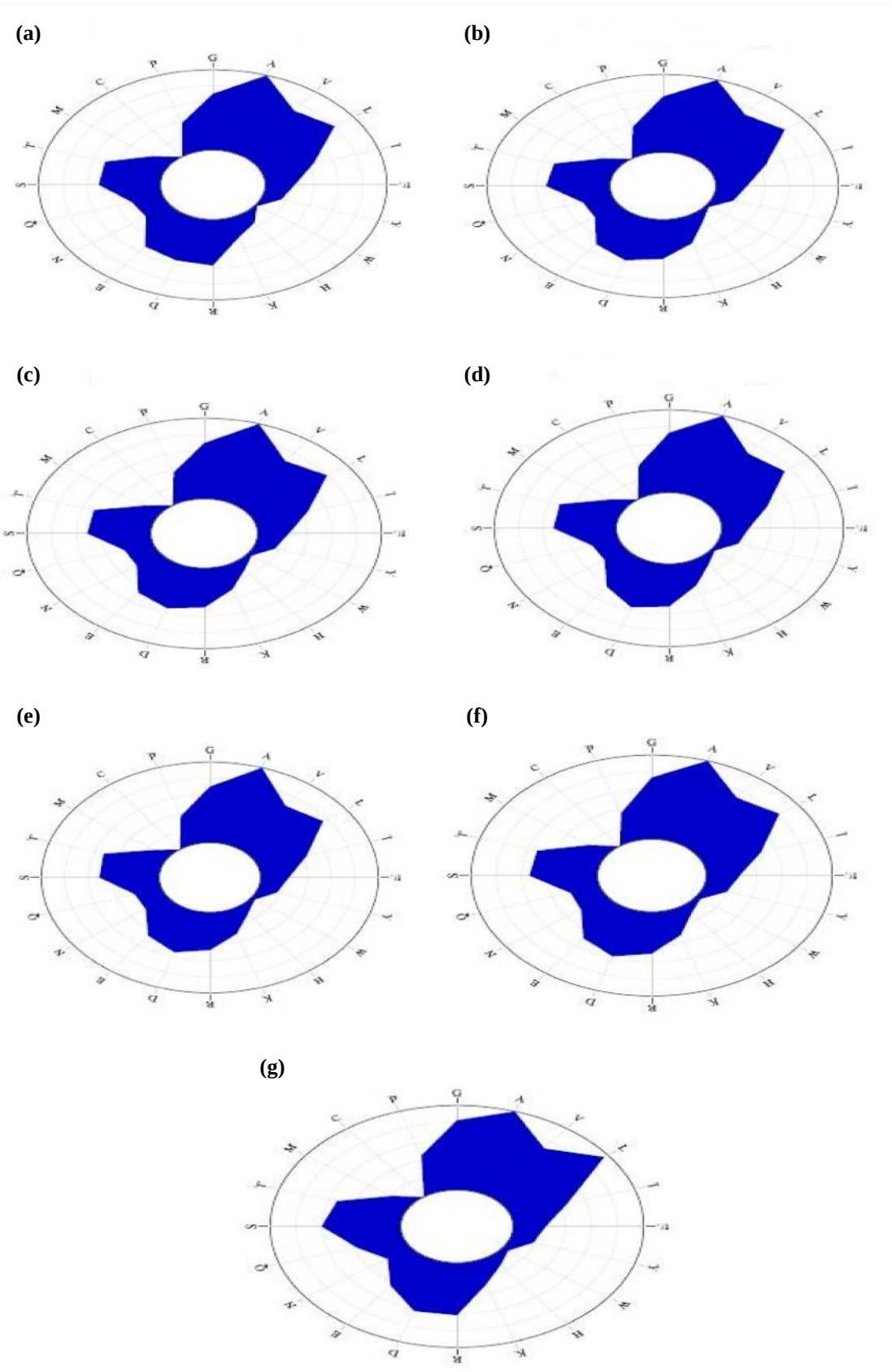

Figure 4.3: Amino acid usage roseplots for **(a**) *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011. Sharp blue peaks represent amino acids used in higher frequencies

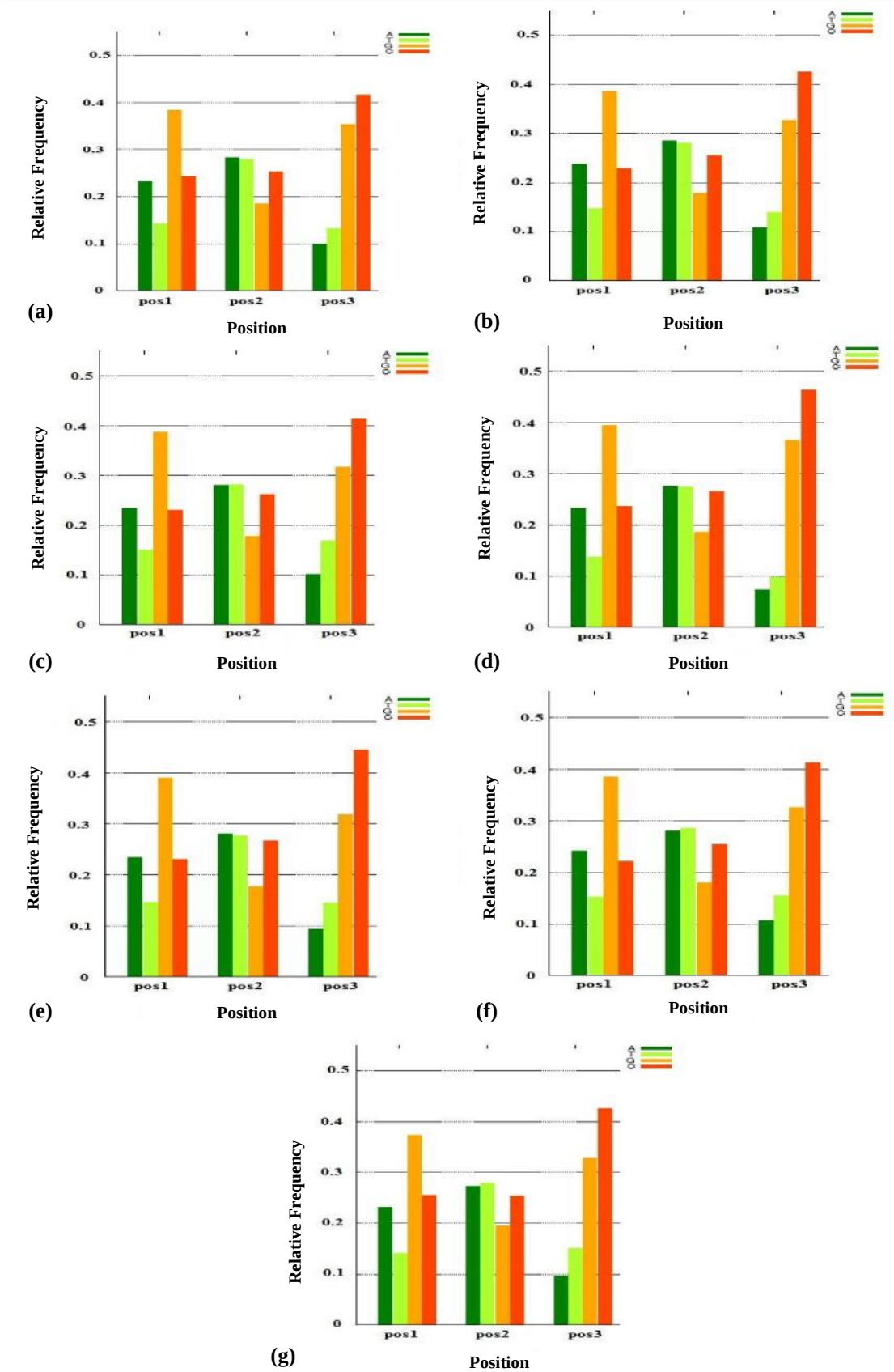

Figure 4.4: Barplots for **(a)** *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011. The color codes for the respective nucleotides have been provided in the upper panel of the figure: A- dark green; T- light green; G- yellow; C- orange

with high RSCU values (RSCU > 1), were found to be GC rich and eighteen among them were found to end with Guanine or Cytosine at the third synonymous codon position. Similar trend was evident for all other strains under study.

4.1.1.2 Factors contributing to codon usage heterogeneity:

Codon usage heatmap [Figure 4.1 (a)] confirmed the inclination of the bifidobacterial genomes towards the usage of GC rich codons like GCC, GGC, CTG, GAC, ACC, GTG and CGC as was reflected by yellow color intensity for these codons in the heatmap [Figure 4.1 (a)]. The intensity of color in the heatmap gradually changed from red to yellow with increase in frequency of a particular codon. Codon usage roseplots of *B. animalis lactis* BB-12, *B. adolescentis* ATCC 15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 (representative strains of all bifidobacterial species considered for present study) have been depicted in Figure 4.2. Roseplot analysis in each case supported our previous observation that the genomes were extremely biased towards the usage of GC rich codons.

It was also well apparent from the amino acid usage heatmap of the bifidobacterial genomes [Figure 4.1 (b)] that the amino acids like Alanine (Ala), Glycine (Gly), Leucine (Leu), Valine (Val), Aspartic acid (Asp) and Arginine (Arg) (represented by green color intensity) were employed in higher frequencies. The intensity of color gradually changed from red to green in the concerned heatmap with increase in frequency of a particular amino acid. Amino acid usage based roseplots for the genomes *B. animalis lactis* BB-12, *B. adolescentis* ATCC 15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 (Figure 4.3) and the rest of the bifidobacterial members further emphasized the fact that GC richness of the genus drove the amino acids to be coded preferentially by their GC rich codons over their AT rich counterparts. Bar plot investigations for *B. animalis lactis* BB-12, *B. adolescentis* ATCC 15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 (Figure 4.4) and the rest of the strains divulged that all of them favored using Guanine or Cytosine in the third synonymous

Table 4.3 Correlation of genomic GC composition with Synonymous Codon Usage Order (SCUO) (Spearman's Rank correlation co-efficient values) for the bifidobacterial strains

Organism	SCUO vs GC (r)
B. longum longum KACC 91563	0.39**
B. longum longum F8	0.34**
B. longum longum JCM 1217	0.49**
B. longum longum JDM301	0.45**
B. longum infantis 157F-NC	0.43**
B. longum infantis ATCC 15697	0.54**
B. longum NCC2705	0.54**
B. longum DJO10A	0.42**
B. breve ACS-071-V-Sch8b	0.47**
B. breve UCC2003	0.56**
B. adolescentis ATCC 15703	0.54**
B. dentium Bd1	0.37**
B. asteroides PRL2011	0.49**
B. bifidum PRL2010	0.18**
B. bifidum S17	0.15**
B. animalis animalis ATCC 25527	0.31**
B. animalis lactis AD011	0.32**
B. animalis lactis BB-12	0.26**
B. animalis lactis Bl-04	0.32**
B. animalis lactis CNCM I-2494	0.25**
B. animalis lactis DSM 10140	0.32**
B. animalis lactis V9	0.32**
B. animalis lactis B420	0.32**
B. animalis lactis BLC1	0.34**

(r): Spearman's Rank correlation co-efficient; ** : statistically significant at $P < 0.01$; GC: GC composition of the genomes

codon position with a distinct preference towards Cytosine at the wobble position.

Synonymous Codon Usage Order (SCUO) was estimated to compare synonymous codon usage bias with the variation of overall GC composition (GC). Correlation analysis of SCUO with GC parameter has been provided in Table 4.3. Significant positive correlations of SCUO with genomic GC content, consistently in all strains signified a notable influence of GC compositional constraint acting on the genomes.

Wright suggested that GC3 versus Nc plots can be effectively employed to investigate the factors underlying variations in codon usage patterns among genes and genomes (Wright, 1990). According to Wright (1990), if the codon usage of a genome is solely governed by GC compositional constraint, then all the genes concerned would fall on the continuous Nc plot curve. GC3 versus Nc plots were generated for all the twenty four strains. The plots for *B. animalis lactis* BB-12, *B. adolescentis* ATCC 15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 have

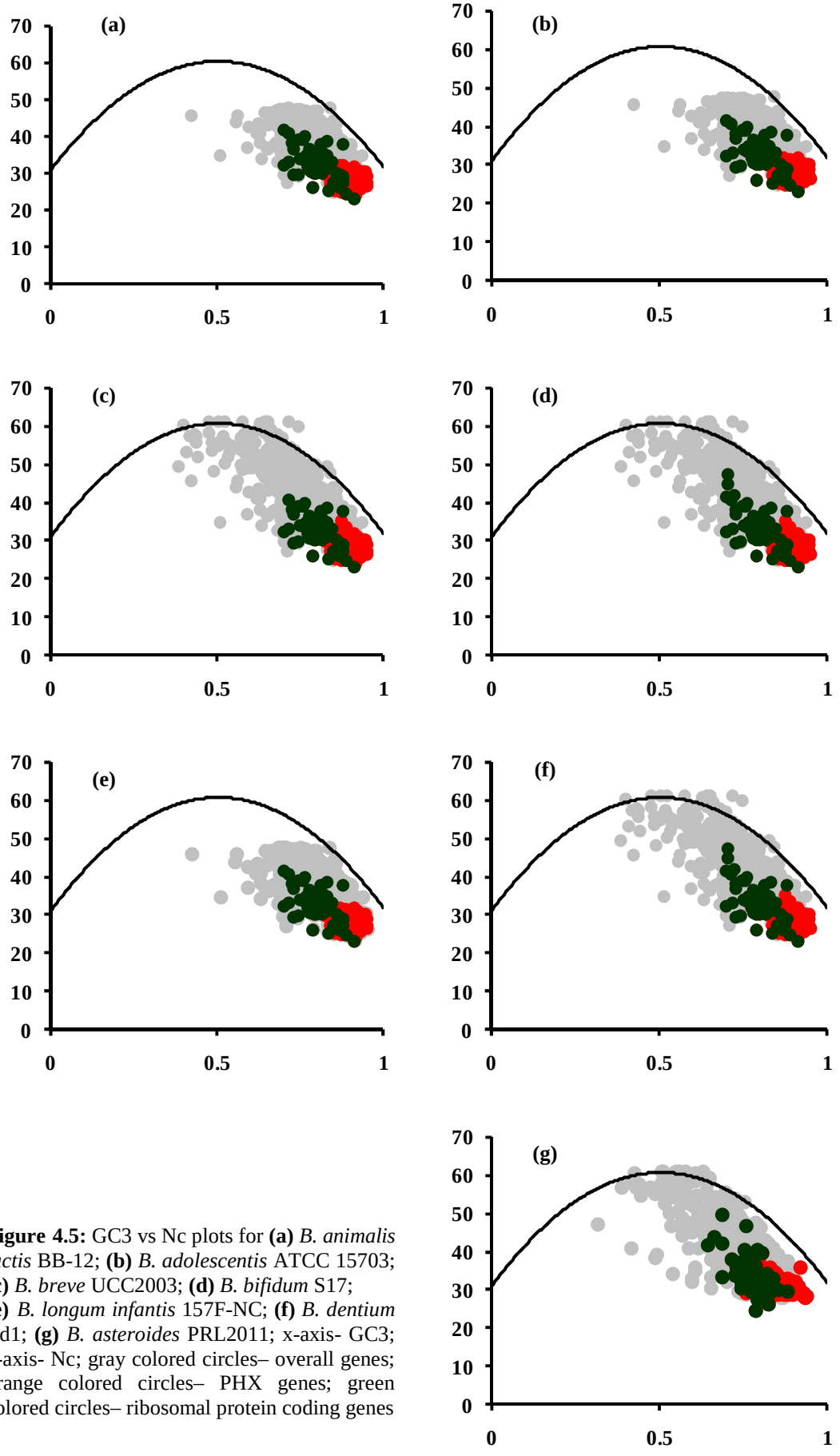

Figure 4.5: GC3 vs Nc plots for **(a)** *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011; x-axis- GC3; y-axis- Nc; gray colored circles– overall genes; orange colored circles– PHX genes; green colored circles– ribosomal protein coding genes

been shown in Figure 4.5. Interestingly, in all cases it was revealed that only few genes were falling on or above the continuous Nc plot. Majority of the genes were found to lie well below the curve signifying that apart from the influence of compositional constraint, there might be other crucial factors that govern the codon usage patterns. A clear trend was also evident from the plots that the ribosomal proteins, expected to show high levels of expression during cell division and growth, clustered at the extreme bottom of the plots. The potential highly expressed (PHX) genes, representing very low Nc values, were found to huddle even below the cluster of ribosomal proteins. It has been proposed by Comeron and Aguade (1998) that genes showing low values of Nc (<40) are under the influence of a strong codon bias, that can merely be explained as an outcome of pure compositional constraint.

RSCU data for all the strains were subjected to multivariate statistical analysis (correspondence analysis) with a motive to address the sources of such complex codon usage tendencies. Scatter plots, on the basis of RSCU values, for *B. animalis lactis* BB-12, *B. adolescentis* ATCC 15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 (Figure 4.6) distinctly reflected an unique pattern of separation of the genes. The ribosomal protein coding genes and the PHX genes, representing the highest CAI values, flocked together at one extreme of Axis 1 (x-axis), whereas, the PLX genes, reflecting the lowest CAI values, were positioned in the opposite extreme of the same axis. Trends were also similar in rest of the strains. Such tendencies of differential clustering of the PHX and PLX genes pointed towards a mighty role of translational selection to be acting on the genes with high expressivity.

A distinct trend was also noticeable on thorough analysis of the anticodon information and tRNA copy number that majority (more than ninety percent) of the optimal triplets, in all strains, were aptly recognized by the most abundant isoacceptor tRNAs and could be inferred to be translationally optimal (frequencies of optimal and translationally optimal codons have been listed in Table 4.2). tRNA Adaptation Index (tAI) values were calculated for the coding sequences in all strains with an objective to further probe the role of translational selection

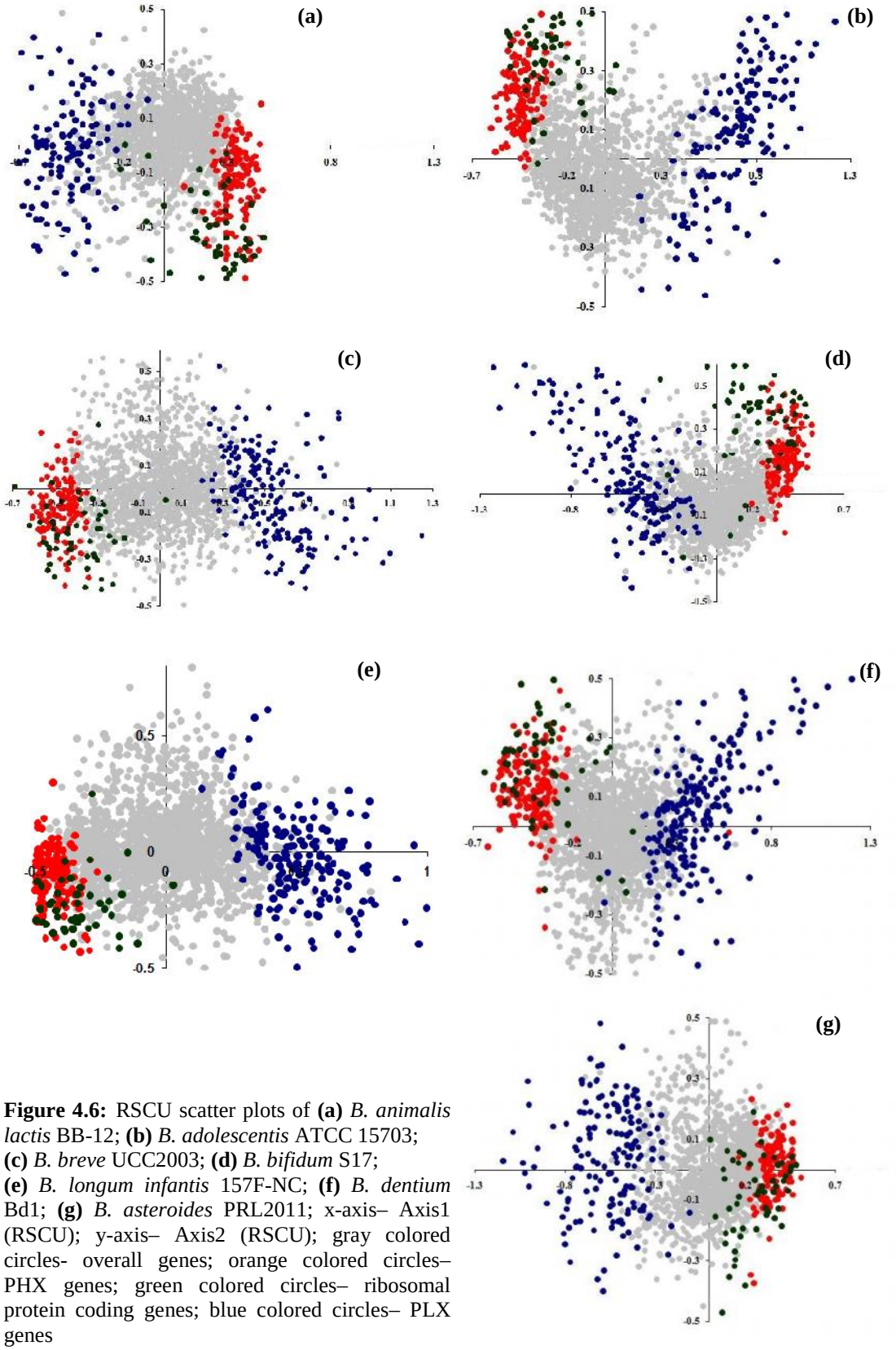

Figure 4.6: RSCU scatter plots of **(a)** *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011; x-axis– Axis1 (RSCU); y-axis– Axis2 (RSCU); gray colored circles- overall genes; orange colored circles– PHX genes; green colored circles– ribosomal protein coding genes; blue colored circles– PLX genes

Table 4.4 Bivariate correlation analysis (Spearman's Rank correlation co-efficient values) among the various codon usage indices in *Bifidobacterium*

Organism		Axis1	CAI	Fop	GC3	Nc	Length	G3	C3	GC	tAI
B. animalis lactis BB-12	**Axis1**	1.00	0.94**	0.73**	0.91**	-0.87**	0.29**	0.33**	0.82**	0.56**	0.47**
	CAI		1.00	0.70**	0.84**	-0.83**	0.27**	0.26**	0.83**	0.45**	0.50**
	Fop			1.00	0.68**	-0.68**	0.16**	0.14**	0.71**	0.37**	0.36**
	GC3				1.00	-0.82**	0.33**	0.46**	0.77**	0.62**	0.37**
	Nc					1.00	-0.16**	-0.75**	-0.31**	-0.51**	0.40**
	Length						1.00	0.22**	0.17**	0.29**	0.12**
	G3							1.00	-0.06**	0.25**	0.18**
	C3								1.00	0.38**	0.27**
	GC									1.00	0.31**
	tAI										1.00
Organism		**Axis1**	**CAI**	**Fop**	**GC3**	**Nc**	**Length**	**G3**	**C3**	**GC**	**tAI**
B. adole-scentis ATCC 15703	**Axis1**	1.00	-0.96**	-0.86**	-0.89**	0.95**	-0.25**	-0.23**	-0.88**	-0.63**	-0.34**
	CAI		1.00	0.82	0.81**	-0.91**	0.24**	0.20**	0.83**	0.54**	0.39**
	Fop			1.00	0.76**	-0.84**	0.26**	0.14**	0.79**	0.53**	0.29**
	GC3				1.00	-0.85**	0.28**	0.39**	0.83**	0.67**	0.20**
	Nc					1.00	-0.23**	-0.26**	-0.81**	-0.61**	-0.34**
	Length						1.00	0.08**	0.25**	0.24**	0.09**
	G3							1.00	-0.05**	0.22**	0.21**
	C3								1.00	0.53**	0.28**
	GC									1.00	0.23**
	tAI										1.00
Organism		**Axis1**	**CAI**	**Fop**	**GC3**	**Nc**	**Length**	**G3**	**C3**	**GC**	**tAI**
B. breve UCC2003	**Axis1**	1.00	-0.96**	-0.83**	-0.84**	0.92**	-0.15**	-0.10**	-0.88**	-0.65**	-0.45**
	CAI		1.00	0.81**	0.78**	-0.89**	0.15**	0.13**	0.81**	0.53**	0.51**
	Fop			1.00	0.72**	-0.78**	0.15**	0.03	0.78**	0.50**	0.40**
	GC3				1.00	-0.86**	0.15**	0.33**	0.84**	0.79**	0.22**
	Nc					1.00	-0.14**	-0.20**	-0.82**	-0.61**	-0.40**
	Length						1.00	0.02	0.14**	0.13**	0.06**
	G3							1.00	-0.10**	0.31**	0.14**
	C3								1.00	0.57**	0.23**
	GC									1.00	0.27**
	tAI										1.00
Organism		**Axis1**	**CAI**	**Fop**	**GC3**	**Nc**	**Length**	**G3**	**C3**	**GC**	**tAI**
B. bifidum S17	**Axis1**	1.00	0.95**	0.71**	0.84**	-0.86**	0.30**	-0.06	0.84**	0.33**	0.41**
	CAI		1.00	0.69**	0.79**	-0.84**	0.28**	-0.03	0.81**	0.26**	0.47**
	Fop			1.00	0.60**	-0.66**	0.26**	-0.11**	0.67**	0.16**	0.30**
	GC3				1.00	-0.81**	0.33**	0.20**	0.73**	0.44	0.25**
	Nc					1.00	-0.23**	-0.05*	-0.73**	-0.32**	-0.36**
	Length						1.00	0.05*	0.21**	0.26**	0.06*
	G3							1.00	-0.37**	0.13**	0.12**
	C3								1.00	0.17**	0.21**
	GC									1.00	0.21**
	tAI										1.00

Contd… to next page.

Table 4.4 *continued from previous page*

Organism		Axis1	CAI	Fop	GC3	Nc	Length	G3	C3	GC	tAI
B. longum infantis 157F-NC	**Axis1**	1.00	-0.96**	-0.80**	-0.84**	0.90**	-0.27**	-0.02	-0.87**	-0.57**	-0.13**
	CAI		1.00	0.78**	0.78**	-0.88**	0.26**	0.06*	0.81**	0.50**	0.19**
	Fop			1.00	0.70**	-0.75**	0.22**	-0.01	0.76**	0.41**	0.12**
	GC3				1.00	-0.83**	0.24**	0.26**	0.83**	0.68**	0.01
	Nc					1.00	-0.20**	-0.10**	-0.81**	-0.57**	-0.12**
	Length						1.00	0.01	0.20**	0.25**	0.04*
	G3							1.00	-0.19**	0.09**	0.18*
	C3								1.00	0.52**	0.20**
	GC									1.00	0.03
	tAI										1.00
Organism		**Axis1**	**CAI**	**Fop**	**GC3**	**Nc**	**Length**	**G3**	**C3**	**GC**	**tAI**
B. dentium Bd1	**Axis1**	1.00	-0.95**	-0.79**	-0.78**	0.88**	-0.17**	0.02	-0.82**	-0.49**	-0.48**
	CAI		1.00	0.76**	0.72**	-0.85**	0.17**	0.02	0.76**	0.44**	0.54**
	Fop			1.00	0.65**	-0.72**	0.17**	-0.09**	0.75**	0.33**	0.39**
	GC3				1.00	-0.78**	0.13**	0.24**	0.77**	0.60**	0.28**
	Nc					1.00	-0.09**	-0.10**	-0.72**	-0.50**	-0.44**
	Length						1.00	-0.02	0.11**	0.13**	0.11**
	G3							1.00	-0.29**	0.10**	0.12**
	C3								1.00	0.40**	0.24**
	GC									1.00	0.21**
	tAI										1.00
Organism		**Axis1**	**CAI**	**Fop**	**GC3**	**Nc**	**Length**	**G3**	**C3**	**GC**	**tAI**
B. asteroides PRL2011	**Axis1**	1.00	0.96**	0.68**	0.95**	-0.93**	0.12**	0.43**	0.87**	0.64**	0.40**
	CAI		1.00	0.66**	0.92**	-0.92**	0.12**	0.46**	0.85**	0.60**	0.41**
	Fop			1.00	0.66**	-0.67**	0.12**	0.09**	0.73**	0.38**	0.46**
	GC3				1.00	-0.90**	0.13**	0.49**	0.87**	0.65**	0.38**
	Nc					1.00	-0.08**	-0.38**	-0.85**	-0.62**	-0.40**
	Length						1.00	-0.03	0.12**	0.18**	0.10**
	G3							1.00	0.14**	0.37**	0.05*
	C3								1.00	0.45**	0.40**
	GC									1.00	0.23**
	tAI										1.00

**: statistically significant at $P < 0.01$; *: statistically significant at $P < 0.05$; Length: Length of coding sequences; Axis 1: Axis 1 of RSCU; Axis 2: Axis 2 of RSCU

on the bifidobacterial genomes. Significant positive correlations of tAI with CAI among all strains, evident from Table 4.4 and Supplementary Table 1 (ST1), indeed reinforced the role of natural selection for efficient translation.

Axis 1 of RSCU was found to display high positive correlation with CAI in case of *B. animalis lactis* BB-12 ($r = 0.94$, $P < 0.01$), *B. bifidum* S17 ($r = 0.95$, $P < 0.01$), *B. asteroides* PRL2011 ($r = 0.96$, $P < 0.01$), *B. animalis animalis* ATCC 25527 ($r = 0.93$, $P < 0.01$) and some other bifidobacterial genomes (Table 4.4 and ST1). Strong positive correlation indicated that the genes with high CAI values including the PHX genes and the ribosomal protein coding genes would fall on the extreme end of the positive quadrant of Axis 1 (x-axis) which in fact was reflected from the respective RSCU scatter plots. However, strong negative correlation between Axis 1 and CAI were evident in case of *B. longum infantis* 157F-NC ($r = -0.96$, $P < 0.01$), *B. longum longum* JDM301 ($r = -0.97$, $P < 0.01$), *B. breve* UCC2003 ($r = -0.96$, $P < 0.01$), *B. animalis lactis* AD011 ($r = -0.93$, $P < 0.01$) and few other bifidobacterial strains. Negative correlation indicated that the genes with high CAI values would cluster on the extreme end of the negative quadrant of Axis 1 (x-axis). Substantial correlations of CAI with Axis 1 established gene expression level to be the most convincing factor in devising the mode of codon usage in *Bifidobacterium*.

Axis 1 of RSCU was also found to correlate strongly with Fop for all the strains, as was perceptible from Table 4.4 and ST1. Such high correlations hinted towards the variation of the genes based on their preference in usage of optimal codons. It could also be concluded from the substantial positive correlations of CAI with Fop (Table 4.4 and ST1) in all the strains that the genes with high levels of expression preferentially used a particular subset of optimal codons to enhance the translational knack.

The putative highly expressed genes were mainly found to be large and small subunit ribosomal protein coding genes, ABC transporter substrate-binding protein coding genes, carbohydrate ABC transporter ATP-binding protein encoding genes, translation elongation factors, kinases, reductases, mutases, and some dehydrogenases associated significantly with the fitness of the

organisms. Some of the PHX genes were also observed to code for the cytosolic proteins. However, majority of the PLX genes were found to code for hypothetical proteins. Functional investigation and domain analysis of these hypothetical proteins revealed that majority of them were CRISPR-associated proteins. Some of the PLX genes coded for cold shock associated proteins. Strong significant positive correlations of CAI with GC content of all the concerned genomes (Table 4.4 and ST1) reflected a distinct pattern that the PHX genes also preferred to be GC rich. Separate RSCU analysis of the PHX and PLX genes also confirmed the trend of GC preference in genes with presumed higher expression levels, whereas, the PLX genes were found to employ the AT rich codons in higher frequency (data not shown).

Differential patterning of codon usage has been attributed to the interactions between overlapping genomic codes, each introducing its own bias. Methodical inspection of RSCU data and the various meaningful correlations between the codon usage parameters pointed towards a significant effect of GC compositional constraint on the codon usage signatures of *Bifidobacterium*.

Codon and amino usage based heatmaps, roseplots and barplots emphasized the tendency of the amino acids to be coded preferentially by their GC rich codon counterparts. GC richness of the genus was distinctly reflected in the codon usage variations. However, GC compositional bias alone did not account for the observed variations. A deep impact of natural selection for efficient translation was perceptible on the potential highly expressed genes and the ribosomal protein coding genes. Substantial role of translational selection was also well evident from the fact that codon usage in *Bifidobacterium* was driven by tRNA population, aimed at enhancing translational adeptness. Cases of codon usage being influenced by cellular tRNA level were reported earlier in organisms like *E. coli* (Ikemura, 1981), *C. elegans* (Duret, 2000), *D. melanogaster* (Moriyama and Powell, 1997), etc. Thus, a combined effect of compositional constraint and selection for translational efficacy were found to govern the codon usage fashion of the genus.

Apart from the definite role of compositional constraint and natural selection, there were other

determinants crucially associated with the preferential usage of codons. Gene expression level played the most vital role in shaping the codon usage patterns in *Bifidobacterium*. Codon usage profiles being dictated by gene expression level were earlier observed in various organisms like *C. elegans* (Duret and Mouchiroud, 1999), *Corynebacterium glutamicum* (Liu *et al*., 2010), *Thermobifida fusca* YX (Nayak, 2009), etc. Genes displaying high expressional levels were inferred to be GC rich and biosynthetically cost -minimized (in terms of protein energetic cost). Additionally, significant inverse correlations were observed between the GC richness and protein energetic costs in all members of *Bifidobacterium* (Table 4.8 and Supplementary table 2 [ST2]). The fact that GC rich codons code for biosynthetically cheaper amino acids was suggested by (Seligmann, 2003) and it might be a probable strategy underlying the GC richness in PHX genes of the bifidobacterial strains. The groups of microbes probably prefer to be economic while synthesizing the proteins associated with vital biological responsibilities and expressed in higher levels. GC preference of PHX genes might not only be a simple reflection of GC richness of the genus but also a manifestation of strategies to employ biosynthetically cheaper (inexpensive) amino acids. Thus, the genus was found to stringently maintain the cost-minimization strategy in the essential sets of highly expressed genes which is a characteristic trademark of a prokaryotic proteomes (Akashi and Gojobori, 2002; Heizer *et al*., 2006).

The genes with high expressivity, employing translationally optimal sets of codons, had a clear proclivity to employ Cytosine over Guanine at the wobble position. Higher correlations of CAI and Fop values with C3 over G3 parameter (Table 4.4 and ST1) indicated the above mentioned fact. Earlier, it was proposed by (Rocha, 2004) that GC content do not alter the constitution of the most prevalent anticodons and the first anticodon position of twofold-degenerated amino acids prefer usage of Guanine over Adenine and Uracil over Cytosine. Consequently, microbial genomes would prefer to employ Cytosine and Adenine at the wobble position to perfectly match the most suitable and abundant isoaccepting tRNAs. We also correlated tAI values with corresponding C3 and G3 parameters to further explore the applicability of

the hypothesis in case of bifidobacterial genomes. tAI values were found to exhibit stronger positive correlations with C3 than the G3 parameter. Thus, a bias among the translationally optimal sets of codons to use Cytosine at the wobble position provided an insight into the efficient strategy of the genus to achieve acme of translational sophistication.

Length of coding sequences was found to bear substantial significant correlations with the Axis 1 of RSCU, unanimously for all the strains involved (Table 4.4 and ST1). It could be inferred, undoubtedly, that length of coding sequences played a contributory part in influencing the codon usage variations. Furthermore, significant positive correlations of the length of coding sequences with the level of gene expression for all the strains specified a general tendency of the potential highly expressed genes to be longer in stretch.

Potential highly expressed genes in all strains of *Bifidobacterium* were found to prefer longer stretch of coding sequences. Similar instances were also reported in *E. coli* (Eyre-Walker, 1996), *Pseudomonas aeruginosa* (Gupta and Ghosh, 2001) and *Streptococcus pneumonia* (Hou and Yang, 2002) where such positive correlations of gene expression level and coding sequence length were attributed to the underlying constraints for avoiding misincorporation and missense errors during translation and maximize translational accuracy in prokaryotic genomes. However, eukaryotic organisms like *D. melanogaster* (Miyasaka, 2002) and *C. elegans* (Marais and Duret, 2001) demonstrated contrary results where length of coding sequences exhibited significant negative correlations with the level of gene expression. In this pretext, (Moriyama and Powell, 1998) suggested that longer proteins tend to be costly in terms of biosynthetic energy and shorter proteins with similar functions are preferred in such a case. In the large eukaryotic genomes, selection constraints act to reduce the length of proteins that are highly expressed as there is always space for such selection to execute. Thus, there has been no definite rule pertaining to the relationship of gene length and codon usage bias in prokaryotic and eukaryotic genomes, which are continuously exposed to varying selective constraints and this specific area remains open for further investigations.

Table 4.5 Frequencies of the favored, neutral and disfavored types of synonymous codon pairs for nine amino acids in *B. longum infantis* 157F-NC

Organism: *B. longum infantis* 157F-NC						
	Isoaccepting			Non-isoaccepting		
Grouped by	Favored	Neutral	Disfavored	Favored	Neutral	Disfavored
Alanine	4 (4 IPs)	0	4	1	1	6
Arginine	10 (6 IPs + 4 Non IPs)	0	2	2	6	16
Glycine	4 (4 IPs)	0	2	2	4	4
Isoleucine	3 (3 IPs)	0	2	0	2	2
Leucine	6 (6 IPs)	4	0	4	10	12
Proline	5 (4 Ips + 1 Non IP)	1	2	3	1	4
Serine	8 (6 IPs +2 Non IPs)	2	0	2	8	16
Threonine	4 (4 Ips)	0	2	2	0	8
Valine	4 (4 Ips)	1	1	1	1	8
Total	48	8	15	17	33	76

IP: Identical codon pair; Non Ip: Non Identical co-tRNA codon pair. Codon pairs were categorized into isoaccepting (sharing a tRNA) and non-isoaccepting (not sharing a tRNA) types. Each specific group was further classified into: Favored (≥ 3 Standard Deviation), Neutral (ranging between -3 and +3 Standard Deviation) and Disfavored (≤ -3 Standard Deviation) categories

4.1.1.3 Usage of successive synonymous codon pairs:

Our approach of exploring translational efficiency now shifted from the aspect of single codon usage to the perspective of successive synonymous codon pair utilization among the genus *Bifidobacterium*. Extensive analysis of successive synonymous codon order in all the bifidobacterial strains revealed a distinct tendency of the genus to employ identical pairs of codons (e.g. GCC-GCC codon pair for Alanine, CGC-CGC codon pair for Arginine, etc.) more than expected. Furthermore, non-identical co-tRNA codon pairs (e.g., GCC-GCT codon pair for Alanine, CGT-CGA codon pair for Arginine, etc.) were also found to occur significantly more than the expected frequencies. Such trends clearly reflected the fact that the bias in successive synonymous codon order has not only been driven by reuse of the same codon but also considerably influenced by the marked existence of non-identical co-tRNA codon pairs. However, the case was different with the non-isoaccepting consecutive codon pairs which were observed to be mostly underrepresented. Frequencies of favored, neutral and disfavored

codon pairs both for the isoaccepting and non-isoaccepting tRNA types for the nine amino acids (with at least two tRNAs), in the strain *B. longum infantis* 157F-NC (taken as representative of the bifidobacterial members), have been enlisted in Table 4.5. Trends were similar in other concerned strains also. Similar instances of such codon pairing bias i.e., favoring the usage of identical codon pairs and non-identical co-tRNA codon pairs and disallowing the practice of codon pairs for non-isoaccepting tRNAs, have already been reported in prokaryotic (Guo *et al*., 2012) and eukaryotic (Cannarozzi *et al*., 2010) datasets. Such codon pair biases have been strongly correlated with the improvement of translational precision. Thus, the genus *Bifidobacterium* has been found to judiciously employ the strategy of biased usage of successive synonymous codon pairs to significantly hone translational finesse.

4.1.1.4 Assessment of replicational-transcriptional selection pressure:

Replicational-transcriptional selection governs genomic trends in microbial genomes that witness strand asymmetry (Das *et al*., 2006; Lafay *et al*., 1999; McInerney, 1998; Guo and Yu, 2007; Guo and Yuan, 2009; Wei and Guo, 2010). Replicational selection is assumed to be accountable for the presence of a higher number of genes transcribed from the leading strand of replication than the lagging one in bacterial genomes where a distinct strand-specific asymmetry is observed (Das *et al*., 2005; Lopez and Philippe, 2001). The frequencies of genes on the leading and lagging strands were calculated to estimate the effect of replicational selection on the bifidobacterial genomes and have been listed in Table 4.6. Protein coding genes were found to be shared almost equally among the complimentary strands of replication in most of the strains. Some strains like *B. longum longum* F8, *B. breve* ACS-071-V-Sch8b and *B. asteroides* PRL2011 even displayed a higher percentage of genes to be present on the lagging strand (Table 4.6). Such results clearly indicated the absence of replication associated mutational pressure on the genomes. In order to further validate our hypothesis, RSCU analysis of the leading and lagging strand associated genes were carried out for all the strains. It was noticeable from the RSCU scatter plots of *B. animalis lactis* BB-12, *B. adolescentis* ATCC

Table 4.6 Frequencies of overall and PHX genes present in the leading and lagging strands of replication (represented in percentage) in *Bifidobacterium*

Organism	Count of Leading strand genes (%)	Count of Lagging strand genes (%)	Count of PHX genes in leading strand (%)	Count of PHX genes in lagging strand (%)
B. longum longum KACC 91563	50.23	49.77	47.72	52.28
B. longum longum F8	38.88	61.12	26.06	73.94
B. longum longum JCM 1217	47.77	52.23	50.00	50.00
B. longum longum JDM301	48.47	51.53	51.28	48.72
B. longum infantis 157F-NC	48.32	51.68	52.76	47.24
B. longum infantis ATCC 15697	53.30	46.70	42.91	57.09
B. longum NCC2705	49.80	50.20	45.66	54.34
B. longum DJO10A	49.48	50.52	47.47	52.53
B. breve ACS-071-V-Sch8b	44.36	55.64	38.67	61.33
B. breve UCC2003	50.05	49.95	48.65	51.35
B. adolescentis ATCC 15703	53.10	46.90	57.93	42.07
B. dentium Bd1	51.90	48.10	59.24	40.76
B. asteroides PRL2011	46.64	53.36	41.32	58.68
B. bifidum PRL2010	50.29	49.71	43.53	56.47
B. bifidum S17	49.92	50.08	44.07	55.93
B. animalis animalis ATCC 25527	49.41	50.59	54.25	45.75
B. animalis lactis AD011	51.57	48.43	55.26	44.74
B. animalis lactis BB-12	49.33	50.67	54.60	45.40
B. animalis lactis Bl-04	50.29	49.71	53.21	46.79
B. animalis lactis CNCM I-2494	49.46	50.54	54.55	45.45
B. animalis lactis DSM 10140	50.32	49.68	53.21	46.79
B. animalis lactis V9	50.19	49.81	53.50	46.50
B. animalis lactis B420	49.26	50.74	51.61	48.39
B. animalis lactis BLC1	49.23	50.77	52.26	47.74

15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 (representative strains of all bifidobacterial species considered for present study) (Figure 4.7) and the rest of the strains (data not shown) that the genes representing the complimentary strands of replication clustered together, i.e., there was no segregation between them. Thus, it was clear from the plots that there was no differential strand-specific codon usage pattern which re-established the absence of

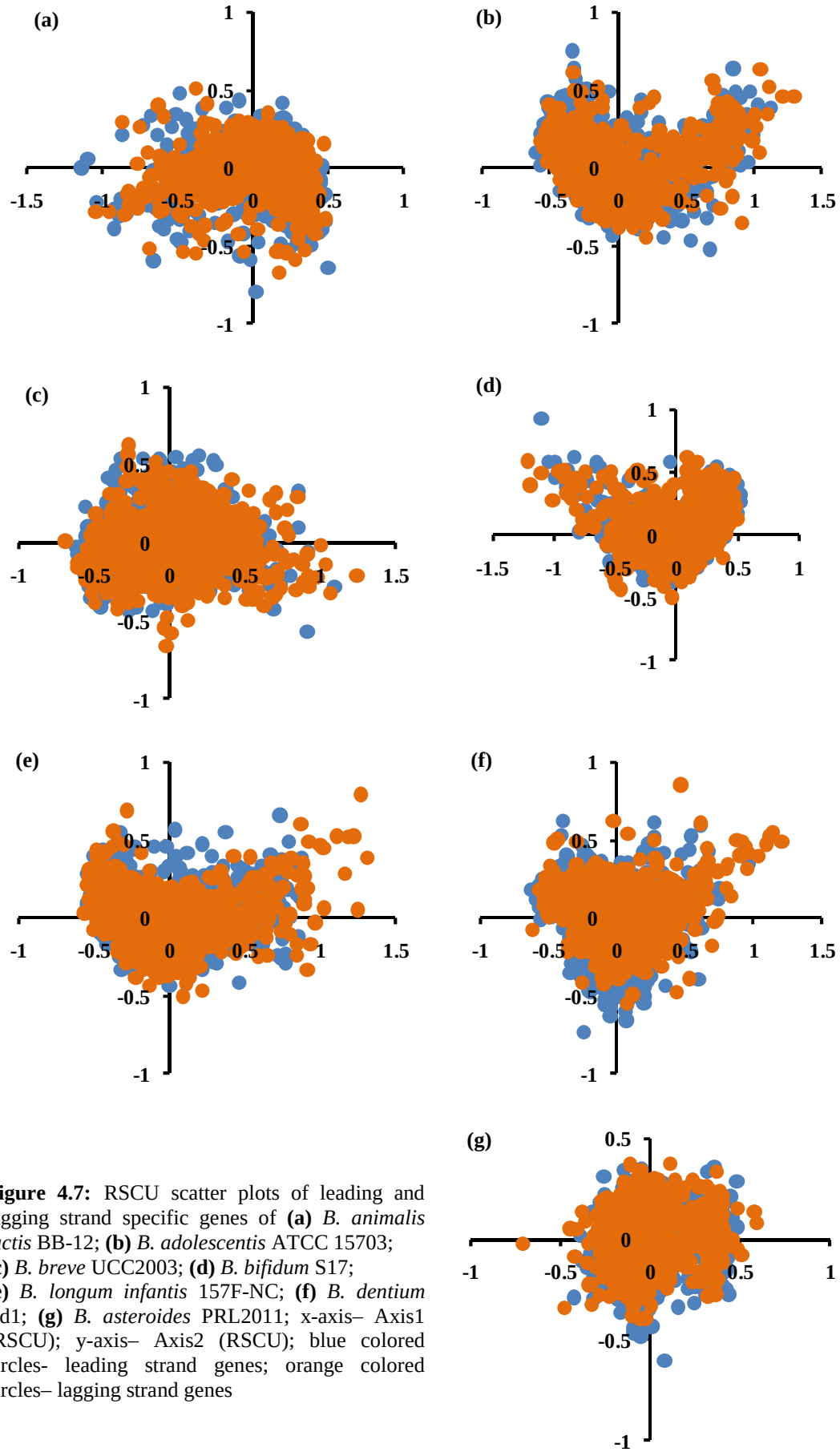

Figure 4.7: RSCU scatter plots of leading and lagging strand specific genes of **(a)** *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011; x-axis– Axis1 (RSCU); y-axis– Axis2 (RSCU); blue colored circles- leading strand genes; orange colored circles– lagging strand genes

replicational selection. Transcriptional selection is assumed to be responsible for a higher share of PHX genes and/or essential genes in the leading strand of replication in microbial genomes with characteristic strand asymmetry (Das *et al*., 2006; Rocha and Danchin, 2003). Frequencies of PHX genes present on the leading and lagging strands of replication in all the bifidobacterial members were calculated to explore the probable role of transcriptional selection pressure on them. The detailed numerical statistics has been provided in Table 4.6. A differential trend was observed among the various members of the genus. Strains representing the species *B. animalis*, *B. dentium*, *B. adolescentis* and some strains of the *B. longum* species (*B. longum infantis* 157F-NC and *B. longum* JDM301) had a higher share of PHX genes on the leading strand of replication. However, most of the strains representing species *B. longum*, *B. breve* and *B. bifidum* displayed higher frequencies of highly expressed genes to be transcribed from the lagging strand. Thus, no definite conclusion could be drawn regarding the effect of transcription associated mutational pressure on the bifidobacterial genomes. Recently, it was suggested that microbial genomes can afford to keep a higher proportion of protein coding genes and even the highly expressed genes in the lagging strand to attain proper mutation-selection balance (Chen and Zhang, 2013). The hypothesis of mutation-selection balance tended to support our observations that the lagging strands in some bifidobacterial strains were enriched with the potential highly expressed genes.

4.1.1.5 Identical codon usage patterns of Bifidobacterium and human host:

Human intestine is an ideal niche for plethora of microbes to thrive and vibe and thus, is an area of complex microbial interactions. *Bifidobacterium* is an important and numerically dominant bacterial genus residing in human intestine and has been associated with crucial beneficiary effects on the physiological “well being” of human host (Ishibashi *et al*., 1997; Lau and Liong, 2014; Wei *et al*., 2012). We executed a thorough comparison of overall codon usage patterns of various bifidobacterial members, representing human gut microflora, in contrast to human host. Codons were defined as over-represented (RSCU > 1.6) and under-represented (RSCU < 0.6) as per

Table 4.7 Comparative codon usage profile of *B. longum infantis* 157F-NC and *H. sapiens*

Codon	*H. sapiens* (RSCU)	*B. longum infantis* 157F-NC (RSCU)	Codon	*H. sapiens* (RSCU)	*B. longum infantis* 157F-NC (RSCU)
UUU	0.93	0.24	GCG	0.42	0.98
UUC	1.07	1.76	**UAU**	0.89	0.61
UUA	0.46	0.07	**UAC**	1.11	1.39
UUG	0.77	0.85	**CAU**	0.84	0.82
CUU	0.79	0.51	**CAC**	1.16	1.18
CUC	1.17	1.49	**CAA**	0.53	0.41
CUA	0.43	0.13	**CAG**	1.47	1.59
CUG	2.37	2.96	AAU	0.94	0.51
AUU	1.08	0.73	**AAC**	1.06	1.49
AUC	1.41	2.15	AAA	0.87	0.43
AUA	0.51	0.12	**AAG**	1.13	1.57
GUU	0.73	0.35	**GAU**	0.93	0.73
GUC	0.95	1.32	**GAC**	1.07	1.27
GUA	0.47	0.29	**GAA**	0.84	0.91
GUG	1.85	2.04	**GAG**	1.16	1.09
UCU	1.13	0.36	UGU	0.91	0.30
UCC	1.31	2.45	UGC	1.09	1.70
UCA	0.90	0.33	CGU	0.48	1.47
UCG	0.33	1.31	CGC	1.10	3.00
CCU	1.15	0.36	CGA	0.65	0.48
CCC	1.29	1.07	**CGG**	1.21	0.65
CCA	1.11	0.29	AGU	0.9	0.27
CCG	0.45	2.27	**AGC**	1.44	1.29
ACU	0.99	0.37	AGA	1.29	0.13
ACC	1.42	2.39	AGG	1.27	0.28
ACA	1.14	0.25	**GGU**	0.65	0.77
ACG	0.46	0.99	GGC	1.35	2.58
GCU	1.06	0.43	GGA	1.00	0.37
GCC	1.60	2.16	GGG	1.00	0.28
GCA	0.91	0.42			

RSCU: Relative Synonymous Codon Usage values; Similarly selected codons are marked in bold

scheme followed by Wong *et al* (2010). Similar codon usage pattern was concluded when a particular codon was found to display RSCU values less than 0.6 or more than 1.6 or found to fall within a range of 0.6 to 1.6 for both human host and concerned bacteria. Comparative codon usage account of *B. longum infantis* 157F-NC and pertaining human host has been provided in Table 4.7. It was observed that 27 out of out of 59 codons (excluding the single synonymous codons AUG [Met] and UGG [Trp] and the three termination codons) were found to be selected similarly in *B. longum infantis* 157F-NC and human host. The trend of selecting similar codons was persistent among other concerned bifidobacterial members and

human host. Various bacterial (Ma *et al*., 2016) and viral pathogens (Zhou *et al*., 2013; Wong *et al*., 2010) have been found to exhibit similar trends of mimicking the codon usage patterns of target host for proper infectivity and pathogenesis. We report similar findings in case of the genus *Bifidobacterium* which is a symbiotic resident of human gut and has been associated with many probiotic activities (Ishibashi *et al*., 1997; Lau and Liong, 2014). Similarities in codon usage behaviour of *Bifidobacterium* and human host might be justified from the aspect of co-evolution and co-existence (Ventura *et al*., 2007; Turroni *et al*., 2014). Identical codon usage patterns probably aid the bifidobacterial members to avert host immune response and thus, pave way for a peaceful residence in human gut.

4.1.1.6 Factors governing amino acid usage variations:

Correspondence analysis of relative amino acid usage (RAAU) data was executed for all the bifidobacterial strains to explore the possible sources of variations at the proteomic level. Correlation analysis between RAAU and several amino acid usage indices were carried out and have been provided in Table 4.8 and ST2. The first two axes (Axis 1 and Axis 2 of RAAU) explained substantial proportion of the total amino acid usage variations in all strains. Both the axes were found to exhibit strong correlations with hydropathicity index (GRAVY) [positive GRAVY (hydrophobic), negative GRAVY (hydrophilic)] of an encoded protein for every concerned strain. Furthermore, Axis 1 and Axis 2 of RAAU were also found to correlate significantly with aromaticity of the encoded proteins. RAAU scatter plots of *B. animalis lactis* BB-12, *B. adolescentis* ATCC 15703, *B. breve* UCC2003, *B. bifidum* S17, *B. longum infantis* 157F-NC, *B. dentium* Bd1 and *B. asteroides* PRL2011 (representative strains of all bifidobacterial species considered for present study) [Figure 4.8] distinctly revealed that the protein coding sequences separated into two discrete groups. One group comprised of sequences coding for protein products with high values of hydrophobicity (high GRAVY) (indicated by green colored circles in Figure 4.8) and the other group comprised of sequences that were found to code for proteins with low values of hydrophobicity (low GRAVY) (indicated by brown colored

Table 4.8 Bivariate correlation analysis (Spearman's Rank correlation co-efficient values) among the various amino acid usage parameters in *Bifidobacterium*

Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
B. animalis lactis BB-12	**Axis1**	1.00	0.02	0.39**	0.13**	-0.50**	0.38**	0.08**
	Axis2		1.00	-0.67**	-0.48**	-0.09**	-0.31**	0.19**
	GRAVY			1.00	0.17**	-0.03	0.18**	0.02
	Aromo				1.00	-0.01	0.80**	-0.29**
	CAI					1.00	-0.14**	0.45**
	PEC						1.00	0.41**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
B. adoles-centis ATCC 15703	**Axis1**	1.00	0.04	0.64**	0.38**	0.14**	0.17**	0.15**
	Axis2		1.00	-0.49**	-0.15**	0.43**	-0.38**	-0.02
	GRAVY			1.00	0.12**	-0.01	0.11**	0.05**
	Aromo				1.00	-0.01	0.80**	0.28**
	CAI					1.00	-0.13**	0.55**
	PEC						1.00	0.39**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
B. breve UCC2003	**Axis1**	1.00	-0.09	0.71**	0.27**	0.13**	0.11**	-0.14**
	Axis2		1.00	-0.49**	-0.18**	0.45**	-0.38**	-0.01
	GRAVY			1.00	0.08**	-0.08**	0.10**	-0.04
	Aromo				1.00	-0.02	0.81**	-0.24**
	CAI					1.00	-0.08**	0.54**
	PEC						1.00	-0.29**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
B. bifidum S17	**Axis1**	1.00	0.01	-0.12**	0.04	0.48**	-0.25**	0.23**
	Axis2		1.00	-0.80**	-0.13**	0.17**	-0.04	-0.05*
	GRAVY			1.00	0.03	-0.03	0.07**	0.09**
	Aromo				1.00	0.01	0.76**	-0.32**
	CAI					1.00	-0.07**	0.26**
	PEC						1.00	-0.42**
	GC							1.00

Contd… to next page.

Table 4.8 *continued from previous page*

Organism		Axis1	Axis2	GRAVY	Aromo	CAI	PEC	GC
B. longum infantis 157F-NC	**Axis1**	1.00	0.03	0.73**	0.21**	0.09**	0.02	-0.19**
	Axis2		1.00	-0.42**	-0.23**	0.37**	-0.48**	-0.07**
	GRAVY			1.00	0.06**	-0.04	0.08**	-0.05**
	Aromo				1.00	-0.03	0.78**	-0.25**
	CAI					1.00	-0.09**	0.49**
	PEC						1.00	-0.28**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
B. dentium Bd1	**Axis1**	1.00	-0.07**	0.86**	0.30**	-0.21**	0.29**	-0.13**
	Axis2		1.00	0.05**	-0.06**	-0.38**	0.23**	0.18**
	GRAVY			1.00	0.88**	-0.11**	0.15**	-0.03
	Aromo				1.00	-0.03	0.81**	-0.34**
	CAI					1.00	-0.15**	0.44**
	PEC						1.00	-0.39**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
B. asteroides PRL2011	**Axis1**	1.00	0.059*	0.82**	0.36**	-0.31*	0.38**	-0.16**
	Axis2		1.00	0.14**	-0.40**	-0.12**	-0.26**	0.50**
	GRAVY			1.00	0.10**	0.14**	0.18**	-0.55*
	Aromo				1.00	-0.07**	0.83**	-0.37**
	CAI					1.00	-0.08**	0.60**
	PEC						1.00	0.41**
	GC							1.00

**: statistically significant at $P < 0.01$; *: statistically significant at $P < 0.05$; Aromo: Aromaticity of encoded proteins; PEC: Protein energetic costs of encoded proteins; Axis 1: Axis 1 of RAAU; Axis 2: Axis 2 of RAAU

circles in Figure 4.8). The group representing the highly hydrophobic gene products included ABC transporter proteins, integral membrane proteins, major facilitator superfamily (MFS) proteins, amino acid permeases and some hypothetical proteins. Some of the hypothetical proteins displaying high GRAVY values were found to be glycosyl transferases and ABC

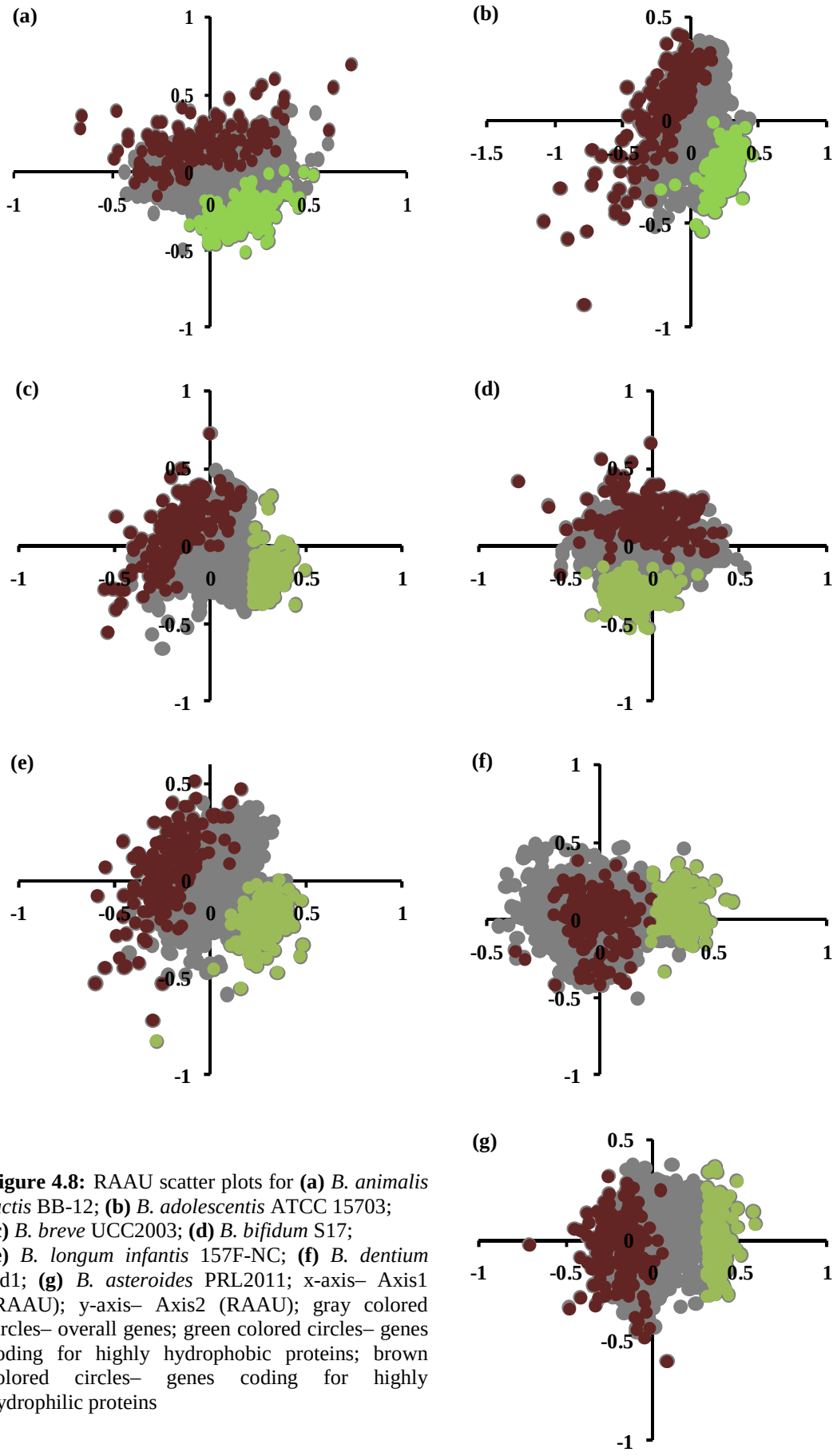

Figure 4.8: RAAU scatter plots for **(a)** *B. animalis lactis* BB-12; **(b)** *B. adolescentis* ATCC 15703; **(c)** *B. breve* UCC2003; **(d)** *B. bifidum* S17; **(e)** *B. longum infantis* 157F-NC; **(f)** *B. dentium* Bd1; **(g)** *B. asteroides* PRL2011; x-axis– Axis1 (RAAU); y-axis– Axis2 (RAAU); gray colored circles– overall genes; green colored circles– genes coding for highly hydrophobic proteins; brown colored circles– genes coding for highly hydrophilic proteins

transporters. Such a trend was consistently found in all the strains. On the contrary, the cluster of sequences with hydrophilic nature (as indicated by the low values of hydrophobicity) included large and small subunit ribosomal proteins, ATP-dependent RNA helicases and some hypothetical proteins. The hypothetical proteins representing the hydrophilic set included DNA-binding proteins and cell division associated proteins.

The major axes of variation (Axis 1 and Axis 2) of RAAU were found to correlate significantly with CAI, the estimator of gene expression level (Table 4.8 and ST2). Thus, gene expressivity not only governed the codon usage patterns, as observed earlier in this study, but also accounted considerably for the heterogeneity in amino acid usage of the genus. Protein energetic cost of the encoded proteins also contributed crucially for the differences in amino acid usage patterns, as could be inferred from the significant correlations with the two major axes of variation (Table 4.8 and ST2).

Protein energetic cost displayed high positive correlations with the aromaticity index (Table 4.8 and ST2) which was in accordance with the fact that the aromatic amino acids tend to be more expensive in terms of biosynthetic expense (Smith and Chapman, 2010). Significant negative correlations of protein energetic cost with CAI, invariably in all bifidobacterial members under scrutiny, reflected that the potential highly expressed genes favored using biosynthetically inexpensive amino acids and avoided usage of expensive ones. Such an observation of curtailing the biosynthetic cost of amino acids among the PHX gene products was further supported by the fact that their protein energetic costs were found to be significantly much reduced than the PLX gene products, throughout the genus (Supplementary Table 3 [ST3]).

Amino acid usage variations appeared to be reflections of the interplay of several crucial factors. Hydrophobic character of the encoded proteins was revealed to be the most crucial determinant accounting for such variations. Considerable influences of aromaticity of gene products and protein energetic cost were also well pronounced on the differential mode of amino acid usage. Apart from the major contributions of hydrophobicity and aromaticity, it was the expression level of the gene products that

influenced the amino acid usage behaviour of the genus considerably.

4.1.2 Codon and amino acid usage signatures of Ruminococcus:

4.1.2.1 Factors governing codon usage heterogeneity:

Members of the genus *Ruminococcus* were found to exhibit moderate AT rich genomic constitution. The average genome size of the genus was found to be around 3.27±0.76 millions of base pairs (Mbp) and the average AT composition was found to be around 56.41±4.02%. All members of the genus, except *Ruminococcus champanellensis* 18P13 (a moderate GC rich organism; GC content of around 52 %), had distinct proclivity towards the usage of AT rich codons. Overall codon usage pattern of *Ruminococcus bromii* L2-63 has been provided in Table 4.9. Twelve amino acids were found to display a bias towards the usage of AT rich codons. In addition to the above fact, it was also observed that 12 out of 23 codons, with high RSCU values (RSCU > 1), were markedly AT rich. Thus, a meticulous inspection of codon usage data signified the dominant role of AT compositional bias in shaping the codon usage patterns among the concerned members of *Ruminococcus*. However, *R. champanellensis* 18P13 was found to behave in a different fashion. Higher GC composition of *R. champanellensis* 18P13, in comparison to the other strains of *Ruminococcus*, might have accounted for such an incongruity.

Codon usage heatmap of *Ruminococcus bromii* L2-63, *Ruminococcus albus* 7, *Ruminococcus champanellensis* 18P13, *Ruminococcus* sp. 80/3, *Ruminococcus torques* L2-14, *Ruminococcus obeum* A2-162 and *Ruminococcus* sp. SR1/5 confirmed that the all the strains, except *R. champanellensis* 18P13, had a predisposition towards the usage of AT rich codons [Figure 4.9 (a)]. AT rich codons like AAA (coding for Lysine), ACA (coding for Threonine), CTT (coding for Leucine), CTG (coding for Leucine), ATA (coding for Isoleucine), ATC (coding for Isoleucine), GAA (coding for Glutamic acid), GTA (coding for Valine) and GTT (coding for Valine) were found to be employed in higher frequencies in comparison to the GC rich codon counterparts. However, in case of *R. champanellensis* 18P13, codons like AAG (coding for Lysine), ACC (coding for Threonine), GAG (coding for Glutamic acid) and GTG (coding

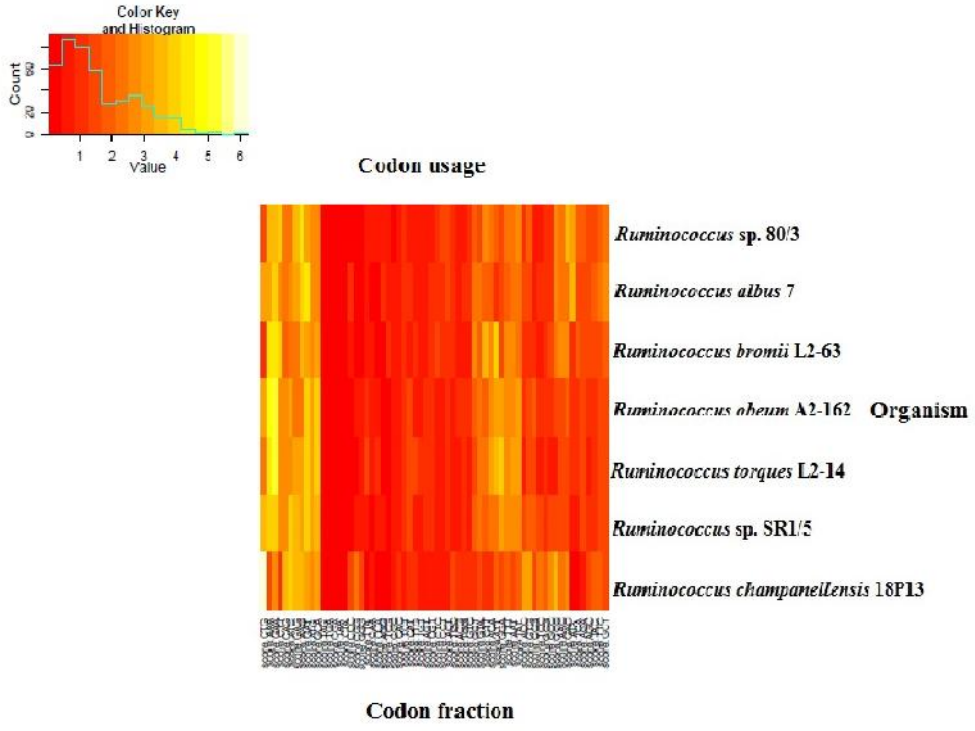

Figure 4.9 (a): Codon usage based two dimensional cluster heatmap. Heatmap based on the usage (frequency) of codons for all the genomes of *Ruminococcus.* Color intensity changes from bright red to yellow with increase in frequency of a particular codon

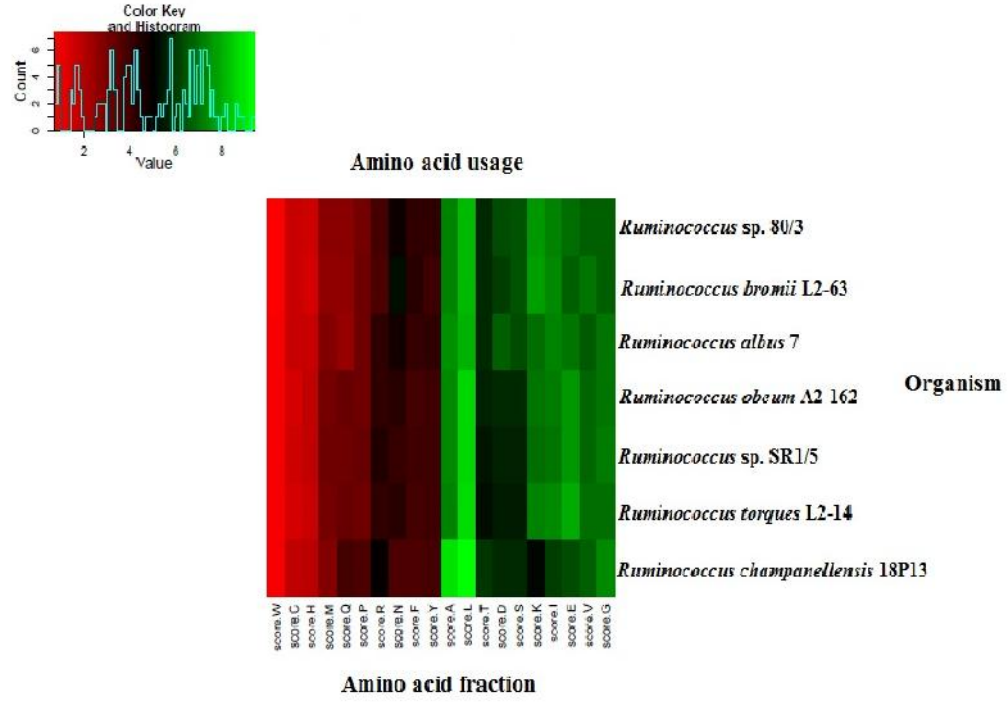

Figure 4.9 (b): Amino acid usage based two dimensional cluster heatmap**.** Heatmap based on the usage of amino acids by the genomes of *Ruminococcus*. Color intensity changes from bright red to green with increase in frequency of a particular amino acid

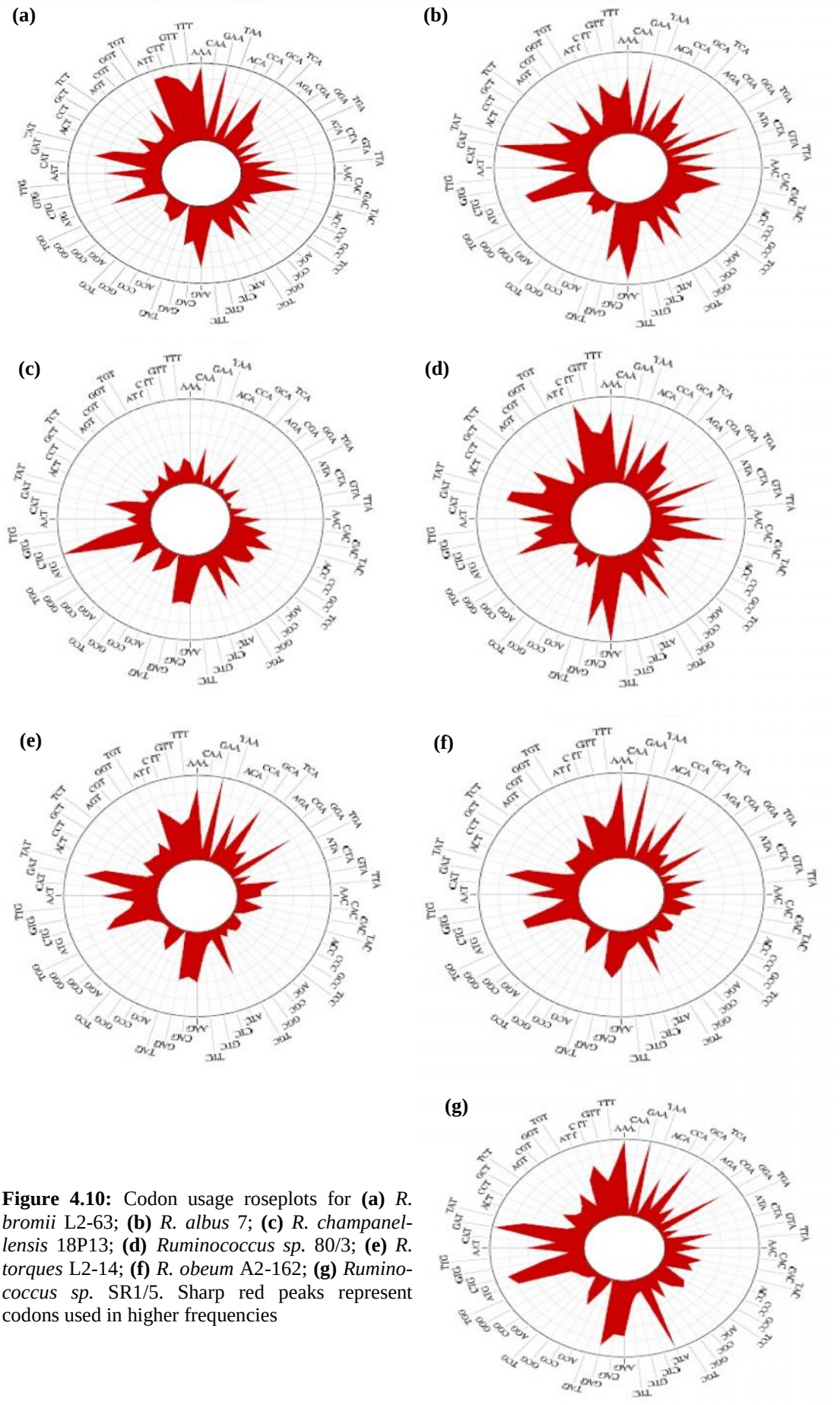

Figure 4.10: Codon usage roseplots for **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus sp.* 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus sp.* SR1/5. Sharp red peaks represent codons used in higher frequencies

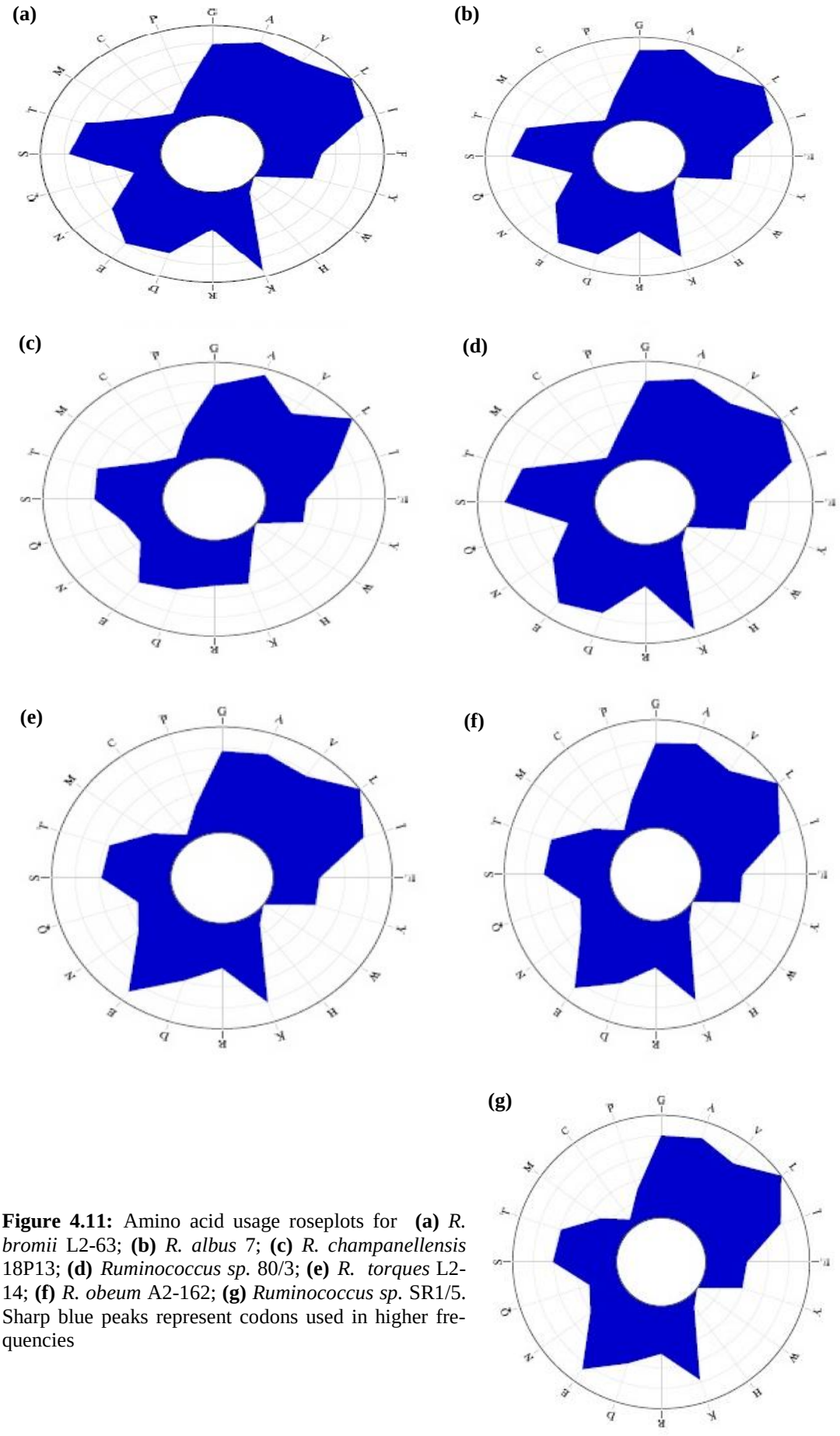

Figure 4.11: Amino acid usage roseplots for **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus sp.* 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus sp.* SR1/5. Sharp blue peaks represent codons used in higher frequencies

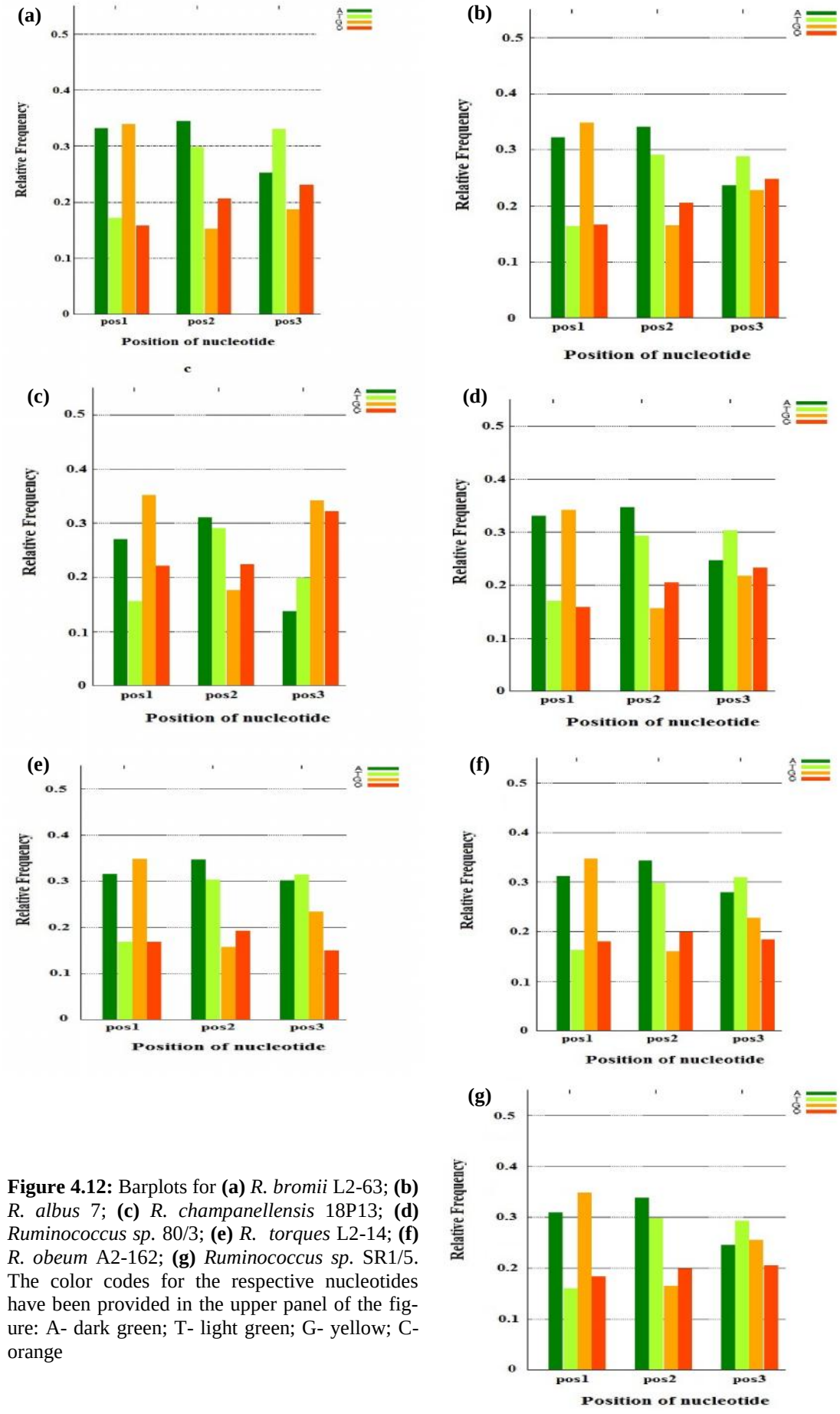

Figure 4.12: Barplots for **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus sp*. 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus sp*. SR1/5. The color codes for the respective nucleotides have been provided in the upper panel of the figure: A- dark green; T- light green; G- yellow; C- orange

Table 4.9 Overall codon usage data of *R. bromii* L2-63

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	**UUU**	15494	1.29	**Ser**	UCU	4724	0.79
	UUC	8463	0.71		UCC	3432	0.58
Leu	UUA	4023	0.53		**UCA**	10503	1.76
	UUG	5236	0.69		UCG	4972	0.83
	CUU	21844	2.88		AGU	4403	0.74
	CUC	8679	1.14		AGC	7737	1.30
	CUA	849	0.11	Cys	UGU	4428	0.95
	CUG	4901	0.65		UGC	4925	1.05
Tyr	**UAU**	13429	1.19	Pro	CCU	4733	1.08
	UAC	9171	0.81		CCC	3980	0.91
His	CAU	3889	0.94		CCA	1905	0.43
	CAC	4343	1.06		CCG	6940	1.58
Gln	CAA	3528	0.48	**Arg**	CGU	5100	1.45
	CAG	11192	1.52		CGC	2473	0.70
Ile	**AUU**	21644	1.58		CGA	1471	0.42
	AUC	12516	0.91		CGG	1048	0.30
	AUA	6999	0.51		**AGA**	8924	2.54
Asn	**AAU**	14884	1.02		AGG	2084	0.59
	AAC	14194	0.98	**Thr**	ACU	5920	0.74
Lys	**AAA**	23247	1.09		ACC	6235	0.78
	AAG	19383	0.91		**ACA**	15640	1.94
Val	**GUU**	19884	2.05		ACG	4382	0.54
	GUC	4275	0.44	Ala	GCU	9615	0.96
	GUA	9516	0.98		GCC	7235	0.72
	GUG	5055	0.52		GCA	18080	1.80
Asp	**GAU**	17964	1.05		GCG	5211	0.52
	GAC	16143	0.95	Gly	GGU	13863	1.51
Glu	**GAA**	24414	1.31		GGC	13425	1.46
	GAG	12728	0.69		GGA	8234	0.89
					GGG	1282	0.14

Amino acids preferentially coded by AU rich codons are marked in bold; RSCU: Cumulative Relative Synonymous Codon Usage values (RSCU) by summation of RSCU of all genes; N: Number of codons; AA: Amino acids

for Valine) were used more frequently than the AT rich codon counterparts. Such an observation appeared a simple reflection of moderate GC richness of the strain. Similar trends were also revealed from the respective codon

Table 4.10 Comparative display of genomic and codon usage data for different strains of *Ruminococcus*

Organism	Genomic Data			Number of opti-mal codons	Number of 'A/ U' rich optimal codons	RSCU	
	Size (Mbp)	A+T (mol %)	tRNAs			Axis 1 (Inertia %)	Axis 2
R. bromii L2-63	2.24	58.95	36	23	15	12.82	9.06
R. albus 7	4.49	56.38	74	25	08	12.22	8.19
R. champanellensis 18P13	2.57	47.79	44	23	05	16.58	7.33
Ruminococcus sp. 80/3	2.97	57.38	55	24	13	12.19	7.78
R. torques L2-14	3.34	59.86	47	21	12	9.88	9.03
R. obeum A2-162	3.76	58.25	46	23	05	11.33	7.52
Ruminococcus sp. SR1/5	3.54	56.24	47	23	12	10.64	6.62

Mbp: millions of base pairs; RSCU: Relative Synonymous Codon Usage

usage based roseplots (Figure 4.10).

It was visible from the amino acid usage based heatmap of the ruminococcal strains that amino acids like Lysine (Lys), Threonine (Thr), Leucine (Leu), Isoleucine (Ile), Glutamic acid (Glu), Valine (Val), Alanine (Ala) and Glycine (Gly) were used in relatively higher proportions than the rest of the amino acids in all the strains [Figure 4.9 (a)]. Amino acid usage based roseplots of the respective strains emphasized the same fact (Figure 4.11).

Barplots (Figure 4.12) revealed that all the ruminococcal strains, except *R. champanellensis* 18P13, were AT rich at the third synonymous codon position. Furthermore, it was also evident that Thymine was preferred over Adenine at the wobble position (third synonymous codon position). *R. champanellensis* 18P13 was found to prefer Guanine or Cytosine at the third synonymous codon position which appeared a consequence of its moderate GC rich genomic architecture. Comparative codon usage features of the concerned members of *Ruminococcus* have been provided in Table 4.10.

Synonymous codon usage order (SCUO) has been defined as an estimate depicting the quantitative relationship between synonymous codon usage bias and genomic composition of an organism (Wan *et al.*, 2004). SCUO was calculated for all the strains of *Ruminococcus*. SCUO data was found to correlate well with AT composition of the concerned strains as was indicated from the strong positive correlations (Table 4.11). However, moderate GC richness of *R. champanellensis* 18P13 resulted in negative correlation between SCUO

Table 4.11 Correlation of genomic AT composition with Synonymous Codon Usage Order (SCUO) (Spearman's Rank correlation co-efficient values) for the ruminococcal strains

Organism	SCUO vs AT (r)
R. bromii L2-63	0.11**
R. albus 7	0.04
R. champanellensis 18P13	-0.24**
Ruminococcus sp. 80/3	0.03
R. torques L2-14	0.23**
R. obeum A2-162	0.32**
Ruminococcus sp. SR1/5	0.30**

(r): Spearman's Rank correlation co-efficient; ** : statistically significant at $P < 0.01$; AT: AT composition

and genomic AT composition. Thus, it was prominent from strong correlation between SCUO data and genomic composition of the organisms that codon usage of the genus was well dominated by genomic constitution.

Correspondence analysis of relative synonymous codon usage (RSCU) data for all the strains of *Ruminococcus* revealed that the Axes 1 and 2 contributed to the maximum codon usage variations observed in the 59-dimensional hyperspace (indicated by high values of inertia). Axis 1 was found to be the principle axis, invariably in all strains, followed by Axis 2 (Table 4.12). Correlation analysis of RSCU data and various codon usage parameters was performed to explore the intricacies of codon usage bias in the genus. Genomic composition of the genus (AT/GC composition) was found to exhibit significant high correlations with both the major axes of separation of genes on the basis of RSCU data i.e., Axis 1 and Axis 2 of RSCU (Table 4.12).

Thus, pronounced impact of compositional bias could be well inferred to be acting in the genus *Ruminococcus* from exhaustive correlation analyses, and thorough inspection of codon usage based heatmaps, roseplots and barplots.

GC3 versus Nc (effective number of codons) plots have been suggested to be instrumental in studying codon usage variations among genes and genomes (Wright, 1990). According to (Wright, 1990), if codon usage of a genome is solely governed by compositional constraint, then all the genes concerned would fall on the continuous Nc plot curve. We analyzed GC3 versus Nc plots for all the concerned strains of *Ruminococcus*

Table 4.12 Bivariate correlation analysis (Spearman's Rank correlation co-efficient values) among the various codon usage indices in *Ruminococcus*

Organism		Axis1	Axis2	GC3	Nc	Length	GC	CAI	tAI
Rumino-coccus bromii L2-63	**Axis1**	1.00	0.11**	-0.09**	-0.65**	0.22**	0.16**	0.93**	0.48**
	Axis2		1.00	0.81**	-0.04	0.12**	0.75**	0.14**	0.46**
	GC3			1.00	0.17**	0.06**	0.76**	-0.06*	0.44**
	Nc				1.00	-0.06*	-0.07**	-0.64**	-0.35**
	Length					1.00	0.15**	0.19**	0.20**
	GC						1.00	0.15**	0.53*
	CAI							1.00	0.47**
	tAI								1.00
Organism		**Axis1**	**Axis2**	**GC3**	**Nc**	**Length**	**GC**	**CAI**	**tAI**
Rumino-coccus albus 7	**Axis1**	1.00	0.01	-0.72**	0.46**	-0.23**	-0.70**	-0.82**	0.03*
	Axis2		1.00	0.45**	0.35**	-0.08**	0.37**	-0.32**	-0.46**
	GC3			1.00	-0.16**	0.09**	0.81**	0.40**	-0.19**
	Nc				1.00	-0.07	-0.18**	-0.50**	-0.11**
	Length					1.00	0.15**	0.31**	-0.01
	GC						1.00	0.42**	-0.20**
	CAI							1.00	0.19**
	tAI								1.00
Organism		**Axis1**	**Axis2**	**GC3**	**Nc**	**Length**	**GC**	**CAI**	**tAI**
Ruminoc occus champan-ellensis 18P13	**Axis1**	1.00	0.01	-0.92**	0.73**	-0.31**	-0.73**	-0.57**	0.44**
	Axis2		1.00	-0.06**	-0.18**	0.10**	-0.22**	0.61**	0.18**
	GC3			1.00	-0.67**	0.26**	0.79**	0.49**	0.44**
	Nc				1.00	-0.14**	-0.50*	-0.65**	-0.35**
	Length					1.00	0.23**	0.24**	0.13*
	GC						1.00	0.26**	0.30**
	CAI							1.00	0.44**
	tAI								1.00
Organism		**Axis1**	**Axis2**	**GC3**	**Nc**	**Length**	**GC**	**CAI**	**tAI**
Rumino-coccus sp. 80/3	**Axis1**	1.00	0.07**	0.10**	-0.66**	0.21**	0.30**	0.90**	0.16**
	Axis2		1.00	-0.78**	-0.18**	-0.06**	-0.58**	0.16**	0.31**
	GC3			1.00	0.06**	0.07**	0.72**	0.02	-0.14**
	Nc				1.00	-0.06**	-0.15**	-0.65**	-0.14**
	Length					1.00	0.13**	0.22**	-0.09**
	GC						1.00	0.19**	-0.07**
	CAI							1.00	0.20**
	tAI								1.00

Contd… to next page.

Table 4.12 *continued from previous page*

Organism		Axis1	Axis2	GC3	Nc	Length	GC	CAI	tAI
Ruminococcus torques L2-14	**Axis1**	1.00	-0.18**	0.11**	0.63**	-0.20**	-0.18**	-0.89**	-0.17**
	Axis2		1.00	0.60**	0.10**	0.11**	0.70**	0.22**	0.39**
	GC3			1.00	0.28**	0.05**	0.71**	-0.14**	0.29**
	Nc				1.00	-0.01	0.03	-0.56**	-0.08**
	Length					1.00	0.18**	0.18**	0.02
	GC						1.00	0.13**	0.21**
	CAI							1.00	0.30**
	tAI								1.00

Organism		Axis1	Axis2	GC3	Nc	Length	GC	CAI	tAI
Ruminococcus obeum A2-162	**Axis1**	1.00	0.06**	-0.50**	0.23**	-0.30**	-0.63**	-0.62**	-0.61**
	Axis2		1.00	0.63**	0.54**	-0.05**	0.39**	-0.61**	-0.20**
	GC3			1.00	0.27**	0.11**	0.76**	-0.13**	0.25**
	Nc				1.00	0.01	0.07**	-0.49**	-0.29**
	Length					1.00	0.23**	0.25**	0.14**
	GC						1.00	0.09**	0.36**
	CAI							1.00	0.60**
	tAI								1.00

Organism		Axis1	Axis2	GC3	Nc	Length	GC	CAI	tAI
Ruminococcus sp. SR1/5	**Axis1**	1.00	0.08**	-0.61**	0.24**	-0.37**	-0.66**	-0.64**	-0.50**
	Axis2		1.00	0.52**	0.32**	-0.03*	0.34**	-0.50**	-0.20**
	GC3			1.00	0.04*	0.19**	0.79**	0.08**	0.21**
	Nc				1.00	0.01	-0.06**	-0.40**	-0.28**
	Length					1.00	0.29**	0.31**	0.14**
	GC						1.00	0.20**	0.30**
	CAI							1.00	0.58**
	tAI								1.00

**: statistically significant at $P < 0.01$; *: statistically significant at $P < 0.05$; Length: Length of coding sequences; Axis1: Axis1 of RSCU; Axis2: Axis2: Axis2 of RSCU

(Figure 4.13) and came to the conclusion that some genes were indeed falling on or above the continuous Nc plot curve. Such an observation signified the dominant signature of compositional constraint to be effective on the concerned members of *Ruminococcus*. However, presence of significant share of genes below the curve implied the presence other factors, apart from compositional constraint. It was distinct from the GC3 versus Nc plots that the ribosomal proteins and potential highly expressed (PHX) genes clustered together. Comeron and Aguade (1998) suggested that low values of Nc (Nc < 40) in case of highly expressed essential sets of genes indicate the presence of factors, other than compositional constraint, to be operating on the genomes. In our case majority of the potential highly expressed genes and ribosomal protein coding genes represented Nc values < 40 which indicated that the codon usage patterns in *Ruminococcus* were not only subject to compositional pressure but also under the influence of other imperative factors.

In order to further elucidate the factors underlying complex codon usage behaviour of the genus *Ruminococcus*, we executed multivariate statistical analysis of relative synonymous codon usage (RSCU) data of all the ruminococcal strains. RSCU scatter plots for the respective strains were generated and have been provided in Figure 4.14. A characteristic pattern was noted that the potential highly expressed (PHX) genes and the ribosomal protein coding genes, representing higher levels of gene expression, were found to huddle at one extreme of the major axis of separation of genes on the basis of RSCU data (either Axis 1 or Axis 2) whereas, the potential lowly expressed (PLX) genes were found to lie in the opposite extreme of the same axis. Such an observation of differential codon usage patterns of PHX and PLX genes hinted towards a probable role of translational selection to be acting on the potential highly expressed genes and the ribosomal proteins of the genus.

Positive correlations of gene expressivity, as indicated by codon adaptation index (CAI), and tRNA adaptation index (tAI), invariably in all strains of *Ruminococcus* (Table 4.12), further re-established the role of translational selection in producing observed codon usage variations.

Systematic scrutiny of RSCU data and

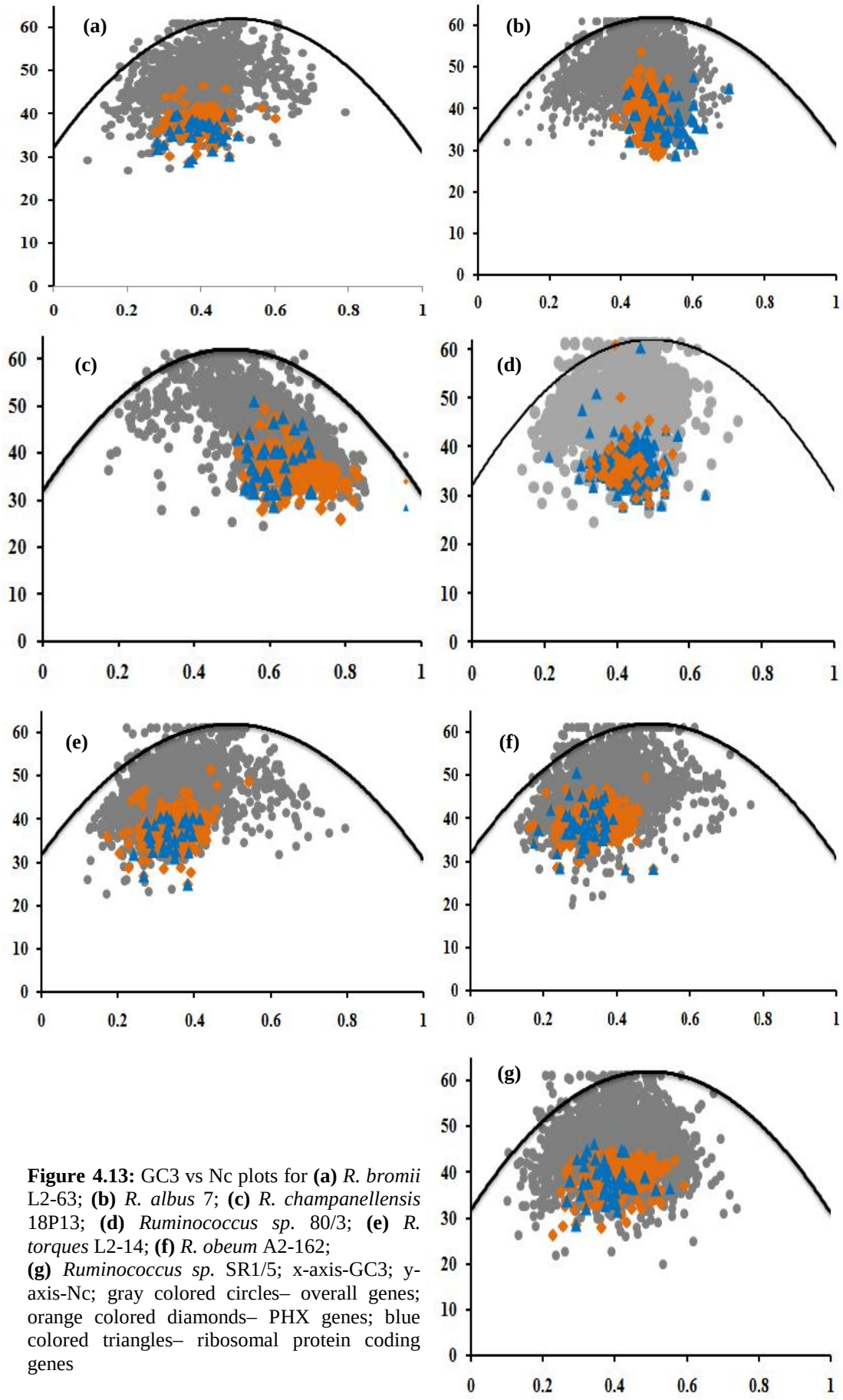

Figure 4.13: GC3 vs Nc plots for **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus sp.* 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus sp.* SR1/5; x-axis-GC3; y-axis-Nc; gray colored circles– overall genes; orange colored diamonds– PHX genes; blue colored triangles– ribosomal protein coding genes

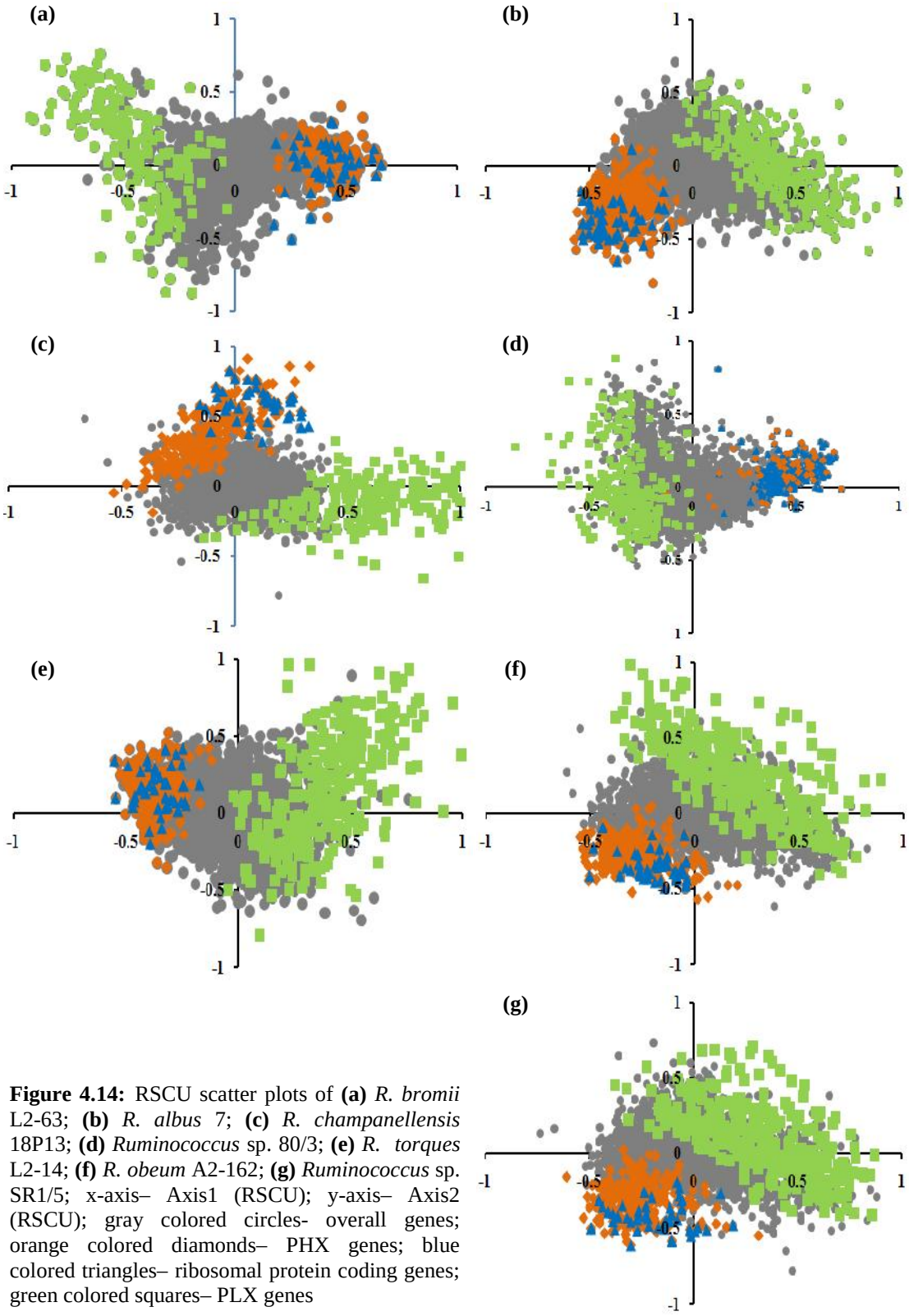

Figure 4.14: RSCU scatter plots of **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus* sp. 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus* sp. SR1/5; x-axis– Axis1 (RSCU); y-axis– Axis2 (RSCU); gray colored circles- overall genes; orange colored diamonds– PHX genes; blue colored triangles– ribosomal protein coding genes; green colored squares– PLX genes

the various significant correlations between the codon usage parameters indicated a mighty effect of AT compositional constraint on the codon usage signatures of *Ruminococcus*. Codon and amino usage based heatmaps, roseplots and barplots emphasized the tendency of the amino acids to be coded preferentially by their AT rich codon counterparts in all the concerned strains of *Ruminococcus*, except *R. champanellensis* 18P13 where there was GC preference among the codon sets due to moderate GC rich genomic composition of the organism. Compositional bias was not the sole factor that determined codon usage behaviour of the genus. Considerable impact of natural selection for efficient translation was perceptible on the potential highly expressed genes and the ribosomal protein coding genes. Role of translational selection was also well evident from the fact that codon usage in *Ruminococcus* was considerably governed by tRNA population, aimed at enhancing translational efficiency. Similar instances of collective influence of compositional bias and selection for translational efficacy have been reported in various other organisms like *Escherichia coli* (Ikemura, 1981), *Caenorhabditis elegans* (Duret, 2000), *Frankia* (Sen *et al.*, 2008), *Bifidobacterium* (Roy *et al.*, 2015), etc. Correlation analysis of RSCU data and the various codon usage indices proved useful in inferring the role of factors other than compositional constraint and natural selection for translation. Strong correlations of codon adaptation index (CAI) with both the major axes of separation of genes on the basis of RSCU data i.e., Axis 1 and Axis 2 of RSCU, among all the concerned ruminococcal members (Table 4.12) established gene expression level to be another vital component affecting the codon usage of the genus. There have been previous reports where gene expression level influenced codon usage signatures in various prokaryotic and eukaryotic organisms like *C. elegans* (Duret and Mouchiroud, 1999), *Lactobacillus* (Nayak, 2012), *Drosophila melanogaster* (Moriyama and Powell, 1997), etc. Potential highly expressed (PHX) genes were mainly found to code for large and small subunit ribosomal proteins. Some were also found encode glycosidases and translation elongation factors. On the other hand, majority of the potential lowly expressed (PLX) genes were found to be hypothetical ones. Some

PLX genes were found to code for protein products associated with secretion.

Length of the protein coding sequences was found to display substantial significant correlations with both the major axes of separation of genes on the basis of RSCU data i.e., Axis 1 and Axis 2 of RSCU, in all the concerned strains of *Ruminococcus* (Table 4.12). Such results clearly indicated that the length of coding sequences had marked impact on the observed codon usage variations. It was also evident from Table 4.12 that gene expression level, as estimated by codon adaptation index (CAI), exhibited significant positive correlations with the length of the coding sequences. It could be deduced from such correlations that the potential highly expressed genes preferred longer stretch of coding sequences.

There have been similar reports of positive correlation between highly expressed gene products and length of coding sequences in other bacterial communities like *Pseudomonas aeruginosa* (Gupta and Ghosh, 2001) and *Streptococcus pneumoniae* (Hou and Yang, 2002) where such correlations have been justified from the perspective of avoiding misincorporation and missense errors during translation in prokaryotic genomes. However, results have been found to be different in case of eukaryotic organisms like *D. melanogaster* (Miyasaka, 2002) and *C. elegans* (Marais and Duret, 2001) where length of coding sequences exhibited significant negative correlations with the level of gene expression. In this context, it has been suggested by Moriyama and Powell (1998) that shorter proteins are always preferred over the longer stretches as the longer proteins tend to be more expensive in terms of biosynthetic energy and selection pressure acts to reduce the length of the highly expressed gene products in the large eukaryotic genomes.

Thus, codon usage fashion of the genus *Ruminococcus* appeared to be governed by several critical factors like compositional bias, natural selection for translation, level of gene expression and length of coding sequences. Gene expressivity was found to be the most crucial determinant affecting codon usage of the genus.

4.1.2.2 Bias in the usage of successive synonymous codon pairs:

Successive synonymous codon pairs are defined as the pairs of synonymous

Table 4.13 Frequencies of the favored, neutral and disfavored types of synonymous codon pairs for nine amino acids in *R. bromii* L2-63

Organism: *R. bromii* L2-63						
	Isoaccepting			Non-isoaccepting		
Grouped by	Favored	Neutral	Disfavored	Favored	Neutral	Disfavored
Alanine	4 (4 IPs)	0	4	1	2	5
Arginine	10 (6 IPs + 4 Non IPs)	0	2	2	4	18
Glycine	6 (4 IPs+ 2 Non IPs)	0	0	2	0	8
Isoleucine	3 (3 IPs)	0	2	0	2	2
Leucine	8 (6 IPs + 2 Non IPs)	2	0	0	0	26
Proline	4 (4 Ips)	2	2	2	2	4
Serine	6 (6 IPs)	2	2	0	5	21
Threonine	4 (4 Ips)	0	2	3	1	6
Valine	4 (4 Ips)	0	2	4	0	6
Total	49	6	16	14	16	96

IP: Identical codon pair; Non Ip: Non Identical co-tRNA codon pair. Codon pairs were categorized into isoaccepting (sharing a tRNA) and non-isoaccepting (not sharing a tRNA) types. Each specific group was further classified into: Favored (≥ 3 Standard Deviation), Neutral (ranging between -3 and +3 Standard Deviation) and Disfavored (≤ -3 Standard Deviation) categories

codons that occur consecutively, irrespective of the number of non-synonymous codons present between them (Cannarozzi *et al.*, 2010). Considerable research on the arrangement patterns of synonymous codons has led to realization of the fact that bias in the usage and distribution of successive synonymous codon pairs improve efficiency of translation (Buchan *et al.*, 2006; Fedorov *et al.*, 2002). In order to corroborate the hypothesis in the genus *Ruminococcus*, we estimated the frequencies of successive synonymous codon pairs among the concerned ruminococcal strains. It was palpable from our investigations that the identical codon pairs (e.g., GTC-GTC codon pair for Valine, CTC-CTC codon pair for Leucine, etc.) and non-identical co-tRNA codon pairs (e.g., GTC-GTT codon pair for Valine, CTC-CTT codon pair for Leucine, etc.) were employed by the ruminococcal members in higher frequencies than expected. However, usage of non-isoaccepting consecutive codon pairs was universally avoided by the ruminococcal strains. Frequencies of

Table 4.14 Frequencies of overall and PHX genes present in the leading and lagging strands of replication (represented in percentage) in *Ruminococcus*

Organism	Count of Leading strand genes (%)	Count of Lagging strand genes (%)	Count of PHX genes in leading strand (%)	Count of PHX genes in lagging strand (%)
R. bromii L2-63	53.78	46.22	61.90	38.10
R. albus 7	55.40	44.60	57.23	42.77
R. champanellensis 18P13	59.25	40.75	64.59	35.41
Ruminococcus sp. 80/3	56.95	43.05	55.13	44.87
R. torques L2-14	62.40	37.60	60.44	39.56
R. obeum A2-162	70.47	29.53	82.47	17.53
Ruminococcus sp. SR1/5	56.05	43.95	47.92	52.08

favored, neutral and disfavored codon pairs both for the isoaccepting and non-isoaccepting tRNA types for the nine amino acids (with at least two tRNAs), in the strain *R. bromii* L2-63 (taken as a representative of the genus *Ruminococcus*), have been enlisted in Table 4.13. There have been previous reports of successive synonymous codon usage bias in prokaryotic (Guo *et al.*, 2012) and eukaryotic (Cannarozzi *et al.*, 2010) organisms and such preferences have been found to facilitate translational exactitude. Thus, observed bias in the usage of successive synonymous codon pairs among members of *Ruminococcus* implied the motif of the genus to achieve acme of translation.

4.1.2.3 Evaluation of replicational-transcriptional selection:

Microbes displaying strand-specific codon usage bias are considered to be under the influence of replication-transcription associated selection constraints (Guo and Yu, 2007); (Guo and Yuan, 2009; McInerney, 1998). Strand-specific asymmetry can often be inferred by the presence of a higher number of protein coding sequences on the leading strand of replication than on the lagging strand and by differential codon usage patterns of the genes representing the complimentary strands of replication (Das *et al.*, 2005; Lopez and Philippe, 2001). Frequencies of genes present on the leading and lagging strands of replication were estimated in all concerned strains of *Ruminococcus* and have been provided in Table 4.14. It was evident from Table 4.14 that all concerned strains displayed a higher percentage of genes on the leading

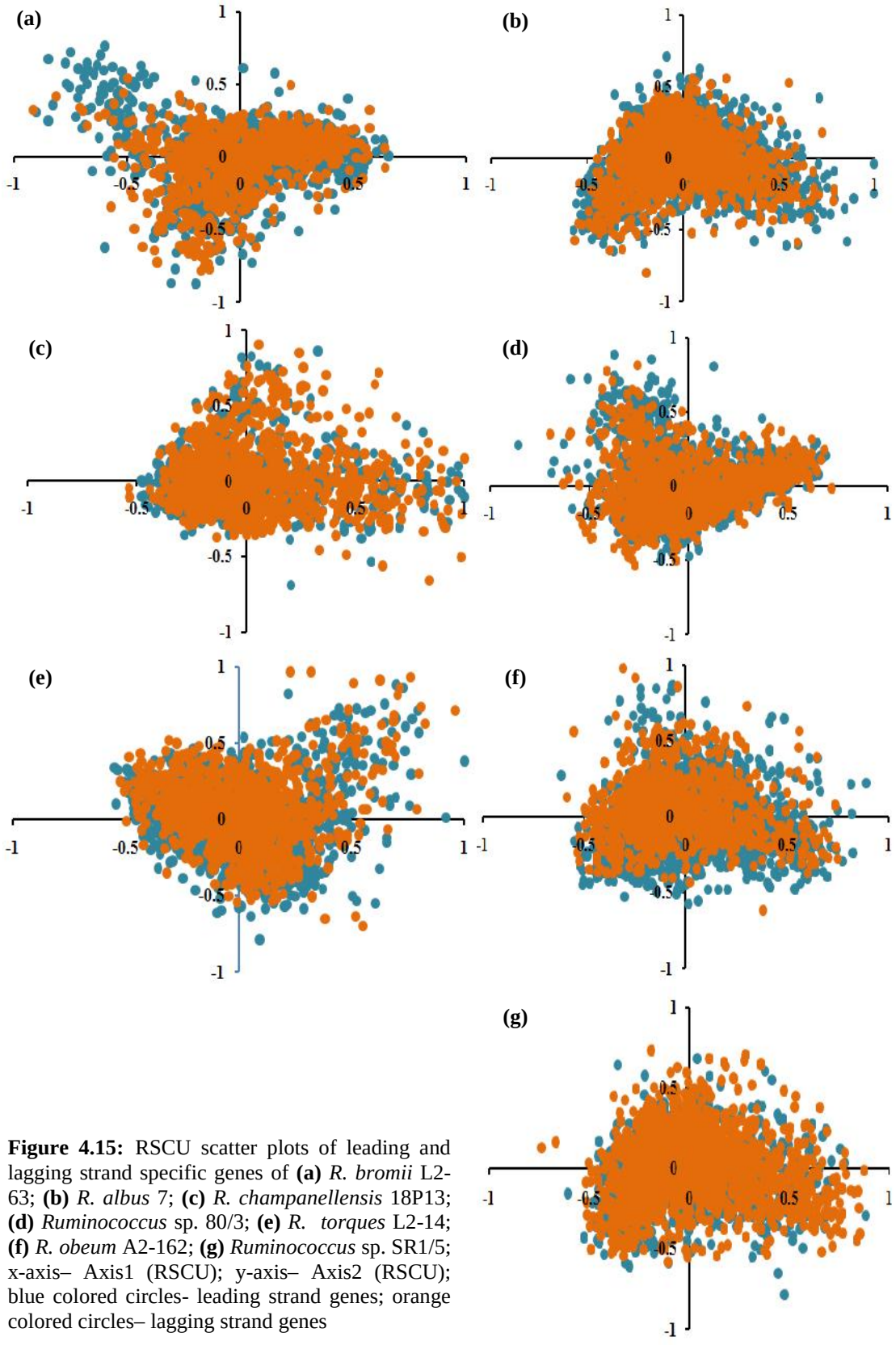

Figure 4.15: RSCU scatter plots of leading and lagging strand specific genes of **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus* sp. 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus* sp. SR1/5; x-axis– Axis1 (RSCU); y-axis– Axis2 (RSCU); blue colored circles- leading strand genes; orange colored circles– lagging strand genes

strand of replication. However, on thorough inspection of the RSCU scatter plots for all the concerned strains (Figure 4.15) it was distinct that the genes present in the leading and lagging strands of replication assembled together without forming separate clusters. Thus, no strand-specific asymmetry was noticed which pointed towards the absence of replication-associated mutational selection in the genus.

Presence of a higher percentage of highly expressed essential genes on the leading strand of replication has been considered to indicate the effect of transcription-associated selection pressure on a microbial genome (Das *et al.*, 2006; Rocha and Danchin, 2003). We calculated the respective frequencies of potential highly expressed (PHX) genes on the complimentary strands of replication and the detailed account has been provided in Table 4.14. All the strains of *Ruminococcus*, except *Ruminococcus* sp. SR1/5, were found to possess higher percentage of potential highly expressed genes on the leading strand of replication. Such findings pointed towards a probable role of transcriptional selection on the concerned strains, except *Ruminococcus* sp. SR1/5. However, further investigations are required to arrive at any definite conclusion.

4.1.2.4 Comparative codon usage analysis of Ruminococcus and human host:

Various members of the genus *Ruminococcus* inhabit human intestinal environment and serve as key players in maintaining a 'steady state' of the gut. Successful adaptation of *Ruminococcus* in human intestine has been a topic of immense research interest. We executed a comparative investigation of codon usage patterns of the concerned genus and related human host (*Homo sapiens*) in order to explore the molecular underpinnings of successful bacterial adaptation in human intestine. Codons were defined as over-represented (RSCU > 1.6) and under-represented (RSCU < 0.6) as per scheme followed by Wong *et al.* (2010). Similar codon usage pattern was concluded when a particular codon was found to display RSCU value less than 0.6 or more than 1.6 or found to fall within a range of 0.6 to 1.6 for both human host and concerned strain of *Ruminococcus*. The comparative tally of codon usage values of *R. bromii* L2-63 (taken as a representative of the genus *Ruminococcus*) and human host

Table 4.15 Comparative codon usage profile of *R. bromii* L2-63 and *H. sapiens*

Codon	*H. sapiens* (RSCU)	*R. bromii* L2-63 (RSCU)	Codon	*H. sapiens* (RSCU)	*R. bromii* L2-63 (RSCU)
UUU	0.93	1.29	**GCG**	0.42	0.52
UUC	1.07	0.71	**UAU**	0.89	1.19
UUA	0.46	0.53	**UAC**	1.11	0.81
UUG	0.77	0.69	**CAU**	0.84	0.94
CUU	0.79	2.88	**CAC**	1.16	1.06
CUC	1.17	1.14	**CAA**	0.53	0.48
CUA	0.43	0.11	**CAG**	1.47	1.52
CUG	2.37	0.65	**AAU**	0.94	1.02
AUU	1.08	1.58	**AAC**	1.06	0.98
AUC	1.41	0.91	**AAA**	0.87	1.09
AUA	0.51	0.51	**AAG**	1.13	0.91
GUU	0.73	2.05	**GAU**	0.93	1.05
GUC	0.95	0.44	**GAC**	1.07	0.95
GUA	0.47	0.98	**GAA**	0.84	1.31
GUG	1.85	0.52	**GAG**	1.16	0.69
UCU	1.13	0.79	**UGU**	0.91	0.95
UCC	1.31	0.58	**UGC**	1.09	1.05
UCA	0.9	1.76	CGU	0.48	1.45
UCG	0.33	0.83	**CGC**	1.10	0.70
CCU	1.15	1.08	CGA	0.65	0.42
CCC	1.29	0.91	CGG	1.21	0.30
CCA	1.11	0.43	**AGU**	0.90	0.74
CCG	0.45	1.58	**AGC**	1.44	1.30
ACU	0.99	0.74	AGA	1.29	2.54
ACC	1.42	0.78	AGG	1.27	0.59
ACA	1.14	1.94	**GGU**	0.65	1.51
ACG	0.46	0.54	**GGC**	1.35	1.46
GCU	1.06	0.96	**GGA**	1.00	0.89
GCC	1.60	0.72	GGG	1.00	0.14
GCA	0.91	1.80			

RSCU: Relative Synonymous Codon Usage values; Similarly selected codons are marked in bold

has been provided in Table 4.15. It was prominent that 40 out of 59 codons (excluding the single synonymous codons AUG [Met] and UGG [Trp] and the three termination codons) were found to be selected similarly between

Table 4.16 Bivariate correlation analysis (Spearman's Rank correlation co-efficient values) among the various amino acid usage parameters in *Ruminococcus*

Organ-ism		Axis1	Axis2	GRAVY	Aromo	CAI	PEC	GC
R. bromii L2-63	**Axis1**	1.00	0.06**	0.74**	0.33**	-0.03	0.13**	-0.22**
	Axis2		1.00	-0.41**	0.02	0.20**	-0.35**	-0.16**
	GRAVY			1.00	0.02	-0.04	0.07**	-0.03
	Aromo				1.00	-0.18**	0.79**	-0.40**
	CAI					1.00	-0.30**	0.15**
	PEC						1.00	-0.46**
	GC							1.00
Organ-ism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
R. albus 7	**Axis1**	1.00	0.01	-0.19**	0.13**	0.29**	-0.25**	-0.17**
	Axis2		1.00	-0.81**	0.31**	-0.13**	0.33**	-0.35**
	GRAVY			1.00	-0.10**	-0.09**	0.01	0.07**
	Aromo				1.00	-0.19**	0.79**	-0.34**
	CAI					1.00	-0.38**	0.42**
	PEC						1.00	-0.42**
	GC							1.00
Organ-ism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
R. champan-ellensis 18P13	**Axis1**	1.00	-0.16**	0.37**	-0.13**	-0.25**	0.05*	0.58**
	Axis2		1.00	-0.61**	-0.30**	0.03	-0.03	-0.27*
	GRAVY			1.00	-0.07**	-0.02	0.01	0.26**
	Aromo				1.00	-0.16**	0.78**	-0.31**
	CAI					1.00	-0.25**	0.26**
	PEC						1.00	-0.38**
	GC							1.00
Organ-ism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
Rumino-coccus sp. 80/3	**Axis1**	1.00	-0.12**	-0.73**	0.05**	-0.16**	0.25**	-0.21**
	Axis2		1.00	0.57**	-0.18**	-0.11**	0.16**	0.31**
	GRAVY			1.00	-0.07**	-0.04*	0.01	0.14**
	Aromo				1.00	-0.22**	0.79**	-0.34**
	CAI					1.00	-0.34**	0.19**
	PEC						1.00	-0.42**
	GC							1.00

Contd… to next page.

Table 4.16 *continued from previous page*

Organism		Axis1	Axis2	GRAVY	Aromo	CAI	PEC	GC
R. torques L2-14	**Axis1**	1.00	0.12**	-0.90**	-0.04**	-0.09**	0.05**	-0.16**
	Axis2		1.00	-0.02	0.70**	-0.47**	0.85**	-0.53**
	GRAVY			1.00	0.01	0.01	0.04	0.07**
	Aromo				1.00	-0.19**	0.82**	-0.36**
	CAI					1.00	-0.35**	0.13**
	PEC						1.00	-0.47**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
R. obeum A2-162	**Axis1**	1.00	-0.19**	0.90**	0.07**	0.10**	0.03	0.16**
	Axis2		1.00	-0.07**	0.44**	-0.45**	0.75**	-0.25**
	GRAVY			1.00	0.01	0.05**	0.06**	0.09**
	Aromo				1.00	-0.16**	0.80**	-0.31**
	CAI					1.00	-0.31**	0.09**
	PEC						1.00	-0.41**
	GC							1.00
Organism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
Ruminococcus sp. SR1/5	**Axis1**	1.00	-0.13**	0.87**	0.06**	0.07**	-0.03	0.21**
	Axis2		1.00	0.03	0.60**	-0.44**	0.83**	-0.34**
	GRAVY			1.00	0.02	-0.03	0.06**	0.15**
	Aromo				1.00	-0.16**	0.81**	-0.31**
	CAI					1.00	-0.31**	0.20**
	PEC						1.00	-0.43**
	GC							1.00

**: statistically significant at $P < 0.01$; *: statistically significant at $P < 0.05$; Aromo: Aromaticity of encoded proteins; PEC: Protein energetic costs of encoded proteins; Axis1: Axis1 of RAAU; Axis2: Axis2 of RAAU

R. bromii L2-63 and human host. Trend was similar when other concerned strains of *Ruminococcus* were compared with human. Selection of similar codons has been reported in bacterial (Ma *et al.*, 2016) and viral pathogens (Zhou *et al.*, 2013; Wong *et al.*, 2010) where the concerned pathogens display codon usage patterns resembling their target host. Such resemblance in codon usage patterns have been associated with successful co-evolution and adaptation of the microbes in the concerned host (Cutter *et al.*, 2006; Zhao *et al.*, 2003; Zhong *et al.*, 2007). Thus, similar modes of

codon usage and selection of identical codons between the members of *Ruminococcu*s and concerned human host provide a clue towards congenial residence of the genus in human gut.

4.1.2.5 Factors dictating amino acid usage behavior:

Amino acid usage parameters like hydropathicity index (GRAVY), aromaticity and protein energetic cost (PEC) were calculated and correlated with relative amino acid usage (RAAU) data (Table 4.16). Hydropathicity index of the encoded protein products was found to possess strong correlations with the major axes of separation of the genes on the basis of RAAU data i.e., Axis 1 and Axis 2 of RAAU. RAAU data based scatter plots for the seven concerned strains of *Ruminococcus* suggested that the protein coding genes were separated into two different clusters- one representing the set with high hydrophobic character (positive GRAVY [hydrophobic]) (indicated by green colored triangles in the respective scatter plots in Figure 4.16) and the other group consisting of the genes coding for the hydrophilic proteins (negative GRAVY [hydrophilic]) (indicated by brown colored diamonds in the respective scatter plots in Figure 4.16). The group of highly hydrophobic gene products consisted of integral membrane proteins, transporter proteins, amino acid permeases and some hypothetical proteins whereas, the hydrophilic group mainly comprised of large and small subunit ribosomal proteins, site-specific recombinases, transcriptional regulators and several hypothetical proteins. Furthermore, aromaticity of the encoded protein products also correlated well with Axis 2 of RAAU data in all the strains, except *R. bromii* L2-63 where it was found to show higher correlation with Axis 1 of RAAU data (Table 4.16). Thus, aromaticity index also contributed to the observed amino acid usage variations. Protein energetic cost was also found to exhibit significant correlations with both Axis 1 and Axis 2 of RAAU data which suggested that apart from hydropathicity index and aromaticity of the gene products, biosynthetic cost of the of the encoded proteins also governed amino acid usage behaviour of the genus. Level of gene expression, as estimated by codon adaptation index (CAI), also correlated well with the two major axes of separation of genes on the basis of RAAU data, invariably among all the

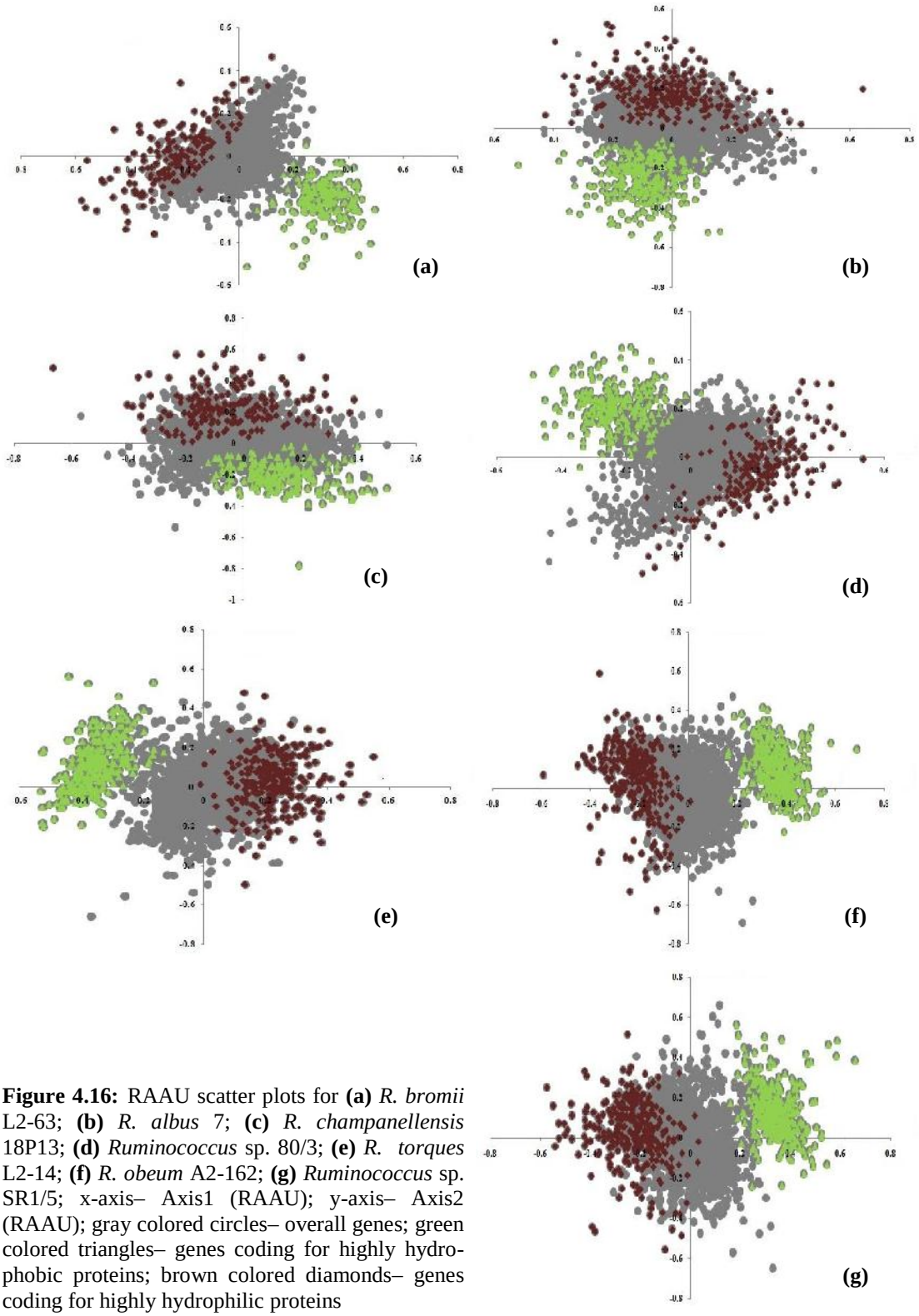

Figure 4.16: RAAU scatter plots for **(a)** *R. bromii* L2-63; **(b)** *R. albus* 7; **(c)** *R. champanellensis* 18P13; **(d)** *Ruminococcus* sp. 80/3; **(e)** *R. torques* L2-14; **(f)** *R. obeum* A2-162; **(g)** *Ruminococcus* sp. SR1/5; x-axis– Axis1 (RAAU); y-axis– Axis2 (RAAU); gray colored circles– overall genes; green colored triangles– genes coding for highly hydrophobic proteins; brown colored diamonds– genes coding for highly hydrophilic proteins

concerned strains (Table 4.16). Besides governing the codon usage patterns, gene expressivity also had significant impact on the amino acid usage of the genus *Ruminococcus*.

Correlation analysis proved handy in inferring close association of protein aromaticity and biosynthetic cost. High positive correlations between the two mentioned parameters (Table 4.16), consistently evident in all the ruminococcal members, re-established the fact that aromatic amino acids tend to be more costly in terms of protein energetic cost (Smith and Chapman, 2010).

Considerable negative correlations between gene expression level and protein energetic cost (Table 4.16) revealed an unanimous trend in the genus that the potential highly expressed genes preferred using economically cheaper amino acids (in terms of protein energetic cost). Protein energetic costs of the potential highly expressed gene products were found to be considerably lower than putative lowly expressed gene products (Supplementary Table 4 [ST4]). Such observations can be well explained in light of the fact that bacterial communities tend to be economic while synthesizing the essential proteins that exhibit high levels of expression and are employed frequently in various vital biological processes (Akashi and Gojobori, 2002).

4.1.3 Fashion of codon and amino acid usage in Helicobacter pylori:

4.1.3.1 Determinants of codon usage variations:

AT rich genomic composition of *Helicobacter pylori* was prominent from the analysis of overall codon usage patterns. The average genome size of the *H. pylori* strains was found to be around 1.59±0.04 millions of base pairs (Mbp) and the average AT composition was observed to be around 61.06±0.25%. It was distinct from overall codon usage pattern of *Helicobacter pylori* 35A (Table 4.17) that 14 amino acids preferred to be encoded by AT rich codons. Moreover, out of 28 codons with high relative synonymous codon usage (RSCU) values (RSCU > 1), 16 were found to be AT rich. Observations were similar in case of all other *H. pylori* strains considered for our analysis. Thus, there was a marked impact of AT compositional constraint on the overall codon usage pattern of *H. pylori*. Pronounced effect of AT compositional bias has already been inferred in *H. pylori* (Lafay *et al.*, 2000) and our

Table 4.17 Overall codon usage data of *H. pylori* 35A

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	**UUU**	20777	1.59	**Ser**	**UCU**	7617	1.40
	UUC	5393	0.41		UCC	2611	0.48
Leu	**UUA**	21149	2.35		UCA	3202	0.59
	UUG	14265	1.58		UCG	1812	0.33
	CUU	8109	0.90		AGU	4669	0.86
	CUC	4625	0.51		AGC	12773	2.34
	CUA	3969	0.44	Cys	UGU	1854	0.68
	CUG	1961	0.22		UGC	3619	1.32
Tyr	**UAU**	12328	1.41	Pro	CCU	7667	1.94
	UAC	5120	0.59		CCC	4182	1.06
His	**CAU**	6992	1.36		CCA	2427	0.61
	CAC	3278	0.64		CCG	1544	0.39
Gln	**CAA**	14701	1.70	**Arg**	CGU	2315	0.82
	CAG	2639	0.30		CGC	4129	1.46
Ile	**AUU**	17471	1.51		CGA	1262	0.45
	AUC	12817	1.10		CGG	542	0.19
	AUA	4530	0.39		**AGA**	4651	1.65
Asn	**AAU**	15803	1.16		**AGG**	4063	1.44
	AAC	11511	0.84	**Thr**	**ACU**	6384	1.22
Lys	**AAA**	33001	1.54		ACC	6493	1.24
	AAG	9858	0.46		ACA	3469	0.66
Val	**GUU**	7410	1.10		ACG	4567	0.87
	GUC	3556	0.53	Ala	GCU	12949	1.58
	GUA	2846	0.42		GCC	6476	0.79
	GUG	13146	1.95		GCA	3929	0.48
Asp	**GAU**	16419	1.48		GCG	9450	1.15
	GAC	5824	0.52	Gly	GGU	4700	0.67
Glu	**GAA**	23752	1.47		GGC	9808	1.40
	GAG	8511	0.53		GGA	3165	0.45
					GGG	10432	1.48

Amino acids preferentially coded by AU rich codons are marked in bold; RSCU: Cumulative Relative Synonymous Codon Usage values (RSCU) by summation of RSCU of all genes; N: Number of codons; AA: Amino acids

results were in complete agreement with the proposed fact.

Codon usage heatmap of *Helicobacter pylori* 35A, *Helicobacter pylori* 83, *Helicobacter pylori* 2017 and *Helicobacter pylori* Cuz20 strains [Figure 4.17 (a)] suggested that the AT rich codons like AAA, GAA, TTA,

Table 4.18 Comparative display of genomic and codon usage data for different strains of *H. pylori*

Organism	Genomic Data			Number of opti-mal codons	Number of 'A/U' rich optimal codons	RSCU	
	Size (Mbp)	A+T (mol %)	tRNAs			Axis 1 (Inertia %)	Axis 2
H. pylori 35A	1.56	61.13	36	24	07	7.88	5.10
H. pylori 83	1.62	61.28	36	22	04	7.76	5.02
H. pylori 2017	1.55	60.70	36	23	04	6.62	4.96
H. pylori Cuz20	1.64	61.14	36	23	05	7.33	4.77

Mbp: millions of base pairs; RSCU: Relative Synonymous Codon Usage

ATT, AAT and GAT were used in higher frequencies (reflected by yellow color intensity for these codons in the heatmap). The intensity of color in the heatmap gradually changed from red to yellow with increase in frequency of a particular codon. Roseplots were generated on the basis of codon usage data to estimate the usage of various codons (Figure 4.18). Roseplot analysis in each case supported the fact that all the *H. pylori* strains were inclined towards the usage of AT rich codons over their GC rich counterparts.

Amino acid usage based heatmap [Figure 4.17 (b)] for all concerned strains of *H. pylori* revealed that the amino acids like Lysine (Lys), Leucine (Leu), Glutamine (Gln), Isoleucine (Ile), Asparagine (Asn) and Aspartic acid (Asp) were used in relatively higher frequencies in comparison to the other amino acids (as was reflected by green color intensity for the mentioned amino acids in the heatmap). The intensity of color gradually changed from red to green in the concerned heatmap with increase in frequency of a particular amino acid. The same fact could be inferred from the amino acid usage based roseplots of the concerned strains (Figure 4.19).

It was evident from barplot analysis (Figure 4.20) that the third synonymous codon position was AT rich with a distinct bias towards the usage of Thymine, invariably in all *H. pylori* strains under study. Thus, it could be inferred from the respective codon and amino acid usage based heatmaps, roseplots and barplots that the strains were under a strong influence of AT compositional constraint. Complete codon usage statistics of the concerned H. pylori strains have been provided in Table 4.18.

Synonymous codon usage order effectively portrays the quantitative relationship between synonymous

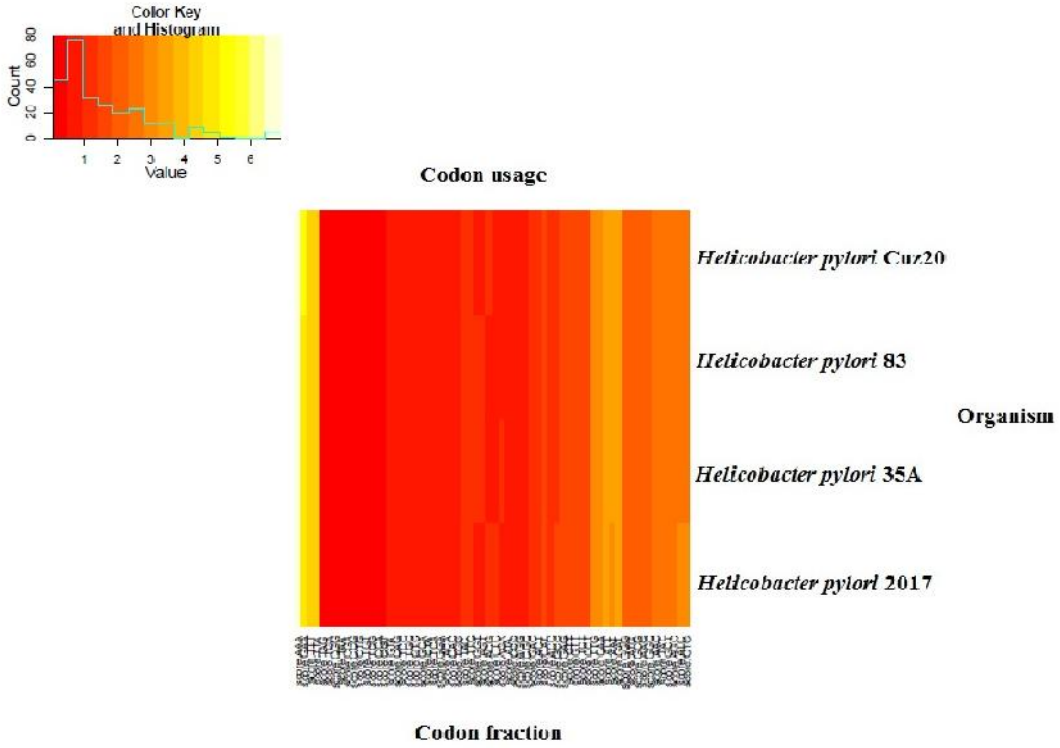

Figure 4.17 (a): Codon usage based two dimensional cluster heatmap. Heatmap based on the usage (frequency) of codons for all the *H. pylori* genomes. Color intensity changes from bright red to yellow with increase in frequency of a particular codon

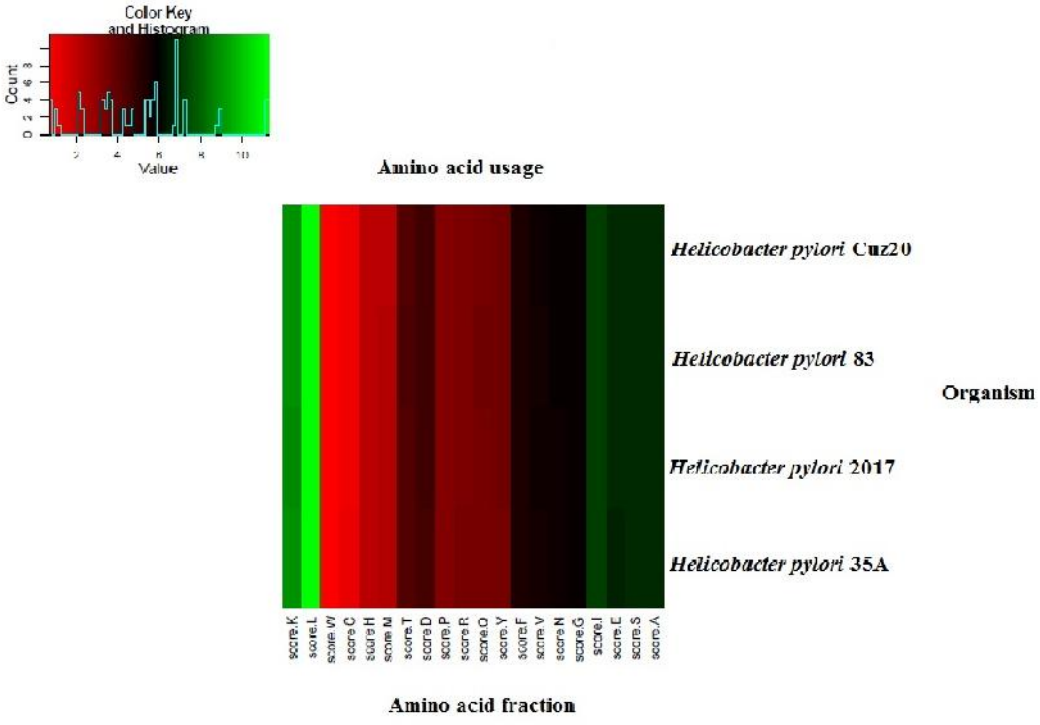

Figure 4.17 (b): Amino acid usage based two dimensional cluster heatmap**.** Heatmap based on the usage of amino acids by all the *H. pylori* genomes. Color intensity changes from bright red to green with increase in frequency of a particular amino acid

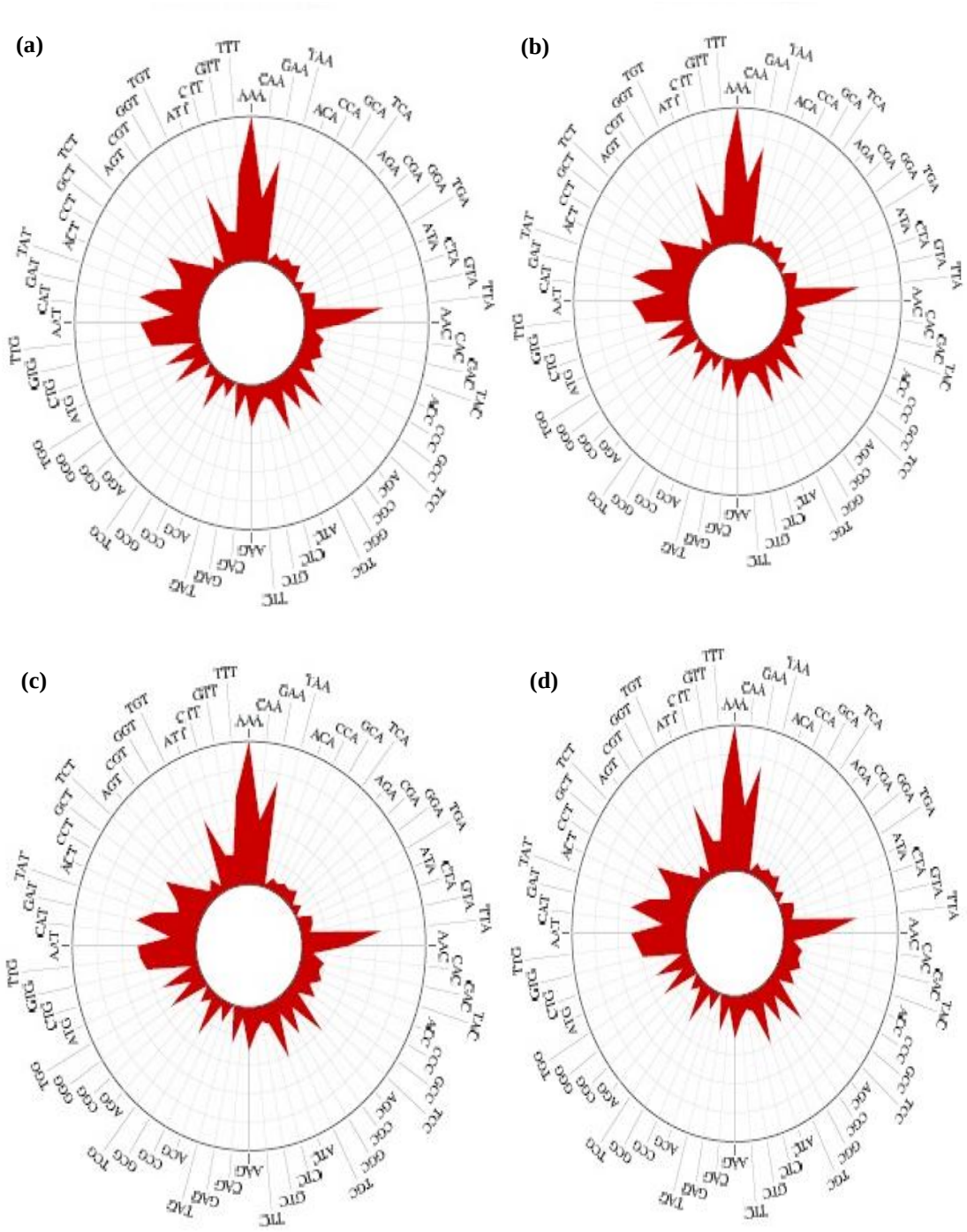

Figure 4.18: Codon usage roseplots for **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; Sharp red peaks represent codons used in higher frequencies

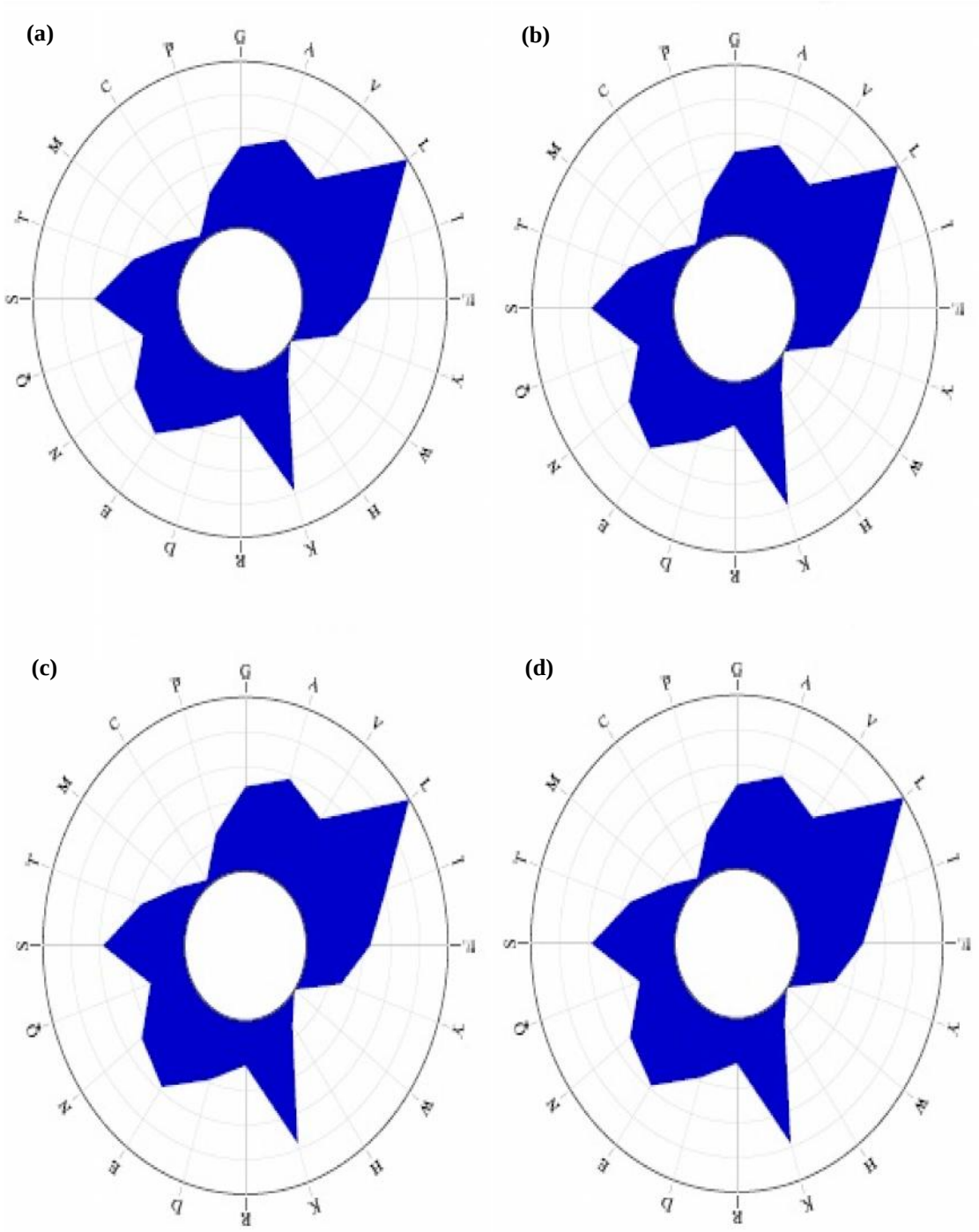

Figure 4.19: Amino acid usage roseplots for **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; Sharp blue peaks represent amino acids used in higher frequencies

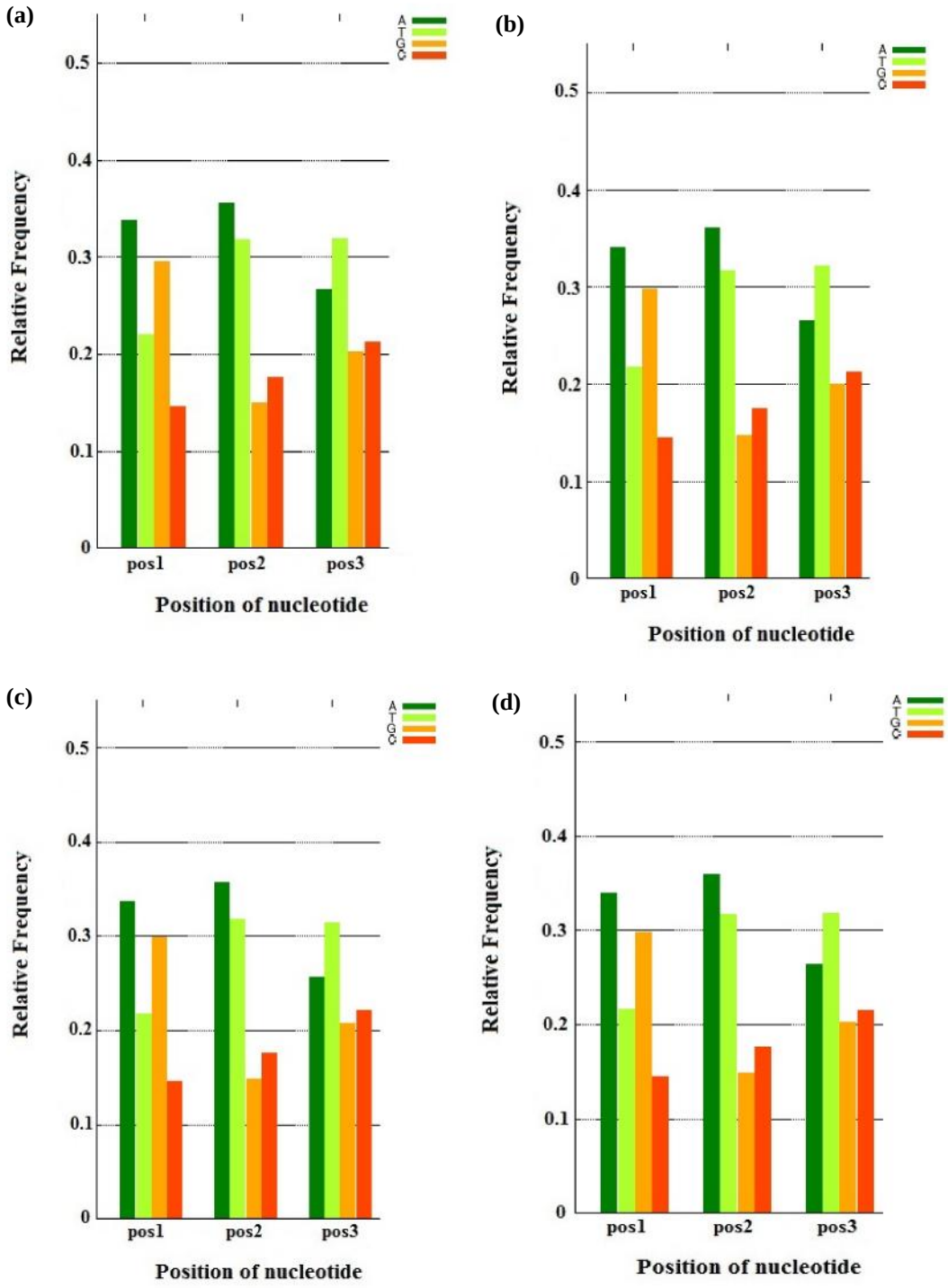

Figure 4.20: Barplots for **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; The color codes for the respective nucleotides have been provided in the upper panel of the figure: A- dark green; T- light green; G- yellow; C- orange

Table 4.19 Correlation of genomic AT composition with Synonymous Codon Usage Order (SCUO) (Spearman's Rank correlation co-efficient values) for *H. pylori* strains

Organism	SCUO vs AT (r)
H. pylori 35A	0.31**
H. pylori 83	0.34**
H. pylori 2017	0.36**
H. pylori Cuz20	0.35**

(r): Spearman's Rank correlation co-efficient; ** : statistically significant at $P < 0.01$; AT: AT composition

codon usage bias and genomic composition of an organism (Wan *et al*. 2004). We estimated SCUO for all the concerned strains and correlated pertaining data with AT composition of the concerned genomes (Table 4.19). Significant positive correlations suggested the pervasive effect of AT compositional constraint on the *H. pylori* strains.

Correspondence analysis of relative synonymous codon usage (RSCU) data, for all the strains, revealed that the Axes 1 and 2 produced major codon usage variations observed in the 59-dimensional hyperspace (indicated by high values of inertia). Axis 1 was found to be the principle axis, invariably for all strains, followed by Axis 2 (Table 4.20). Substantial correlations of Axis 1 of RSCU data with genomic (AT/GC) composition, in all concerned strains of *H. pylori*, further signified the marked influence of compositional bias (Table 4.20).

GC3 versus Nc (effective number of codons) plots have been suggested to be useful in proper investigation of codon usage data (Wright, 1990). Wright (1990) proposed that if codon usage of an organism is solely governed by genomic compositional constraint, then all the concerned genes would fall on the continuous Nc plot curve. GC3 versus Nc plots, for all the strains of *H. pylori* under scrutiny (Figure 4.21), revealed that significant share of genes were found to lie on or above the continuous Nc plot curve. Such an observation accentuated the undeniable effect of AT compositional bias on the concerned genomes. Interestingly, the ribosomal protein coding genes and the potential highly expressed (PHX) genes, in all the respective GC3 versus Nc plots, were found to lie below the continuous Nc plot curve which suggested the influence of other factors, apart from compositional bias.

In order to investigate the role of translational selection, RSCU data

Table 4.20 Bivariate correlation analysis (Spearman's Rank correlation co-efficient values) among the various codon usage indices in *H. pylori* strains

Organism		Axis1	Axis2	GC3	Nc	Length	GC	CAI	tAI
Helicobacter pylori 35A	**Axis1**	1.00	0.15**	0.48**	-0.15**	0.23**	0.22**	0.20**	0.29*
	Axis2		1.00	-0.05	0.16**	-0.07**	-0.09**	-0.29**	0.01
	GC3			1.00	0.05*	0.25**	0.73**	-0.07**	0.40**
	Nc				1.00	0.06*	0.06*	-0.46**	-0.08
	Length					1.00	0.23**	0.20**	0.27**
	GC						1.00	-0.05	0.27**
	CAI							1.00	0.25**
	tAI								1.00
Organism		**Axis1**	**Axis2**	**GC3**	**Nc**	**Length**	**GC**	**CAI**	**tAI**
Helicobacter pylori 83	**Axis1**	1.00	-0.02	-0.55**	0.11**	-0.23**	-0.31**	-0.16**	-0.27**
	Axis2		1.00	-0.01	-0.03	0.01	-0.05	-0.02	0.17**
	GC3			1.00	0.06*	0.22**	0.75**	-0.07**	0.40**
	Nc				1.00	0.10*	0.06*	-0.42**	-0.05
	Length					1.00	0.23**	0.20**	0.26**
	GC						1.00	-0.02	0.28**
	CAI							1.00	0.26**
	tAI								1.00
Organism		**Axis1**	**Axis2**	**GC3**	**Nc**	**Length**	**GC**	**CAI**	**tAI**
Helicobacter pylori 2017	**Axis1**	1.00	-0.05*	0.55**	-0.13**	0.28**	0.30**	0.20**	0.23**
	Axis2		1.00	0.11**	-0.18**	-0.03	0.07**	0.13**	0.07**
	GC3			1.00	0.03	0.28**	0.73**	-0.04	0.38**
	Nc				1.00	0.11**	0.06*	-0.45**	-0.07*
	Length					1.00	0.26**	0.20**	0.21**
	GC						1.00	-0.04	0.26**
	CAI							1.00	0.25**
	tAI								1.00
Organism		**Axis1**	**Axis2**	**GC3**	**Nc**	**Length**	**GC**	**CAI**	**tAI**
Helicobacter pylori Cuz20	**Axis1**	1.00	0.02	0.51**	-0.07**	0.17**	0.26**	0.13**	0.27**
	Axis2		1.00	0.24**	-0.17**	0.09**	0.19**	0.12**	0.10**
	GC3			1.00	0.09**	0.23**	0.74**	-0.08**	0.51**
	Nc				1.00	0.09**	0.07**	-0.45**	-0.02
	Length					1.00	0.22**	0.19**	0.25*
	GC						1.00	-0.05*	0.30**
	CAI							1.00	0.24**
	tAI								1.00

**: statistically significant at P < 0.01; *: statistically significant at P < 0.05; Length: Length of coding sequences; Axis1: Axis1 of RSCU; Axis2– Axis2 of RSCU

based scatter plots were generated for all the concerned *H. pylori* strains (Figure 4.22). Surprisingly, the potential highly expressed (PHX) genes and the potential lowly expressed (PLX) genes clustered together. This was a unanimous trend among all the strains under subject. Identical codon usage patterns of the PHX and PLX genes indicated the absence of translational selection on the highly expressed gene sets. In order to validate such an observation, we correlated tRNA adaptation index (tAI) with effective number of codons (Nc) (Table 4.20). Insignificant correlations re-established the absence of selection pressure for translational efficacy on the potential highly expressed genes. Similar results have been reported in *H. pylori* where low tRNA count (only a small set of 36 tRNA genes) has been attributed to the absence of translational selection (Lafay *et al.*, 2000; dos Reis *et al.*, 2004).

Gene expression, as estimated by codon adaptation index (CAI), was found display considerable correlations with Axis 1 of RSCU data (Table 4.20). Such significant correlations established gene expressivity as an important determinant of codon usage patterns. Gene expression has been observed to dictate the mode of codon usage patterns in various other organisms like *Mycobacterium tuberculosis* (Andersson and Sharp, 1996), *Bacillus subtilis* (Kanaya *et al.*, 1999), *Sinorhizobium meliloti* (Peixoto *et al.*, 2003), etc.

Majority of the potential highly expressed (PHX) gene products were found to be large and small subunit ribosomal proteins, kinases, dehydrogenases and restriction enzymes associated with vital biological functions. On the contrary, potential lowly expressed (PLX) gene products were mainly found to be hypothetical ones with unknown functions. However, some PLX genes were found to be associated with pathogenicity (cag pathogenicity island associated genes). Few PLX genes were also found to encode membrane proteins.

Length of the protein coding sequences was found to correlate significantly with Axis 1 of RSCU data for all the concerned strains of *H. pylori* (Table 4.20) and thus, could be attributed as a contributory factor for the variations observed in codon usage. Strong positive correlations between length of the coding sequences and CAI pointed towards the tendency of the potential

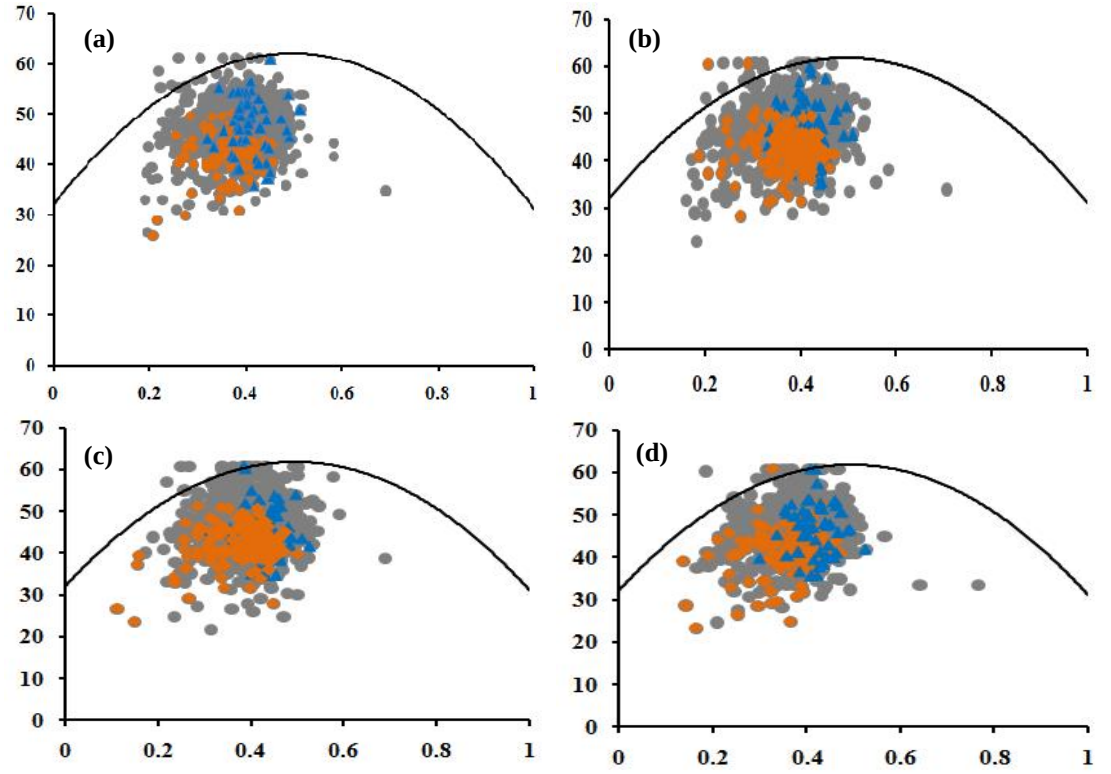

Figure 4.21: GC3 vs Nc plots for **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; x-axis-GC3; y-axis-Nc; gray colored circles– overall genes; orange colored diamonds– PHX genes; blue colored triangles– ribosomal protein coding genes

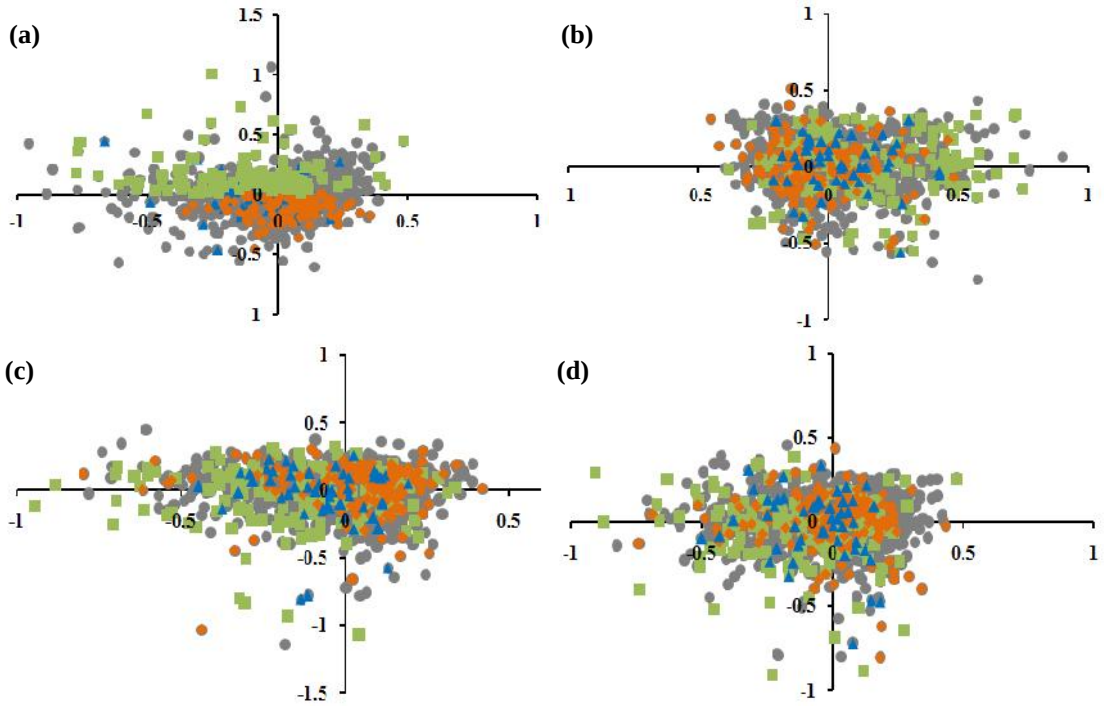

Figure 4.22: RSCU scatter plots of **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; x-axis– Axis1 (RSCU); y-axis– Axis2 (RSCU); gray colored circles- overall genes; orange colored diamonds– PHX genes; blue colored triangles– ribosomal protein coding genes; green colored squares– PLX genes

Table 4.21 Frequencies of the favored, neutral and disfavored types of synonymous codon pairs for nine amino acids in *H. pylori* 35A

Organism: *H. pylori* 35A						
	Isoaccepting			Non-isoaccepting		
Grouped by	Favored	Neutral	Disfavored	Favored	Neutral	Disfavored
Alanine	6 (6 IPs)	0	2	2	0	6
Arginine	10 (6 IPs + 4 Non IPs)	0	2	2	2	20
Glycine	4 (4 IPs)	0	2	0	2	8
Isoleucine	3 (3 IPs)	0	2	0	2	2
Leucine	8 (6 IPs + 2 Non IPs)	0	2	0	6	20
Proline	4 (4 Ips)	4	0	2	2	4
Serine	6 (6 IPs)	2	2	2	6	18
Threonine	4 (4 Ips)	0	2	2	2	6
Valine	4 (4 Ips)	0	2	0	2	8
Total	49	6	16	10	24	92

IP: Identical codon pair; Non Ip: Non Identical co-tRNA codon pair. Codon pairs were categorized into isoaccepting (sharing a tRNA) and non-isoaccepting (not sharing a tRNA) types. Each specific group was further classified into: Favored (≥ 3 Standard Deviation), Neutral (ranging between -3 and +3 Standard Deviation) and Disfavored (≤ -3 Standard Deviation) categories

highly expressed genes to be longer in stretch. It has been reported in other prokaryotic organisms like *Escherichia coli* (Eyre-Walker, 1996), *Bifidobacterium* (Roy *et al.*, 2015), etc., and it has been suggested that essential genes displaying high expression levels in bacterial genomes prefer to be longer in order avoid misincorporation errors during translation.

Thus, AT compositional bias was found to be the most crucial factor accounting for codon usage heterogeneity of the concerned *H. pylori* strains. Level of gene expression and length of the protein coding sequences also played significant role in shaping the codon usage patterns.

4.1.3.2 Predisposition in the usage of successive synonymous codon pairs:

Preferences in the usage of successive synonymous codon pairs considerably administer translational accuracy (Gutman and Hatfield, 1989; Irwin *et al.*, 1995). Pairs of synonymous codons that appear consecutively, regardless of the number of codons that code for different amino acids (non-synonymous codons) occurring between them, are termed as successive synonymous codon pairs (Cannarozzi

et al., 2010). Frequencies of successive synonymous codon pairs were estimated among all the *H. pylori* strains. It was perceptible through extensive scrutiny that the strains had a clear predilection towards the usage of identical pairs of codons (e.g., GGC-GGC codon pair for Glycine, CCC-CCC codon pair for Proline, etc.). Non-identical co-tRNA codon pairs (e.g., GGC-GGT codon pair for Glycine, CCC-CCT codon pair for Proline, etc.) were also employed in higher frequencies than expected. However, non-isoaccepting consecutive codon pairs appeared mostly under-represented. Frequencies of favored, neutral and disfavored codon pairs both for the isoaccepting and non-isoaccepting tRNA types for the nine amino acids (with at least two tRNAs), in the strain *H. pylori* 35A (taken as a representative of the *H. pylori* members), have been enlisted in Table 4.21. Bias in the usage of successive synonymous codon pairs i.e., preferring the usage of identical codon pairs and non-identical co-tRNA codon pairs and disallowing the practice of codon pairs for non-isoaccepting tRNAs, have already been reported in prokaryotic (Guo *et al.*, 2012) and eukaryotic (Cannarozzi *et al.*, 2010) organisms. Such predilections in the usage and arrangement of successive synonymous codon pairs have been found to associate strongly with translation precision. Thus, *H. pylori* strains, considered in our study, were found to exhibit specific bias in the usage and arrangement of successive synonymous codon pairs to enhance translation knack.

4.1.3.3 Estimation of replicational-transcriptional selection:

It has been reported that replicational-transcriptional selection dictates codon usage patterns in microbial genomes that exhibit strand asymmetry (Das *et al.*, 2006; Guo and Yuan, 2009; Wei and Guo, 2010). Presence of a higher number of genes on the leading strand of replication in comparison to the lagging strand and differential mode of codon usage of the genes residing on the leading and lagging strands of replication signify the effect of replicational selection to be operative in bacterial genomes (Das *et al.*, 2005; Lopez and Philippe, 2001). In order to examine the presence of replicational selection on the *H. pylori* strains, frequencies of genes on the leading and lagging strands were calculated and have been listed in Table 4.22. Protein coding genes were found to be shared

Table 4.22 Frequencies of overall and PHX genes present in the leading and lagging strands of replication (represented in percentage) in *H. pylori*

Organism	Count of Leading strand genes (%)	Count of Lagging strand genes (%)	Count of PHX genes in leading strand (%)	Count of PHX genes in lagging strand (%)
H. pylori 35A	49.90	50.10	53.42	46.58
H. pylori 83	51.84	48.16	55.97	44.16
H. pylori 2017	49.39	50.61	50.64	49.36
H. pylori Cuz20	47.73	52.27	47.71	52.29

almost equally among the complimentary strands of replication in all the concerned strains. It could be concluded from such an observation that no replication associated mutational pressure was instrumental on the codon usage behavior of the genomes under analysis. Relative synonymous codon usage (RSCU) analysis of the genes present on the leading and lagging strands of replication confirmed that the absence of replication associated selection pressure. It was clear from the RSCU scatter plots of *H. pylori* 35A, *H. pylori* 83, *H. pylori* 2017 and *H. pylori* Cuz20 (Figure 4.23) that the genes representing the complimentary strands of replication flocked together without forming discrete groups. Occurrence of a higher share of potential highly expressed (PHX) genes and/or essential genes on the leading strand of replication indicates transcriptional selection to be influential on microbial genomes (Das *et al.*, 2006; Rocha and Danchin, 2003). Frequencies of PHX genes present on the leading and lagging strands of replication in all *H. pylori* strains were calculated to investigate the effect of transcriptional selection pressure on them. Detailed numerical statistics has been provided in Table 4.22. A differential trend was observed among the concerned strains. *H. pylori* 35A and *H. pylori* 83 strains were found to exhibit a higher share of PHX genes on the leading strand of replication in comparison to the lagging strand. However, *H. pylori* Cuz20 strain displayed higher frequency of highly expressed genes to be transcribed from the lagging strand. Interestingly, there was an even share of PHX genes on the complimentary strands of replication in case of *H. pylori* 2017. Thus, no precise conclusion could be drawn regarding the effect of transcription-associated mutational pressure on the *H. pylori*

Table 4.23 Comparative codon usage profile of *H. pylori* 35A and *H. sapiens*

Codon	*H. sapiens* (RSCU)	*H. pylori 35A* (RSCU)	Codon	*H. sapiens* (RSCU)	*H. pylori 35A* (RSCU)
UUU	0.93	1.59	GCG	0.42	1.15
UUC	1.07	0.41	**UAU**	0.89	1.41
UUA	0.46	2.35	UAC	1.11	0.59
UUG	0.77	1.58	**CAU**	0.84	1.36
CUU	0.79	0.90	**CAC**	1.16	0.64
CUC	1.17	0.51	CAA	0.53	1.70
CUA	0.43	0.44	CAG	1.47	0.30
CUG	2.37	0.22	**AAU**	0.94	1.16
AUU	1.08	1.51	**AAC**	1.06	0.84
AUC	1.41	1.10	**AAA**	0.87	1.54
AUA	0.51	0.39	AAG	1.13	0.46
GUU	0.73	1.10	**GAU**	0.93	1.48
GUC	0.95	0.53	GAC	1.07	0.52
GUA	0.47	0.42	**GAA**	0.84	1.47
GUG	1.85	1.95	GAG	1.16	0.53
UCU	1.13	1.40	**UGU**	0.91	0.68
UCC	1.31	0.48	**UGC**	1.09	1.32
UCA	0.9	0.59	CGU	0.48	0.82
UCG	0.33	0.33	**CGC**	1.10	1.46
CCU	1.15	1.94	CGA	0.65	0.45
CCC	1.29	1.06	CGG	1.21	0.19
CCA	1.11	0.61	**AGU**	0.90	0.86
CCG	0.45	0.39	AGC	1.44	2.34
ACU	0.99	1.22	AGA	1.29	1.65
ACC	1.42	1.24	**AGG**	1.27	1.44
ACA	1.14	0.66	**GGU**	0.65	0.67
ACG	0.46	0.87	**GGC**	1.35	1.40
GCU	1.06	1.58	GGA	1.00	0.45
GCC	1.60	0.79	**GGG**	1.00	1.48
GCA	0.91	0.48			

RSCU: Relative Synonymous Codon Usage values; Similarly selected codons are marked in bold

strains.

4.1.3.4 Host specific codon usage signatures of Helicobacter pylori:

Helicobacter pylori is a notorious pathogen associated with gastritis and peptic ulcers in human beings (Lafay *et al.*, 2000; Kelly, 1998). Several strains of *H. pylori* reside in human gut and these organisms exhibit pathogenesis with dysbiosis of human intestine. In order to explicate the complex strategies exerted by *H. pylori* for proper adaptation and subsequent infection in human host, we analyzed the codon usage patterns of the pathogen and related human host. Codons were considered to be over-represented when RSCU > 1.6, and under-represented when RSCU < 0.6. Similar codon usage pattern was

inferred between concerned bacterium and human host (*Homo sapiens*) when a particular codon was found to display RSCU values less than 0.6 or more than 1.6 or found to fall within a range of 0.6 to 1.6 in both the cases. The complete details of codon usage of *H. pylori* 35A (taken as a representative of *H. pylori* members) and human has been provided in Table 4.23. It was interesting to note that 36 out of 59 codons (excluding the single synonymous codons AUG [Met] and UGG [Trp] and the three termination codons) were found to be selected similarly between *H. pylori* 35A and *H. sapiens*. This trend was persistent when codon usage patterns of other *H. pylori* strains (subject of this work) were compared with that of pertaining human host. Thus, *H. pylori* was found to resemble the codon usage patterns of its human host. Certain bacterial and viral pathogens adopt the codon usage patterns of their respective hosts to dodge host immune responses and to establish proper infection (Ma *et al.*, 2016; Zhou *et al.*, 2013; Wong *et al.*, 2010). Identical codon usage signatures might provide *H. pylori* an advantage to elude host immune responses and properly colonize and infect human gut.

4.1.3.5 Factors governing amino acid usage:

Correspondence analysis was performed on relative amino acid usage (RAAU) data of the concerned *H. pylori* strains with an objective to reveal the contributions of various factors on the observed amino acid usage variations. Correlation analysis of various amino acid usage parameters have been enlisted in Table 4.24. Axis 1 and Axis 2 of RAAU were found to display highest inertia values and accordingly proved to be the major axes of separation of genes for all the concerned strains. Hydropathicity index (GRAVY) (positive GRAVY [hydrophobic], negative GRAVY [hydrophilic]) was found to correlate strongly with Axis 1 and Axis 2 of RAAU in all the strains (Table). It was also noticeable from Table 4.24, that aromaticity of the encoded protein products exhibited significant correlations with Axis 1 of RAAU in all cases. RAAU based scatter plots of *H. pylori* 35A, *H. pylori* 83, *H. pylori* 2017 and *H. pylori* Cuz20 (Figure 4.24) revealed a distinct trend that the protein coding sequences segregated in two discrete groups. One group consisted of genes coding for protein products with high values of

Table 4.24 Bivariate correlation analysis (Spearman's Rank correlation co-efficient values) among the various amino usage parameters in *H. pylori* strains

Organ-ism		Axis1	Axis2	GRAVY	Aromo	CAI	PEC	GC
H. pylori 35A	**Axis1**	1.00	-0.13**	0.66**	0.51**	-0.16**	0.32**	0.27**
	Axis2		1.00	-0.59**	-0.03	0.04	-0.25**	-0.19**
	GRAVY			1.00	0.20**	-0.17**	0.24**	0.17**
	Aromo				1.00	-0.05	0.83**	-0.32**
	CAI					1.00	-0.08**	-0.05
	PEC						1.00	-0.43**
	GC							1.00
Organ-ism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
H. pylori 83	**Axis1**	1.00	0.15**	0.66**	0.49**	-0.16**	0.31**	0.30**
	Axis2		1.00	0.62**	-0.03	-0.06*	0.21**	0.25**
	GRAVY			1.00	0.19**	-0.18**	0.24**	0.20**
	Aromo				1.00	-0.04	0.85**	-0.31**
	CAI					1.00	-0.10**	-0.04
	PEC						1.00	-0.42**
	GC							1.00
Organ-ism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
H. pylori 2017	**Axis1**	1.00	0.17**	0.65**	0.52**	-0.20**	0.34**	0.26**
	Axis2		1.00	0.62**	0.14**	-0.13**	0.34**	0.12**
	GRAVY			1.00	0.23**	-0.20*	0.25**	0.14**
	Aromo				1.00	-0.08**	0.85**	-0.32**
	CAI					1.00	-0.12**	-0.04
	PEC						1.00	-0.45**
	GC							1.00
Organ-ism		**Axis1**	**Axis2**	**GRAVY**	**Aromo**	**CAI**	**PEC**	**GC**
H. pylori Cuz20	**Axis1**	1.00	-0.15**	-0.67**	-0.50**	0.16**	-0.31**	-0.31**
	Axis2		1.00	0.59**	0.05*	-0.04	0.29**	0.18**
	GRAVY			1.00	0.22**	-0.17**	0.25**	0.19**
	Aromo				1.00	-0.02	0.84**	-0.30**
	CAI					1.00	-0.08**	-0.05*
	PEC						1.00	-0.43**
	GC							1.00

**: statistically significant at $P < 0.01$; *: statistically significant at $P < 0.05$; Aromo: Aromaticity of encoded proteins; PEC: Protein energetic costs of encoded proteins; Axis1: Axis1 of RAAU; Axis2: Axis2: Axis2 of RAAU

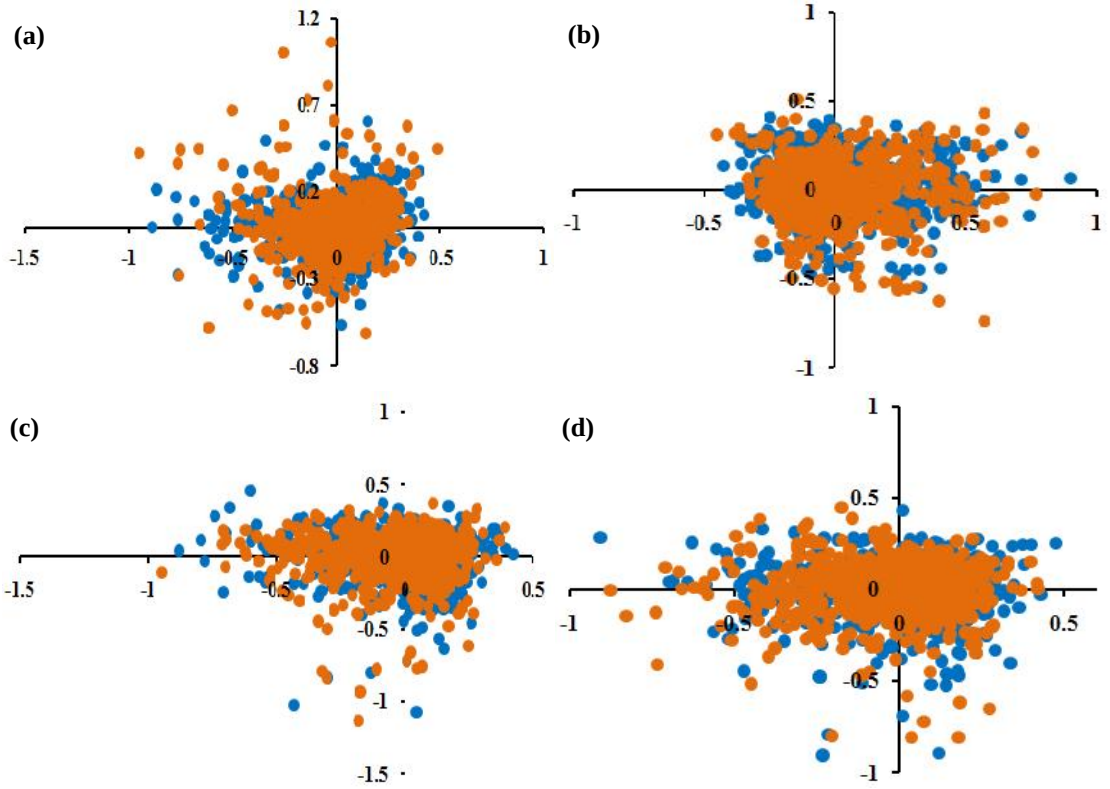

Figure 4.23: RSCU scatter plots of leading and lagging strand specific genes of **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; x-axis– Axis1 (RSCU); y-axis– Axis2 (RSCU); blue colored circles- leading strand genes; orange colored circles– lagging strand genes

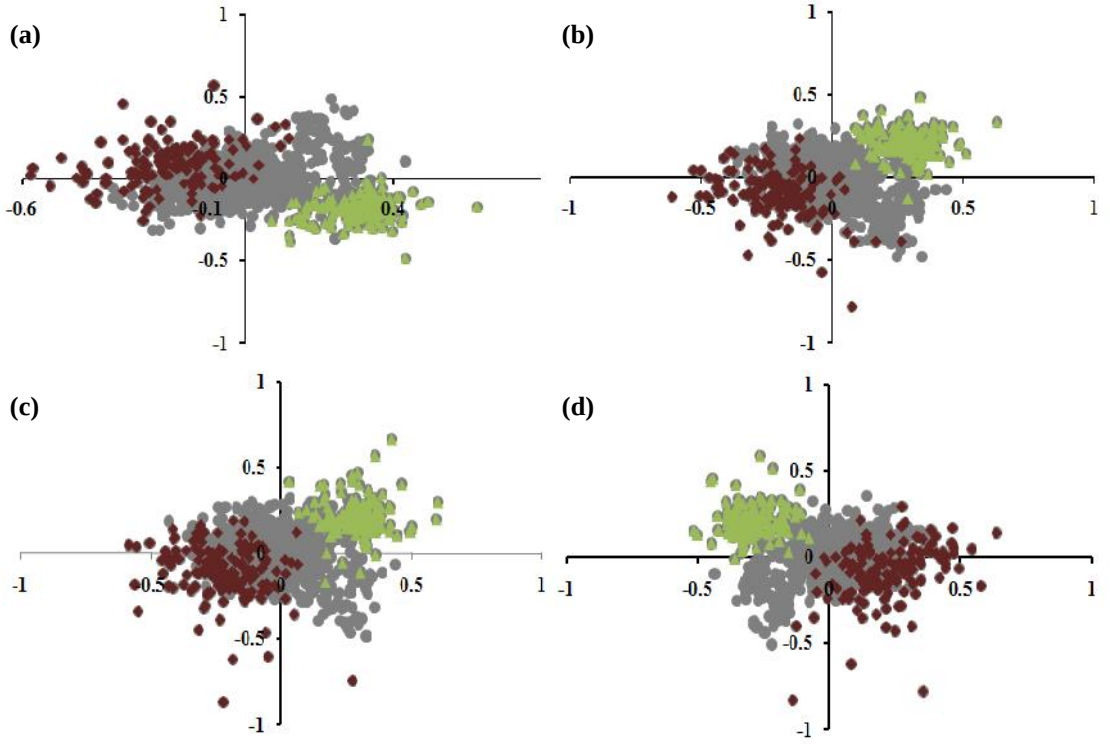

Figure 4.24: RAAU scatter plots for **(a)** *H. pylori* 35A; **(b)** *H. pylori* 83; **(c)** *H. pylori* 2017; **(d)** *H. pylori* Cuz20; x-axis– Axis1 (RAAU); y-axis– Axis2 (RAAU); gray colored circles– overall genes; green colored triangles– genes coding for highly hydrophobic proteins; brown colored diamonds– genes coding for highly hydrophilic proteins

hydrophobicity (high GRAVY) (indicated by green colored triangles in Figure 4.24) and the other group represented sequences that were found to code for proteins with low values of hydrophobicity (low GRAVY) (indicated by brown colored diamonds in Figure 4.24).

Substantial correlations were also perceptible between Axis 1 of RAAU and codon adaptation index (CAI), an estimator of gene expression level, for all genomes of *H. pylori* (Table 4.24). Apart from influencing the codon usage patterns, gene expression level was also found to shape amino acid usage patterns in the *H. pylori* strains. Biosynthetic cost of the encoded protein products (protein energetic cost) was found to display meaningful correlations with Axis 1 of RAAU (Table 4.24) for all the strains of interest. Such an observation pointed to the fact that energetic cost of the encoded protein products also had enough say on the differential mode of amino acid usage.

Strong positive correlations of aromaticity index of the proteins and the protein biosynthetic cost among the *H. pylori* strains (Table 4.24) re-established the fact that aromatic amino acids are economically more expensive (Smith and Chapman 2010). Another interesting inference that could be drawn from the strong negative correlations of protein energetic cost and CAI was that the potential highly expressed gene products had a distinct preference towards the usage of less costly (inexpensive) amino acids. Our results emphasized the fact that the *H. pylori* strains strictly maintain economy while synthesizing the essential highly expressed gene products. Such a tendency seemed justified from the aspect of bacterial cost-minimization strategy (Akashi and Gojobori, 2002).

Thus, various determinants like hydrophobicity, aromaticity and biosynthetic cost of the encoded proteins were observed to dictate the amino acid usage of the concerned *H. pylori* strains. Gene expression level was also found to be operative on the amino acid usage behaviour of *H. pylori*.

4.2 Comparative genomics:

Rapid advancement of Human Microbiome Project has generated enormous genomic and proteomic data that demand fruitful analysis. Comparative genomics and proteomics based research have been extremely instrumental in unleashing various

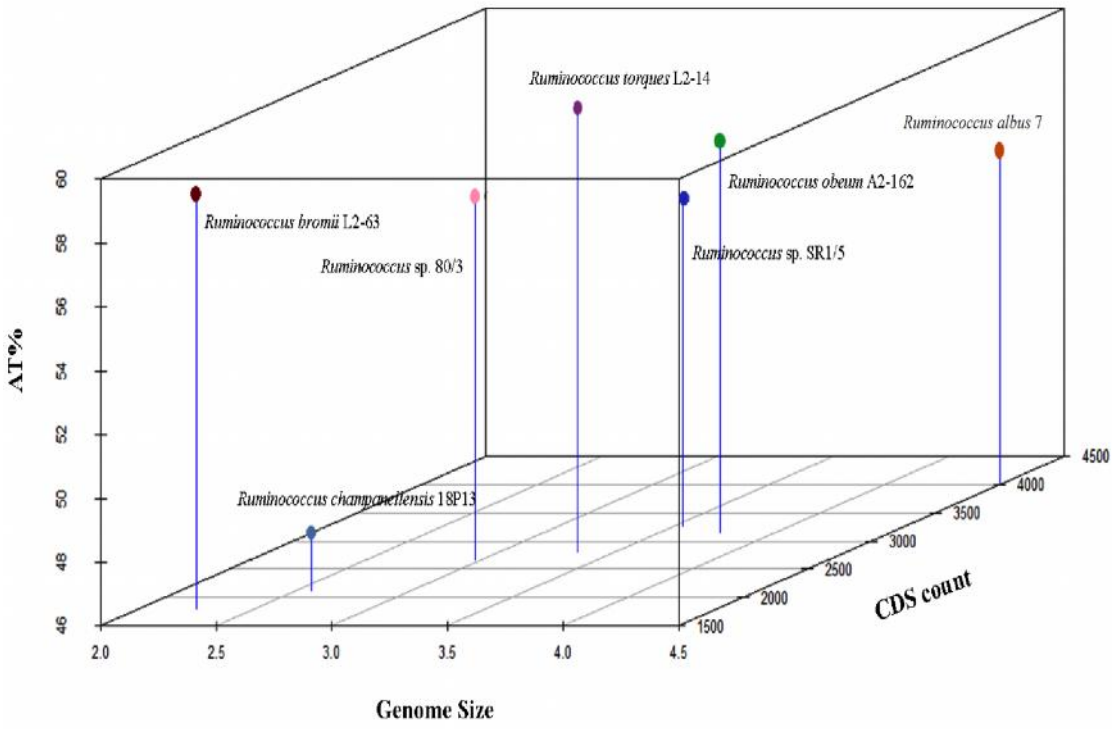

Figure 4.25: 3D plot of *Ruminococcus*– 3D plot represents genomic composition (in terms of percentage of Adenine and Thymine [AT%]), genome size and frequency of coding sequences (CDS count) of the respective ruminococcal members in three dimensional hyperspace

genomic and proteomic features of the gut microbes. Detailed comparative genomics and proteomics based analysis of various microbial members of human gut like *Bifidobacterium*, *Lactobacillus*, *Escherichia coli*, *Enterococcus*, *Veillonella*, etc., have already been accomplished leading to proper realization of their adaptive tactics for smooth survival in gut environment (Lukjancenko *et al.*, 2012; Lukjancenko *et al.*, 2010; Palmer *et al.*, 2012; Vesth *et al.*, 2013). However, comparative genomic and proteomic scrutiny of various important genera like *Ruminococcus*, *Bacteroides* and *Eubacterium* still remain unanalyzed. Accordingly, in this present approach we plan to execute extensive genomic and proteomic profiling of members of the concerned genera to gain a deeper insight into the mechanisms of successful residence of these bacterial members in the human intestinal niche. We also aim an extensive pan-genome and core genome analysis of concerned bacterial genera to elucidate the basic functional traits and symbiotic mechanisms that the microbes exhibit with human host.

Table 4.25 Comparative genomic statistics of various members of *Ruminococcus*

Organism	Genome Size (Mbp)	CDS Count	AT%	Unique gene families	Core genome Size	Pan-genome size
R. bromii L2-63	2.24	1811	58.95	1014	224	12757
R. albus 7	4.49	4014	56.38	2245		
R. champanellensis 18P13	2.57	2114	47.79	1201		
Ruminococcus sp. 80/3	2.97	2670	57.38	1196		
R. torques L2-14	3.34	2798	59.86	1335		
R. obeum A2-162	3.76	3155	58.25	1328		
Ruminococcus sp. SR1/5	3.54	3260	56.24	1568		

CDS: Coding sequences; Mbp: millions of base pairs

4.2.1 Comparative genomics of Ruminococcus:

4.2.1.1 Genomic architecture of Ruminococcus:

Members of the genus *Ruminococcus* were found to display AT rich genomic constitution. *Ruminococcus bromii* L2-63 (Ze *et al.*, 2012), *Ruminococcus albus* 7 (Suen *et al.*, 2011), *Ruminococcus torques* L2-14 (Lawson and Finegold, 2015), *Ruminococcus obeum* A2-162 (Lawson and Finegold, 2015), *Ruminococcus* sp. 80/3 (Wegmann *et al.*, 2014) and *Ruminococcus* sp. SR1/5 strains have been reported to be AT rich. However, *Ruminococcus champanellensis* 18P13 has been found to exhibit moderate GC richness (Collins *et al.*, 1994). A three dimensional plot (3D plot), displaying genomic composition (in terms of percentage of Adenine and Thymine [AT%]), genome size and frequency of coding sequences (CDS count) of the respective ruminococcal members was generated by plotting the mentioned parameters in three-dimensional hyperspace (Figure 4.25). It was well evident from the 3D plot that there was significant variation in genome size among the members of the genus. Genome size of *R. bromii* L2-63 was relatively smaller compared to the other members of the genus. The 3D plot has also been instrumental in revisiting the fact that the genus prefers to be moderately AT rich with *R. champanellensis* 18P13 as an exception (as observed earlier in this thesis in Section 4.1.2).

4.2.1.2 Multi-locus sequence analysis (MLSA) based phylogeny of Ruminococcus:

Multi-locus sequence analysis (MLSA) based phylogenetic approach uses concatenated multiple housekeeping

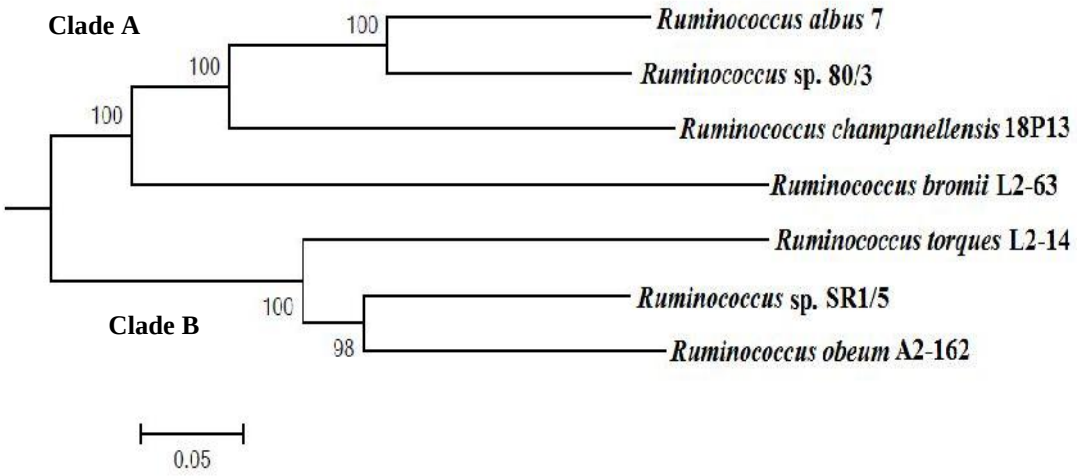

Figure 4.26: MLSA based phylogenetic tree of *Ruminococcus*

gene sequences to identify the phylogenetic aspects of desired species and investigates the patterns of genotypes between each species cluster. It was prominent from the MLSA based phylogenetic tree (Figure 4.26) that the ruminococcal members were segregated in two distinct clades. Clade A consisted of *R. albus* 7, *R. bromii* L2-63, *R. champanellensis* 18P13 and *Ruminococcus* sp. 80/3. Clade B comprised of *R. torques* L2-14, *R. obeum* A2-162 and *Ruminococcus* sp. SR1/5. It has been reported earlier that the genus *Ruminococcus* is not monophyletic and can be divided into two phylogenetically separate groups on performing 16S rRNA based phylogeny (Krause *et al.*, 1999a; Chassard *et al.*, 2010). The cellulolytic members like *R. albus* 7, *R. champanellensis* 18P13 and amylolytic member *R. bromii* L2-63 represent clostridial rRNA cluster IV whereas, the non-fibrolytic members like *R. torques* L2-14, *R. obeum* A2-162 and *Ruminococcus* sp. SR1/5 represent clostridial rRNA cluster XIVa (Collins *et al.*, 1994; Krause *et al.*, 1999b). Our results were in absolute accordance with such observations where Clades A and B in Figure 4.26 resembled clostridial rRNA cluster IV and clostridial rRNA cluster XIVa respectively.

4.2.1.3 Pan-Core genome plot and petal plot analysis:

The seven concerned strains of *Ruminococcus* produced a set of 12,757 gene families in the pan-genome and 224 gene families in the

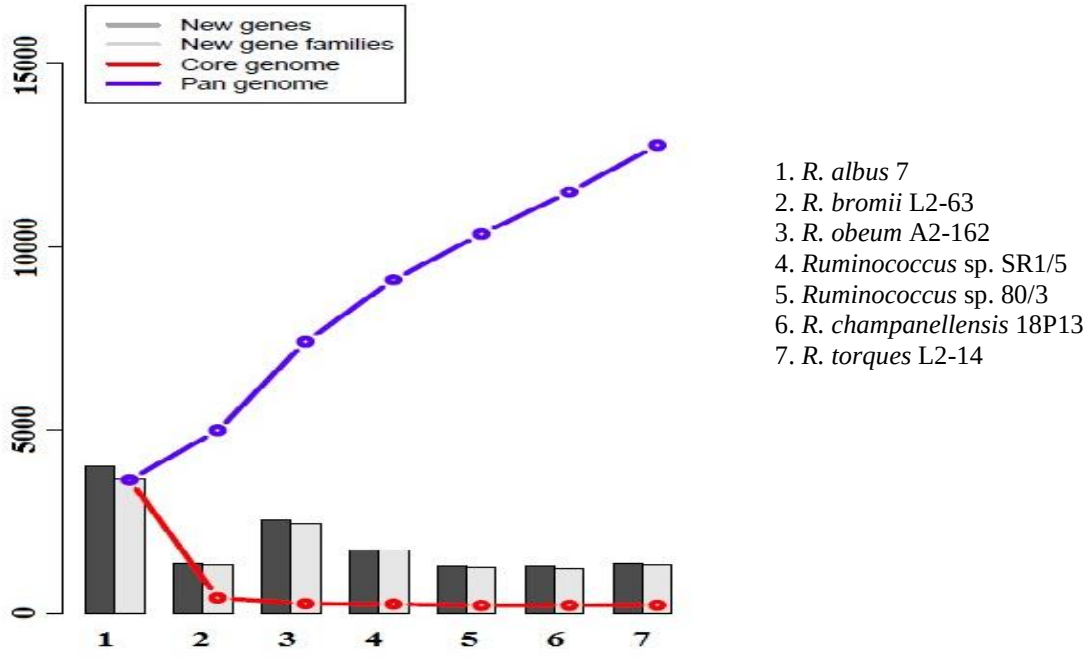

Figure 4.27: Pan-Core genome plot for *Ruminococcus*- The pan-genome (blue line) and core genome (red line) plot for *Ruminococcus* genomes. The number of discovered novel genes (dark bars) and novel gene families (light-grey bars) are also shown for each added genome

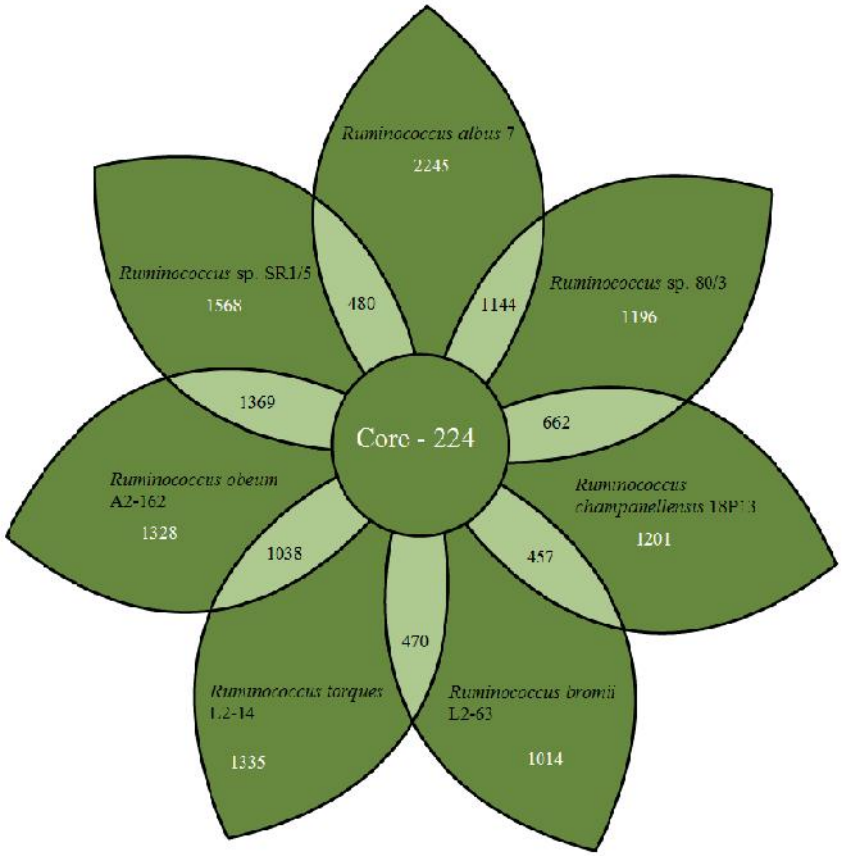

Figure 4.28: Petal plot of *Ruminococcus*- Petal plot of *Ruminococcus* displaying frequencies of core gene families and unique gene families

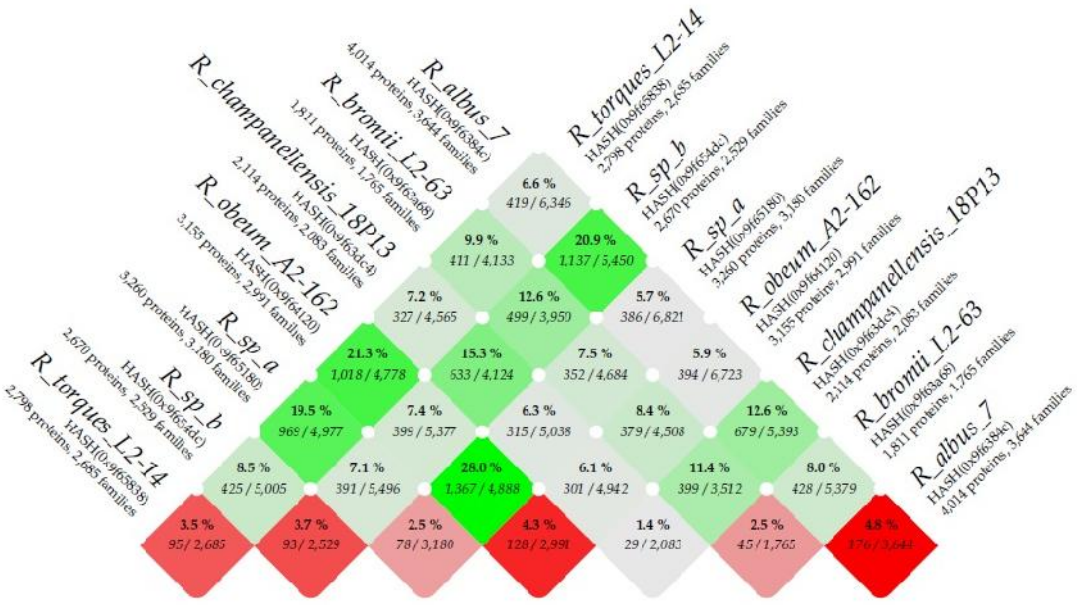

Figure 4.29: BLAST Matrix of *Ruminococcus*– To the side, the total number of protein genes and gene families are listed for each genome. In the matrix cells, the shared protein genes are given as a percentage, based on the ratio of the core genome and pan-genome of each pair; Green color represents homology between proteomes; Red color represents homology within proteomes; R_sp_a- *Ruminococcus* sp. SR1/5; R_sp_b- *Ruminococcus* sp. 80/3

core genome (Table 4.25). It was evident from the Pan-Core genome plot (Figure 4.27) that there was a sharp fall in the core genome curve (indicated by red color) when there was shift from *R. albus* 7 to *R. bromii* L2-63. Reduced genome size for *R. bromii* L2-63 might have accounted for such an observation. As ruminococcal genomes were subsequently added for analysis, the Pan-Core genome plot displayed significant changes in both the pan-genome and core genome curves. There was considerable jump in both the curves. The size of pan-genome was found to increase and the size of core genome was found to decrease with the addition of new ruminococcal strains. The frequencies of unique gene families of all ruminococcal strains were also estimated and have been displayed in the form of a petal plot (Figure 4.28). Ruminococcal members representing clostridial rRNA cluster IV were found to display a high share of common gene families (intersecting gene families in Figure 4.28) among them i.e., 1,144 gene families among *R. albus* 7 and *Ruminococcus* sp. 80/3, 662 gene families among *Ruminococcus* sp. 80/3 and *R. champanellensis* 18P13 and 457 gene families among *R. champanellensis* 18P13 and *R. bromii* L2-63. Same was the case with the ruminococcal members reflecting clostridial rRNA

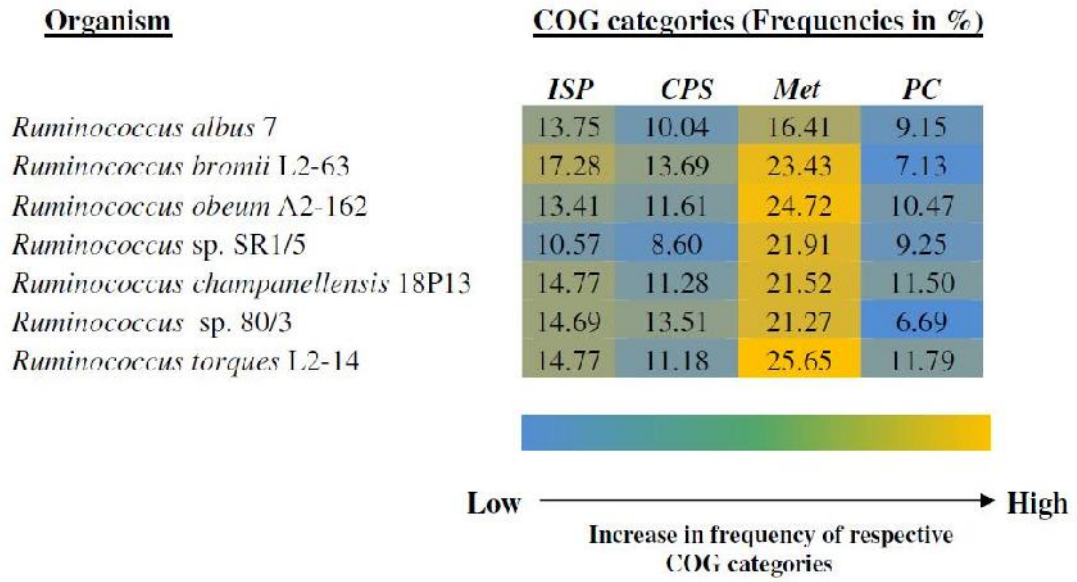

Figure 4.30: COG category heatmap of *Ruminococcus*- Heatmap based on the frequencies (in percentage) of various COG categories of ruminococcal proteomes. In the heatmap, *ISP* refers to *Information Storage and Processing*, *CPS* refers to *Cellular Processes and Signaling*, *Met* refers to *Metabolism*, *PC* refers to *Poorly Characterized* COG categories. Change of color intensity from blue to bright yellow refers to increase in frequency of a particular COG category (color scale provided below the heatmap)

cluster XIVa where *R. torques* L2-14 and *R. obeum* A2-162 shared 1,038 common gene families and *R. obeum* A2-162 and *Ruminococcus* sp. SR1/5 shared 1,369 common gene families.

4.2.1.4 BLAST matrix analysis of Ruminococcus:

BLAST matrix, displaying the similarity between the ruminococcal proteomes based on a pair-wise comparative scrutiny, was depicted in Figure 4.29. The percentage of similarity between proteomes, for each combination, has been presented by green color and the percentage of similarity within the same proteome has been depicted by red color in Figure 4.29. It was prominent from the BLAST matrix that ruminococcal strains like *R. albus* 7, *R. bromii* L2-63, *R. champanellensis* 18P13 and *Ruminococcus* sp. 80/3 (members of Clade A in Figure 4.26), representing clostridial rRNA cluster IV, shared higher degree of similarity among them. Same was the case among the non-fibrolytic ruminococcal members like *R. torques* L2-14, *R. obeum* A2-162 and *Ruminococcus* sp. SR1/5 (members of Clade B in Figure 4.26), reflecting clostridial rRNA cluster XIVa.

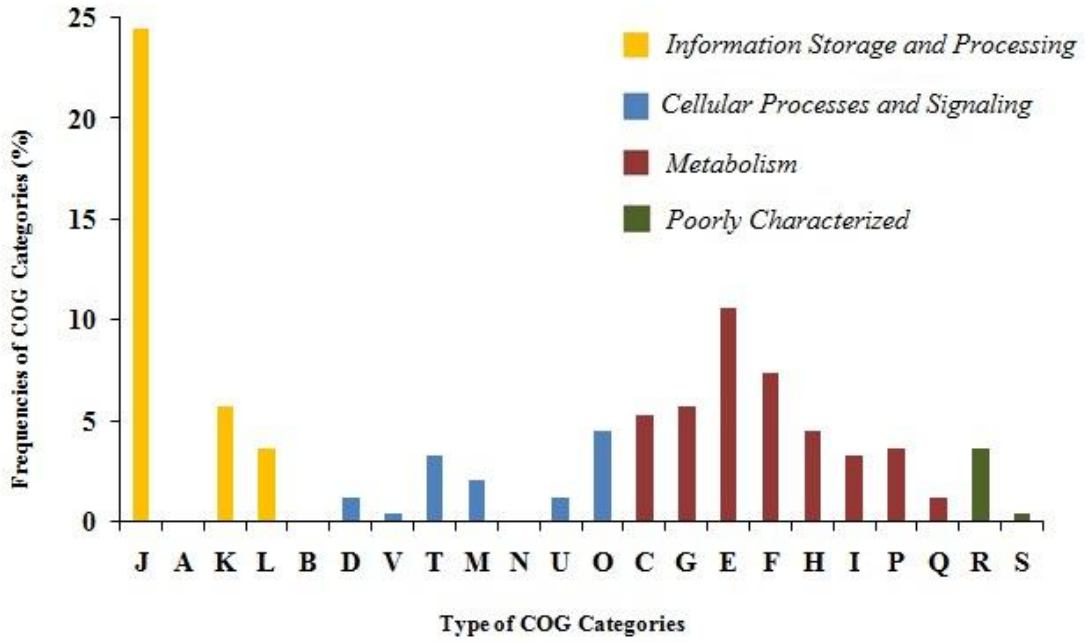

Figure 4.31: Bar diagram displaying COG categories of core proteome of *Ruminococcus*

4.2.1.5 Investigation of COG categories among the members of Ruminococcus:

COG categories of proteins present in all strains of *Ruminococcus* were estimated and displayed in the form of a heatmap (Figure 4.30). There was a gradual change of color shade from blue to bright yellow with increase in frequency of a particular COG category. It was markedly visible that the '*Metabolism'* COG category represented highest frequencies of proteins, invariably in all concerned strains. Considerable proportions of proteins were also found to fall in the '*Information Storage and Processing'* COG category. These proteins were the key components involved in crucial biological processes like translation, transcription and biogenesis of the organisms. Few proteins were poorly categorized. Hypothetical proteins with undetermined COG categories were not taken into consideration.

Further COG category inspection of the ruminococcal proteomes revealed that '*Energy production and conversion* [C]', '*Carbohydrate transport and metabolism* [G]', '*Amino acid transport and metabolism* [E]' and '*Nucleotide transport and metabolism* [F]' sub-categories, representing the '*Metabolism*' COG category, were found to be prevalent among the core proteomic set of *Ruminococcus* (Figure

4.31). '*Translation, ribosomal structure and biogenesis* [J]' sub-category representing the '*Information Storage and Processing*' COG category was also predominant among the core proteome of *Ruminococcus*.

Large proportion of proteins representing '*Metabolism*' COG category implied that the proteomic architecture of *Ruminococcus* has been apposite for metabolic mechanism and subsequent acclimatization by proper utilization of available undigested nutrients in the human gastrointestinal tract. It has been well known that various ruminococcal members utilize host-indigestible complex carbohydrates that remain non-degraded in the large intestine of the host (Flint *et al.*, 2012; Ze *et al.*, 2012; Chassard *et al.*, 2012; Chassard *et al.*, 2010; Ze *et al.*, 2015). Our results were in agreement with previous reports suggesting that significant share of proteins in various strains of *Ruminococcus* are involved in carbohydrate metabolizing functions (White *et al.*, 2014; Tasse *et al.*, 2010; Flint *et al.*, 2012; Chassard *et al.*, 2010; Ze *et al.*, 2012). Proper degradation of undigested carbohydrates, dietary fibers and amino acids render a healthy environment in human gut aiding in proper digestion and maintenance of a healthy state. Metabolically sophisticated proteomic machinery not only provides the ruminococcal members a competitive advantage over the other bacterial members that compete to establish their existence in the human intestinal niche, but also enhances the host's energy balance (Flint *et al.*, 2012; Ze *et al.*, 2012).

4.2.1.6 Carbohydrate degradation pathways of Ruminococcus:

Comparative scrutiny of carbohydrate degradation pathways in the genus *Ruminococcus* and host *Homo sapiens* (human) was conducted to properly elucidate the mechanisms of carbohydrate degradation employed by the ruminococcal members. The complete list of carbohydrate degradation pathways present in *Ruminococcus* but absent in *H. sapiens* has been provided in Table 4.26. Melibiose degradation pathway and xylose degradation I pathway, absent in *H. sapiens*, were present in all ruminococcal members, except *R. bromii* L2-63. Glycogen degradation I pathway and sucrose degradation II (sucrose synthase) pathway, absent in human, were found to be occurring in all members of the genus *Ruminococcus*, except *Ruminococcus*

Table 4.26 List of carbohydrate degradation pathways present in *Ruminococcus* but absent in *H. sapiens*

	Organism						
Carbohydrate Degradation Pathway	R_bro	R_alb	R_cha	R_sp. 80/3	R_tor	R_obe	R_sp. SR1/5
N-acetylneuraminate and N-acetylmannosamine degradation	X	X	X	X	√	√	X
Cellulose degradation II (fungi)	X	√	X	X	X	X	X
Chitin degradation II	√	X	X	X	X	X	√
Chitobiose degradation	X	√	X	X	X	X	X
Fructose degradation	√	X	X	√	√	√	√
Fucose degradation	X	√	X	X	√	√	√
Galactose degradation III	√	X	X	X	√	√	√
Glucose and glucose-1-phosphate degradation	X	X	√	X	X	√	√
Glucuronoarabinoxylan degradation	X	√	X	X	X	X	X
Glycogen degradation I	√	√	√	X	√	√	√
Homogalacturonan degradation	X	√	X	X	X	X	X
L-arabinose degradation I	X	X	X	X	√	√	√
L-rhamnose degradation I	X	X	X	X	√	X	X
Lactose and galactose degradation I	X	√	X	X	X	X	X
Melibiose degradation	X	√	√	√	√	√	√
Pectin degradation III	X	√	X	X	X	X	X
Sucrose degradation I (sucrose phosphotransferase)	X	X	X	X	X	X	√
Sucrose degradation II (sucrose synthase)	√	√	√	X	√	√	√
Sucrose degradation III (sucrose invertase)	X	X	X	X	√	√	√
Xylose degradation I	X	√	√	√	√	√	√

R_bro: *R. bromii* L2-63; R_alb: *R. albus* 7; R_cha: *R. champanellensis* 18P13; R_sp. 80/3: *Ruminococcus* sp. 80/3; R_tor: *R. torques* L2-14; R_obe: *R. obeum* A2-162; R_sp. SR1/5: *Ruminococcus* sp. SR1/5; √: Present; X: Absent

sp. 80/3. *R. albus* 7 was noted to possess lactose and galactose degradation I pathway, pectin degradation III pathway and homogalacturonan degradation pathway that were not only absent in *H. sapiens* but also in other members of the genus *Ruminococcus*. Chitin degradation II pathway was found to occur only in *R. bromii* L2-63 and *Ruminococcus* sp. SR1/5. Thus, it was perceptible that members of the genus possess specific carbohydrate degradation pathways that account for

Table 4.27 Frequencies of various CAZyme families in *Ruminococcus*

Organism	GH	GT	PL	CE	CBM
R. bromii L2-63	21	17	-	2	11
R. albus 7	98	24	7	18	131
R. champanellensis 18P13	54	12	9	12	35
Ruminococcus sp. 80/3	55	16	-	6	15
R. torques L2-14	42	25	-	4	4
R. obeum A2-162	33	26	-	3	5
Ruminococcus sp. SR1/5	45	26	-	3	4

GH: Glycoside hydrolases; GT: Glycosyltransferases; PL: Polysaccharide lyases; CE: Carbohydrate esterases; CBM: Carbohydrate-binding modules; -: Not present

their cellulolytic (Christopherson *et al.*, 2014) and amylolytic (Ze *et al.*, 2012) behaviour and facilitate proper breakdown of complex carbohydrates and starch in human intestine (Bhattacharya *et al.*, 2015).

4.2.1.7 Carbohydrate-Active enZymes (CAZymes) of Ruminococcus:

Carbohydrate-Active enZymes (CAZymes) are associated with the breakdown undigested carbohydrate components and are inevitably required for digesting complex dietary polysaccharides. CAZymes of the gut microbes are of prime importance for human host as the human genome codes for very few carbohydrate degrading enzymes and thus, rely heavily on the gut microbial pool for proper digestion. Members of the genus *Ruminococcus* were found to be enriched with Carbohydrate-Active enZymes (CAZymes) serving as key players in degradation of complex dietary components in intestinal environment of human host. Families representing the major class of modules- Glycoside hydrolases (GH), Glycosyltransferases (GT), Polysaccharide lyases (PL), Carbohydrate esterases (CE) and additional Carbohydrate-binding modules (CBM) were abundant among the ruminococcal members. Glycoside hydrolases hydrolyze the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety, such as a protein, or a lipid (Cantarel *et al.*, 2009). Glycosyltransferases catalyze the transfer of sugar moieties from activated donor to specific glycosyl acceptor molecules and establish natural glycosidic linkages (Williams and Thorson, 2009). Transfer of

Table 4.28 List and function of CAZyme families constituting core CAZyme set in *Ruminococcus*

Core CAZyme Families	Functions
GH3	β-glucosidase
GH13	α-amylase
GH31	α-glucosidase
GT2	Cellulose synthase
GT4	Sucrose synthase
GT28	1,2-diacylglycerol 3-β-galactosyltransferase
GT35	Glycogen or starch phosphorylase
CE4	Acetyl xylan esterase
CBM48	Glycogen-binding function

Information pertaining to function has been taken from CAZy database

glycosyl groups result in the biosynthesis of several disaccharides, oligosaccharides and polysaccharides (Lairson *et al.*, 2008). Polysaccharide lyases have been reported to cleave uronic acid-containing polysaccharide chains and generate unsaturated hexenuronic acid residue (Lombard *et al.*, 2010). Carbohydrate esterases are the enzymes that catalyze the de-O or de-N-acylation of esters or amides and other substituted saccharides (Biely, 2012). Carbohydrate-binding modules, associated with CAZymes, are non-catalytic domains that exhibit carbohydrate-binding activities. The frequencies of respective CAZyme classes have been provided in Table 4.27. GH class was found to be most prevalent among the ruminococcal members.

Apart from estimating the frequencies, we also filtered the sets of core CAZyme families, present universally in all strains of *Ruminococcus*. The complete list of the CAZyme families constituting the core set and their functions has been provided in Table 4.28. It was noticeable from Table 4.28 that families GH3, GH13, GH31, GT2, GT4, GT28, GT35, CE4 and CBM48 constituted the core CAZyme set.

Ruminococcus employs sophisticated CAZyme machinery to establish a strong symbiotic platform with human host. Repertoire of diverse sets of enzymes, efficient in fermentation of undigested carbohydrates and fibers in the human gut, ameliorates intestinal environment (Flint *et al.*, 2012). A large stock of CAZymes among the ruminococcal members allow them to access different energy sources and provide them additional competitive

advantage over other members of the human gut microbial community (Ben David *et al.*, 2015; White *et al.*, 2014). Besides conferring adaptive benefits, CAZymes also initiate degradation of complex carbohydrates like dietary fibers and resistant starch in certain gut microbial members (Flint *et al.*, 2012). *R. bromii* L2-63 is an established 'keystone' starch degrader which initiates breakdown of recalcitrant starch (Ze *et al.*, 2012) and paves way for efficient degradation by other bacterial members like *Eubacterium rectale* and *Bacteroides thetaiotaomicron* (Ze *et al.*, 2012; Ze *et al.*, 2013). Human genome encodes only few carbohydrate degrading enzymes and thus, depends on the CAZyme pool of gut microbes for proper digestion of the dietary components (Bhattacharya *et al.*, 2015). CAZymes serve as indispensable tools which gut microbes employ for proficient energy harvest in human host system.

4.2.2 Comparative genomics of Bacteroides:

4.2.2.1 Genome composition of Bacteroides:

A consistent trend of AT preference was observed among the members of the genus *Bacteroides*. The average genome size of the genus was found to be around 5.21±0.77 millions of base pairs (Mbp) and the average AT composition was found to be around 56.67±1.71%. Genome composition (in terms of AT percentage), genome size and frequency of coding sequences (CDS count), for all concerned strains of *Bacteroides*, were plotted in three-dimensional hyperspace and has been represented in the form of a three dimensional plot (3D plot) (Figure 4.32). There was considerable variation in genome size among the concerned bacterial members. *Bacteroides thetaiotaomicron* VPI-5482 was found to display the highest genome size containing 4,938 coding sequences. *Bacteroides helcogenes* P 36-108 was found to be the smallest in terms of genome size containing 3,353 coding sequences. It was evident form Figure 4.32 and Table 4.29 that there was a moderate bias towards AT composition among the genomes of interest.

4.2.2.2 Gene families representing pan - and core genome in Bacteroides:

Members of the genus *Bacteroides* i.e., *Bacteroides xylanisolvens* XB1A, *Bacteroides fragilis* 638R, *Bacteroides fragilis* NCTC 9343, *Bacteroides fragilis* YCH46, *Bacteroides thetaiotaomicron* VPI-5482,

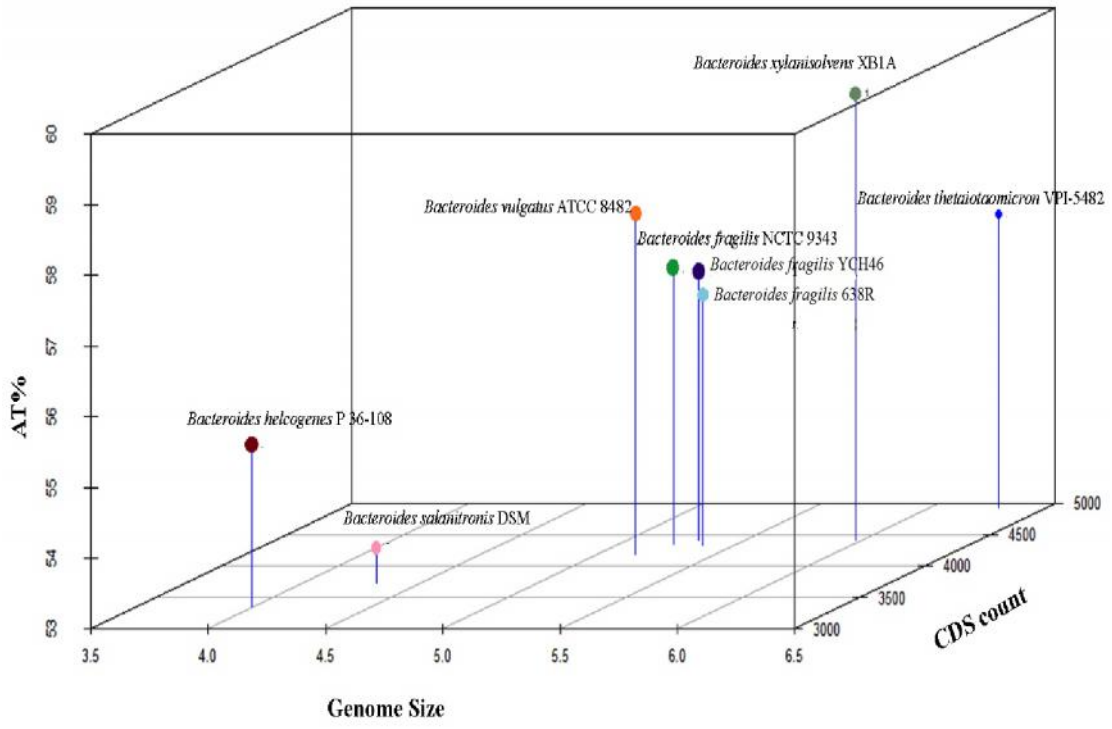

Figure 4.32: 3D plot of *Bacteroides*– 3D plot represents genomic composition (in terms of percentage of Adenine and Thymine [AT%]), genome size and frequency of coding sequences (CDS count) of the respective members of *Bacteroides* in three dimensional hyperspace

Bacteroides vulgatus ATCC 8482, *Bacteroides helcogenes* P 36-108 and *Bacteroides salanitronis* DSM 18170 displayed a set of 12,601 gene families in the pan-genome and 1,155 gene families in the core genome (Table 4.29). Significant changes were observed in the pan-genome and core genome curves of the Pan-Core genome plot (Figure 4.33) with subsequent addition of genomes. The pan-genome was found to increase in size whereas, the core genome reduced in size on addition of new genomes. Frequencies of unique gene families of all concerned strains of *Bacteroides* and common gene families shared between them were also estimated and the respective frequencies have been displayed in the form of a petal plot (Figure 4.34).

4.2.2.3 Multi-locus sequence analysis (MLSA) based phylogeny and BLAST matrix of Bacteroides:

Multi-locus sequence analysis (MLSA) based phylogenetic tree was constructed employing selected housekeeping genes of the concerned strains of *Bacteroides*. It was evident from the phylogenetic tree (Figure 4.35) that the concerned strains separated into two distinct clades - one clade representing the members like *B. fragilis* 638R, *B. fragilis* NCTC 9343,

Table 4.29 Comparative genomic statistics of various members of *Bacteroides*

Organism	Genome Size (Mbp)	CDS Count	AT%	Unique gene families	Core genome Size	Pan-genome size
B. xylanisolvens XB1A	5.98	4407	59.33	444	1155	12601
B. fragilis 638R	5.37	4326	56.58	403		
B. fragilis NCTC 9343	5.24	4342	56.89	383		
B. fragilis YCH46	5.31	4407	56.8	981		
B. thetaiotaomicron VPI-5482	6.29	4938	57.14	1316		
B. vulgatus ATCC 8482	5.16	4192	57.8	1278		
B. helcogenes P 36-108	3.99	3353	55.28	1287		
B. salanitronis DSM 18170	4.31	3737	53.51	1270		

CDS: Coding sequences; Mbp: millions of base pairs

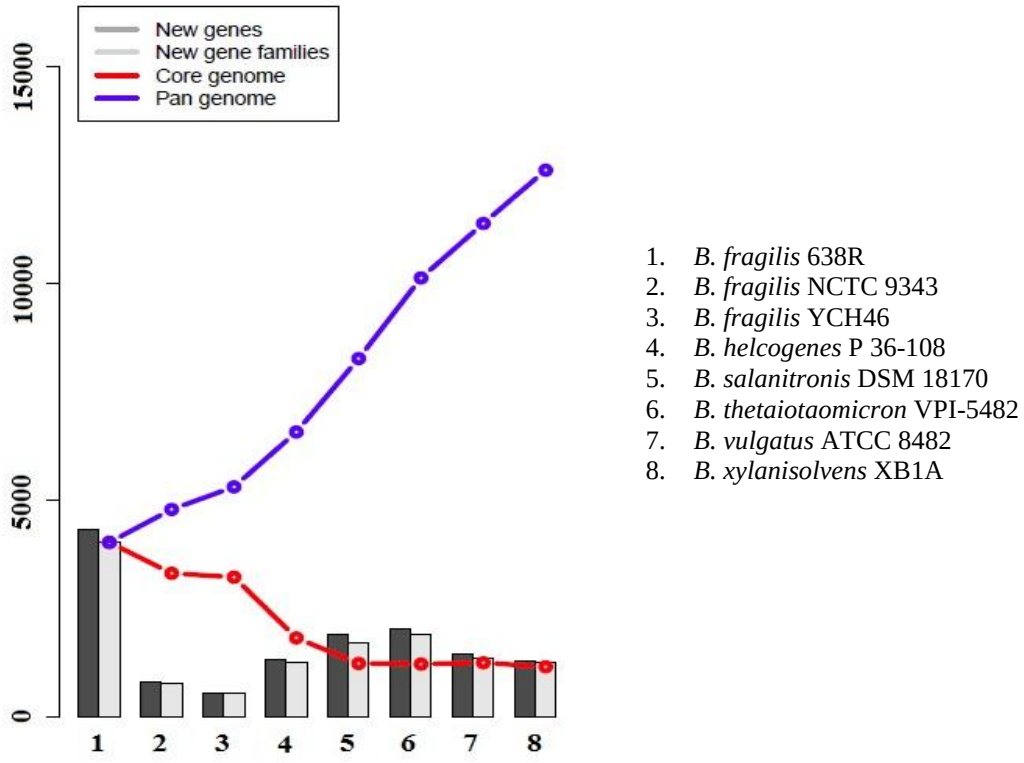

Figure 4.33: Pan-Core genome plot for *Bacteroides*- The pan-genome (blue line) and core genome (red line) plot for *Bacteroides* genomes. The number of discovered novel genes (dark bars) and novel gene families (light-grey bars) are also shown for each added genome

B. fragilis YCH46, *B. thetaiotaomicron* VPI-5482, *B. xylanisolvens* XB1A and *B. helcogenes* P 36-108 and the other clade reflecting the members like *B. vulgatus* ATCC 8482 and *B. salanitronis* DSM 18170. MLSA based phylogenetic tree was found to yield similar results with 16s rRNA based tree generated by Karlsson and colleagues (Karlsson *et al.*, 2011)

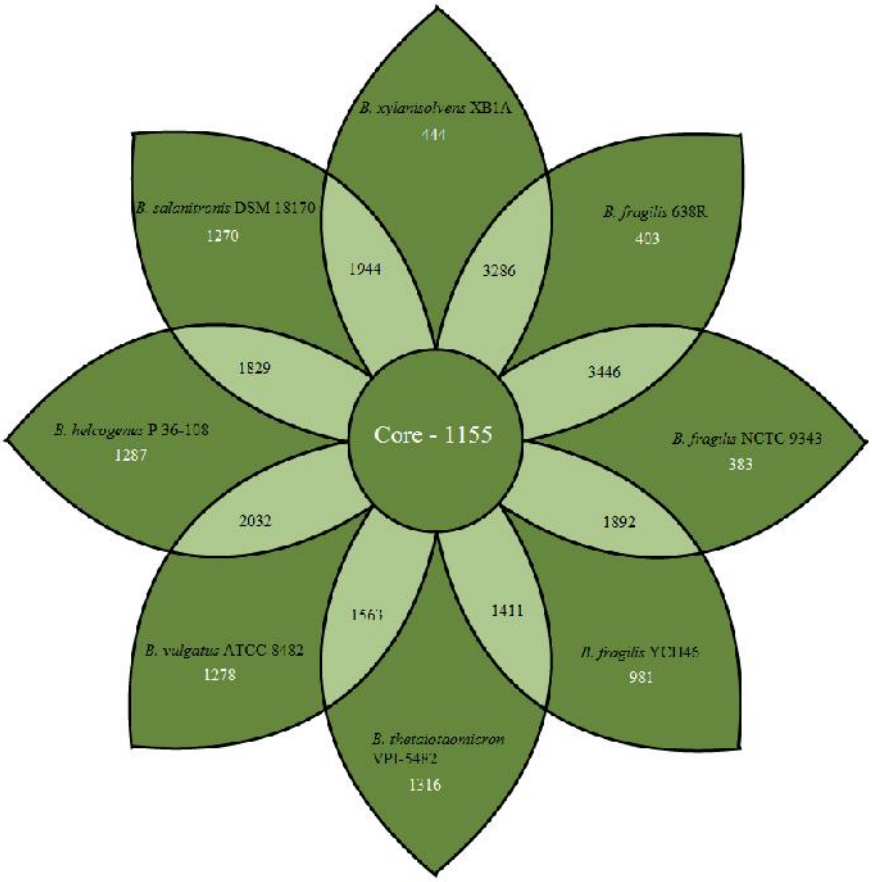

Figure 4.34: Petal plot of *Bacteroides*- Petal plot of *Bacteroides* displaying frequencies of core gene families and unique gene families

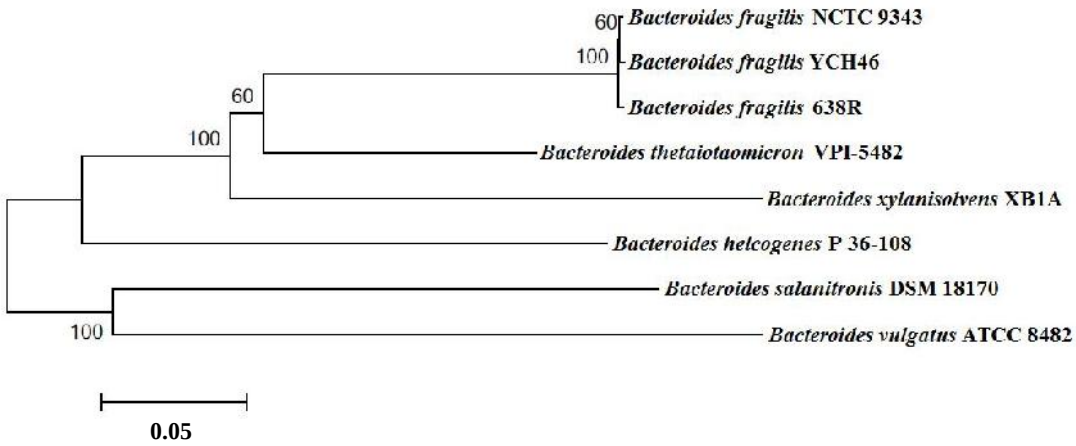

Figure 4.35: MLSA based phylogenetic tree of *Bacteroides*

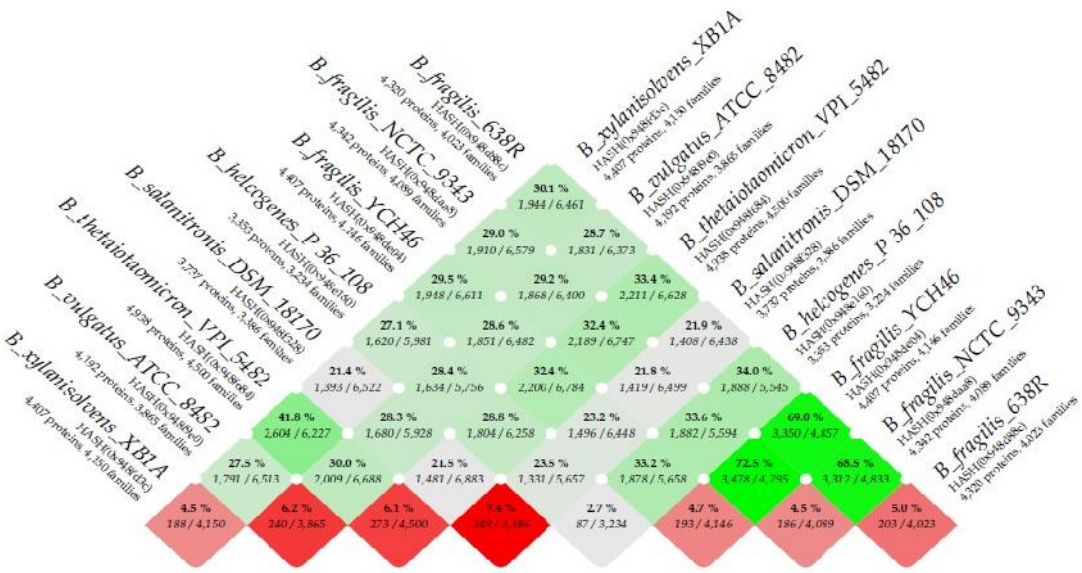

Figure 4.36: BLAST Matrix of *Bacteroides*– To the side, the total number of protein genes and gene families are listed for each genome. In the matrix cells, the shared protein genes are given as a percentage, based on the ratio of the core genome and pan-genome of each pair; Green color represents homology between proteomes; Red color represents homology within proteomes

where *B. fragilis* and *B. vulgatus* were found to lie in different phylogenetic branches. BLAST matrix of the concerned proteomes of *Bacteroides* was generated in order to estimate the degrees of similarity between them (Figure 4.36). The percentage of similarity between proteomes, for each combination, has been presented by green color and the percentage of similarity within the same proteome has been depicted by red color in Figure 4.36. It was distinctly perceptible from Figure 4.36 that the strains representing same species had the highest degrees of similarity between them i.e., strains reflecting *B. fragilis* species showed high proteomic resemblance. *B. thetaiotaomicron* VPI-5482 and *B. xylanisolvens* XB1A were found to display high similarity at the proteomic level. Such an observation seemed well justified from the fact that both the strains were found to group in the same clade in MLSA based phylogenetic tree (Figure 4.35).

4.2.2.4 COG category estimation in Bacteroides:

Estimation of COG categories of the concerned members of *Bacteroides* revealed that majority of the proteins were found to represent '*Metabolism*' COG category (Figure 4.37). Significant share of proteins also reflected '*Information Storage and Processing*' and '*Cellular Processes and Signaling*' COG categories (Figure 4.37). Some proteins were found to be

Organism	COG categories (Frequencies in %)			
	ISP	*CPS*	*Met*	*PC*
B. xylanisolvens XB1A	10.37	11.13	21.00	6.13
B. fragilis 638R	10.52	15.31	21.24	6.09
B. fragilis NCTC 9343	10.40	15.00	21.33	6.13
B. fragilis YCH46	10.33	14.77	21.17	6.03
B. thetaiotaomicron VPI-5482	10.16	13.45	21.34	5.92
B. vulgatus ATCC 8482	10.30	14.30	22.60	5.68
B. helcogenes P 36-108	11.57	14.53	23.14	5.98
B. salanitronis DSM 18170	10.81	11.30	19.23	5.93

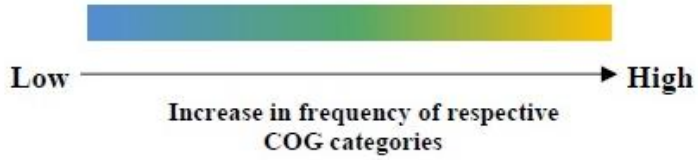

Figure 4.37: COG category heatmap of *Bacteroides*- Heatmap based on the frequencies (in percentage) of various COG categories of the proteomes of *Bacteroides*. In the heatmap, *ISP* refers to *Information Storage and Processing*, *CPS* refers to *Cellular Processes and Signaling*, *Met* refers to *Metabolism*, *PC* refers to *Poorly Characterized* COG categories. Change of color intensity from blue to bright yellow refers to increase in frequency of a particular COG category (color scale provided below the heatmap)

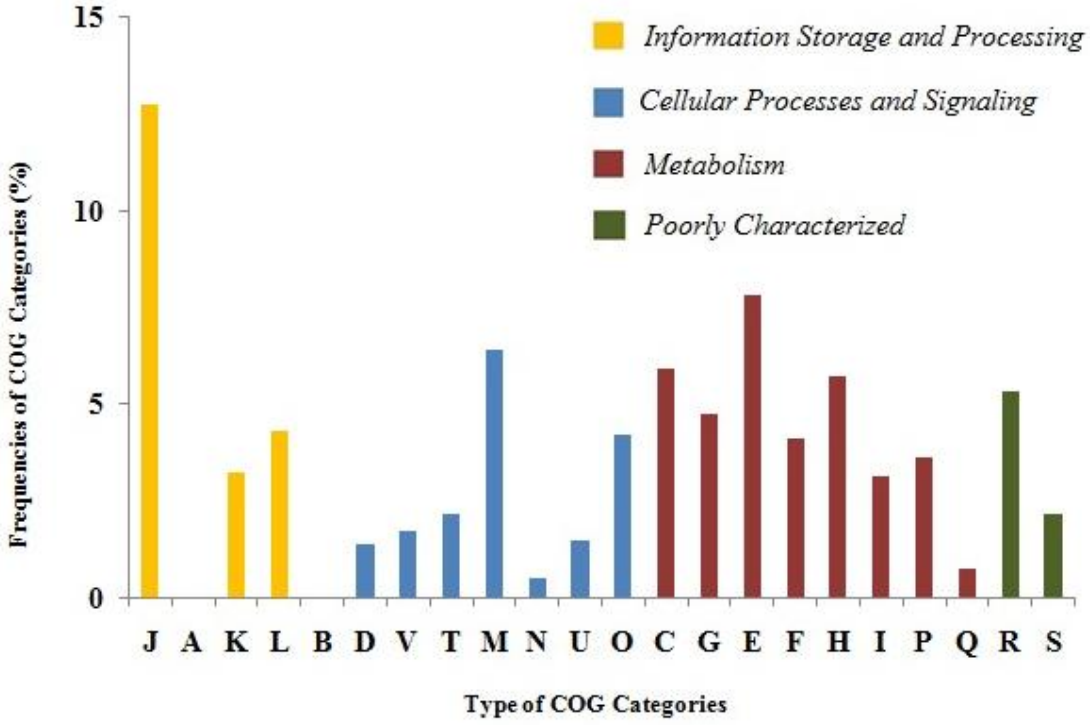

Figure 4.38: Bar diagram displaying COG categories of core proteome of *Bacteroides*

Table 4.30 List of carbohydrate degradation pathways in present in *Bacteroides* but absent in *H. sapiens*

	Organism							
Carbohydrate Degradation Pathway	Bac 1	Bac 2	Bac 3	Bac 4	Bac 5	Bac 6	Bac 7	Bac 8
2-O-α-mannosyl-D-glycerate degradation	√	√	X	√	√	X	X	X
N-acetylneuraminate and N-acetylmannosamine degradation	√	√	√	√	X	√	√	X
Chitin degradation II	√	√	√	√	√	√	√	√
Chondroitin sulfate and dermatan sulfate degradation I (bacterial)	√	√	√	√	√	X	√	X
Fructose degradation	X	√	√	√	X	X	X	X
Fucose degradation	√	√	√	√	√	√	√	√
Galactose degradation I (Leloir pathway)	√	X	X	√	√	√	X	X
Gellan degradation	√	X	X	X	X	X	X	X
Glucose and glucose-1-phosphate degradation	X	X	X	X	X	X	√	√
Glycogen degradation I	√	√	√	√	√	√	√	√
Homogalacturonan degradation	√	X	X	X	√	√	X	√
Kojibiose degradation	X	X	X	X	X	X	X	√
L-arabinose degradation I	√	X	X	X	√	√	X	√
L-rhamnose degradation I	√	X	X	X	√	√	X	√
Mannitol degradation II	X	X	X	X	X	X	X	√
Melibiose degradation	√	√	√	√	√	√	√	√
Pectin degradation II	X	X	X	X	√	X	X	√
Pectin degradation III	√	X	X	X	√	√	X	√
Starch degradation V	X	X	X	√	√	X	√	√
Sucrose degradation II (sucrose synthase)	√	√	√	√	√	√	√	√
Xylose degradation I	√	√	√	√	√	√	√	√

Bac 1: *B. xylanisolvens* XB1A; Bac 2: *B. fragilis* 638R; Bac 3: *B. fragilis* NCTC 9343; Bac 4: *B. fragilis* YCH46; Bac 5: *B. thetaiotaomicron* VPI-5482; Bac 6: *B. vulgatus* ATCC 8482; Bac 7: *B. helcogenes* P 36-108; Bac 8: *B. salanitronis* DSM 18170; √: Present; X: Absent

functionally uncharacterized. It was noticeable from investigation of COG sub-categories that '*Energy production and conversion* [C]', '*Carbohydrate transport and metabolism* [G]', '*Amino acid transport and metabolism* [E]', '*Nucleotide transport and metabolism* [F]', '*Coenzyme transport and metabolism* [H]' and '*Lipid transport and metabolism* [I]' sub-categories representing '*Metabolism'* COG category were dominant in the core proteome set of *Bacteroides*. '*Translation, ribosomal structure and*

biogenesis [J]' sub-category representing the '*Information Storage and Processing*' COG category and '*Cell wall/membrane/envelope biogenesis* [M]' sub-category signifying '*Cellular Processes and Signaling*' COG category were also found in significant proportions in the core proteome of *Bacteroides* (Figure 4.38).

It has been perceptible from COG category analysis that the genus *Bacteroides* possesses a rich collection of proteins associated with vital metabolic activities like carbohydrate, amino acid and nucleotide metabolism. Members of the genus have been reported to be efficient carbohydrate degraders and have been found to aid digestion in human intestinal tract (Martens *et al.*, 2009). Our results emphasized the fact that the genus has been well equipped with rich stock of proteins associated with diverse metabolic functions (Flint *et al.*, 2012). Presence of such metabolically efficient proteins confers them an advantage to properly adjust in human intestinal environment.

4.2.2.5 Carbohydrate degradation pathways of Bacteroides:

Comparative profiling of carbohydrate degradation pathways in the genus *Bacteroides* and human (*Homo sapiens*) host was performed in order to gain a clear perception as how various members of *Bacteroides* facilitate proper carbohydrate digestion in human intestine. Interestingly, many vital pathways associated with carbohydrate degradation were found to be absent in *H. sapiens*. The complete list of carbohydrate degradation pathways present in concerned members of *Bacteroides* but absent in *H. sapiens* has been provided in Table 4.30. Chitin degradation II pathway, fucose degradation pathway, glycogen degradation I pathway, melibiose degradation pathway, sucrose degradation II (sucrose synthase) pathway and xylose degradation I pathway, absent in *H. sapiens*, were found to present universally among all the members of *Bacteroides*. Members of the species *B. fragilis* were found to possess fructose degradation pathway which was not only absent in *H. sapiens* but also in other members of *Bacteroides*. Mannitol degradation II pathway was found to be occurring specifically in *B. salanitronis* DSM 18170 whereas, gellan degradation pathway was noted to be present only in *B. xylanisolvens* XB1A.

Table 4.31 Frequencies of various CAZyme families in *Bacteroides*

Organism	GH	GT	PL	CE	CBM
B. xylanisolvens XB1A	232	75	21	19	20
B. fragilis 638R	142	78	1	6	15
B. fragilis NCTC 9343	136	83	2	6	17
B. fragilis YCH46	144	83	1	5	17
B. thetaiotaomicron VPI-5482	271	91	18	20	31
B. vulgatus ATCC 8482	187	82	8	19	18
B. helcogenes P 36-108	100	70	6	8	11
B. salanitronis DSM 18170	126	44	8	17	9

GH: Glycoside hydrolases; GT: Glycosyltransferases; PL: Polysaccharide lyases; CE: Carbohydrate esterases; CBM: Carbohydrate-binding modules; -: Not present

It was evident from our analysis that various species of *Bacteroides*, that reside in human gut, possess many crucial carbohydrate degradation pathways that are absent in human host. These bacterial members have been associated with breakdown of complex dietary carbohydrate components which human beings are incapable of degradation by their own metabolic apparatus (Mirande *et al.*, 2010; El Kaoutari *et al.*, 2013).

4.2.2.6 Carbohydrate-Active enZymes (CAZymes) of Bacteroides:

Carbohydrate-Active enZymes (CAZymes) refer to special enzyme sets of human gut microflora that break down complex carbohydrate components of human diet into simpler forms and aid digestion. Major modules of CAZymes mainly include Glycoside hydrolases (GH), Glycosyltransferases (GT), Polysaccharide lyases (PL), Carbohydrate esterases (CE) and Carbohydrate-binding modules (CBM). Estimation of CAZyme families in *Bacteroides* revealed abundance of these vital enzyme components in the concerned genus. All strains of *Bacteroides* were found to be endowed with the major CAZyme modules. Respective frequencies of the major modules have been provided in Table 4.31. Glycoside hydrolases are the enzymes associated with hydrolysis of glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety (Cantarel *et al.*, 2009). Glycosyltransferases have been reported to be associated with transfer of sugar moieties from activated donor to specific glycosyl acceptor molecules

Table 4.32 List and function of CAZyme families constituting core CAZyme set in *Bacteroides*

Core CAZyme Families	Functions
GH2	β-galactosidase
GH3	β-glucosidase
GH13	α-amylase
GH16	Xyloglucosyltransferase
GH18	Chitinase
GH23	Lysozyme type G
GH25	Lysozyme
GH27	α-galactosidase
GH29	α-L-fucosidase
GH31	α-glucosidase
GH32	Invertase
GH33	Sialidase, Neuraminidase
GH35	β-galactosidase
GH36	α-galactosidase
GH43	β-xylosidase
GH57	α-amylase
GH73	Lysozyme
GH77	Amylomaltase
GH88	d-4,5-unsaturated β-glucuronyl hydrolase
GH89	α-N-acetylglucosaminidase
GH92	Mannosyl-oligosaccharide α-1,2-mannosidase
GH95	α-L-fucosidase
GH97	Glucoamylase
GH109	α-N-acetylgalactosaminidase
GH130	β-1,4-mannosylglucose phosphorylase
GT2	Cellulose synthase
GT3	Glycogen synthase
GT4	Sucrose synthase
GT5	UDP-Glc: glycogen glucosyltransferase
GT19	Lipid-A-disaccharide synthase
GT28	1,2-diacylglycerol 3-β-galactosyltransferase
GT35	Glycogen or starch phosphorylase
GT51	Murein polymerase
CE1	Acetyl xylan esterase
CE11	UDP-3-0-acyl N-acetylglucosamine deacetylase
CBM20	Starch-binding domains
CBM32	Binding to galactose and lactose
CBM50	Cleave chitin or peptidoglycan

Information pertaining to function has been taken from CAZy database

(Williams and Thorson, 2009). Polysaccharide lyases catalyze the cleavage of uronic acid-containing polysaccharide chains to generate unsaturated hexenuronic acid residue (Lombard *et al.*, 2010). Carbohydrate esterases facilitate de-O or de-N-acylation of esters or amides and other substituted saccharides (Biely, 2012). Carbohydrate-binding modules (CBM) have been reported to be appended with CAZymes that display carbohydrate-binding activities (Tomme *et al.*, 1988). GH class was found to be most abundant among the members of *Bacteroides*. We also constituted a core CAZyme set which comprised of the CAZyme families

present globally in all concerned strains of *Bacteroides*. The core set of CAZymes has been enlisted in Table 4.32.

Bacteroides is an important microbial genus that resides in human gut and confers several beneficial effects associated with human intestinal homeostasis (Martens *et al.*, 2009). Members of the genus have been noted to effectively utilize diverse range of complex polysaccharides (Flint *et al.*, 2012). *B. thetaiotaomicron* has been reported to encode plenty of Glycoside hydrolases and has been found to degrade glycan proficiently (Anderson and Salyers, 1989; Ravcheev *et al.*, 2013). Our results were in complete agreement with the fact that the genus *Bacteroides* possesses a rich repository of varied CAZyme families that bestow an adaptive advantage to the concerned bacterial genus and simultaneously, enhance the digestive abilities of human host.

4.2.3 Comparative genomics of Eubacterium:

4.2.3.1 Genomic constitution of Eubacterium:

Genomic architecture of the genus *Eubacterium* was found to be AT rich. The average genome size of the genus was found to be around 3.13±0.38 millions of base pairs (Mbp) and the average AT composition was found to be around 59.01±2.89%. A three-dimensional plot (3D plot), representing genomic composition (in terms of AT percentage), genome size and number of coding sequences (CDS count) in three dimensional coordinates, was generated considering all members of *Eubacterium* under analysis (Figure 4.39). It was noticeable that AT compositional bias was predominant in all concerned bacterial strains. *Eubacterium eligens* ATCC 27750 was found to display highest AT composition (around 62%) among the concerned members. There was considerable variation in genome size and frequency of coding sequences among the members of the genus. *Eubacterium siraeum* V10Sc8a was found to be comparatively smaller than the other members of *Eubacterium*, in terms of genome size and was found to contain 2,211 coding sequences (Figure 4.39 and Table 4.33). *Eubacterium rectale* M104/1 was observed to bear a highest number of 3,212 coding sequences (Figure 4.39 and Table 4.33).

4.2.3.2 Pan-genome and core genome analysis:

Members of the genus *Eubacterium*

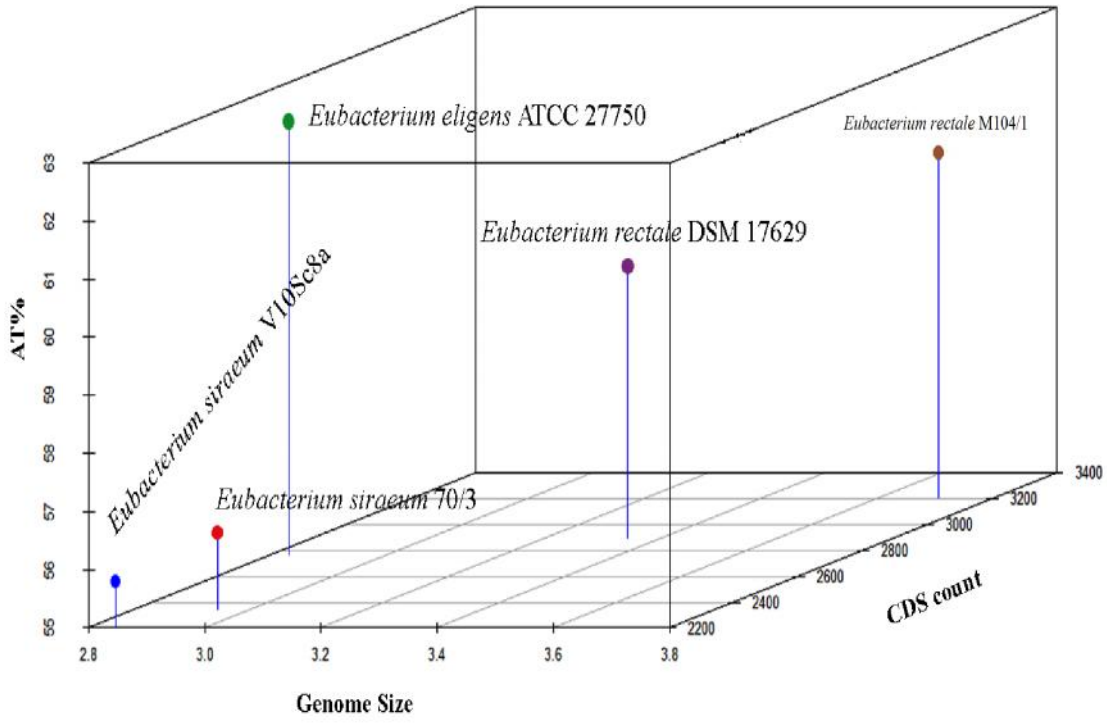

Figure 4.39: 3D plot of *Eubacterium* – 3D plot represents genomic composition (in terms of percentage of Adenine and Thymine [AT%]), genome size and frequency of coding sequences (CDS count) of the respective members of *Eubacterium* in three dimensional hyperspace

i.e., *Eubacterium eligens* ATCC 27750, *Eubacterium rectale* DSM 17629, *Eubacterium rectale* M104/1, *Eubacterium siraeum* 70/3 and *Eubacterium siraeum* V10Sc8a produced a set of 7,802 gene families reflecting the pan-genome and 271 gene families representing the core genome (Table 4.33). As genomes were subsequently added for analysis, the Pan-Core genome plot (Figure 4.40) displayed significant changes in both the pan-genome and core genome curves. There was considerable jump in both the curves. There was significant increase in the pan-genome size and considerable decrease in the size of the core genome with the addition of new members of *Eubacterium*. We also estimated the frequencies of unique and common gene families between the concerned strains of *Eubacterium* and represented them in the form of a petal plot (Figure 4.41). Strains representing same species were found to display a higher share of common gene families between them. *E. rectale* DSM 17629 and *E. rectale* M104/1 shared 2,172 common gene families. Members representing *E. siraeum* species shared 1,641 common families of genes. The frequencies of intersecting gene families were found to reduce in case of an interspecific comparison. *E.*

Table 4.33 Comparative genomic statistics of various members of *Eubacterium*

Organism	Genome Size (Mbp)	CDS Count	AT%	Unique gene families	Core genome Size	Pan-genome size
E. eligens ATCC 27750	2.83	2765	62.43	1690	271	7802
E. rectale DSM 17629	3.34	2898	59.65	477		
E. rectale M104/1	3.7	3212	60.9	867		
E. siraeum 70/3	2.94	2347	56.3	600		
E. siraeum V10Sc8a	2.84	2211	55.77	501		

CDS: Coding sequences; Mbp: millions of base pairs

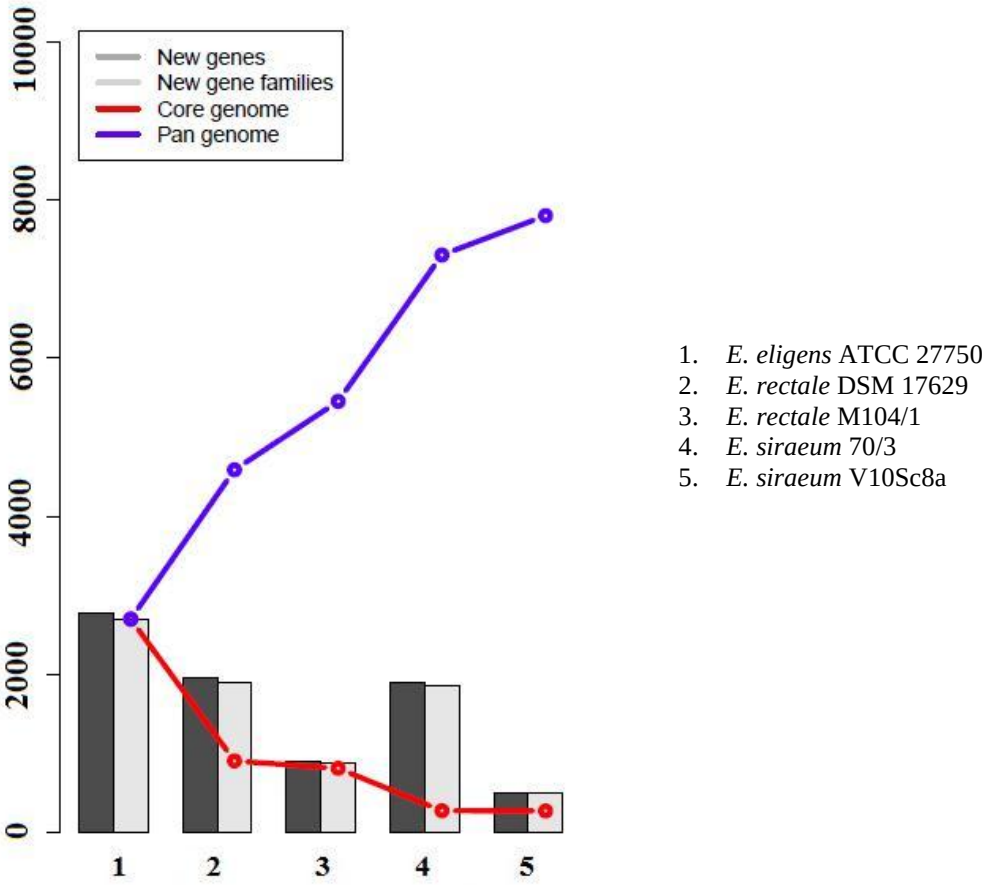

Figure 4.40: Pan-Core genome plot for *Eubacterium*- The pan-genome (blue line) and core genome (red line) plot for *Eubacterium* genomes. The number of discovered novel genes (dark bars) and novel gene families (light-grey bars) are also shown for each added genome

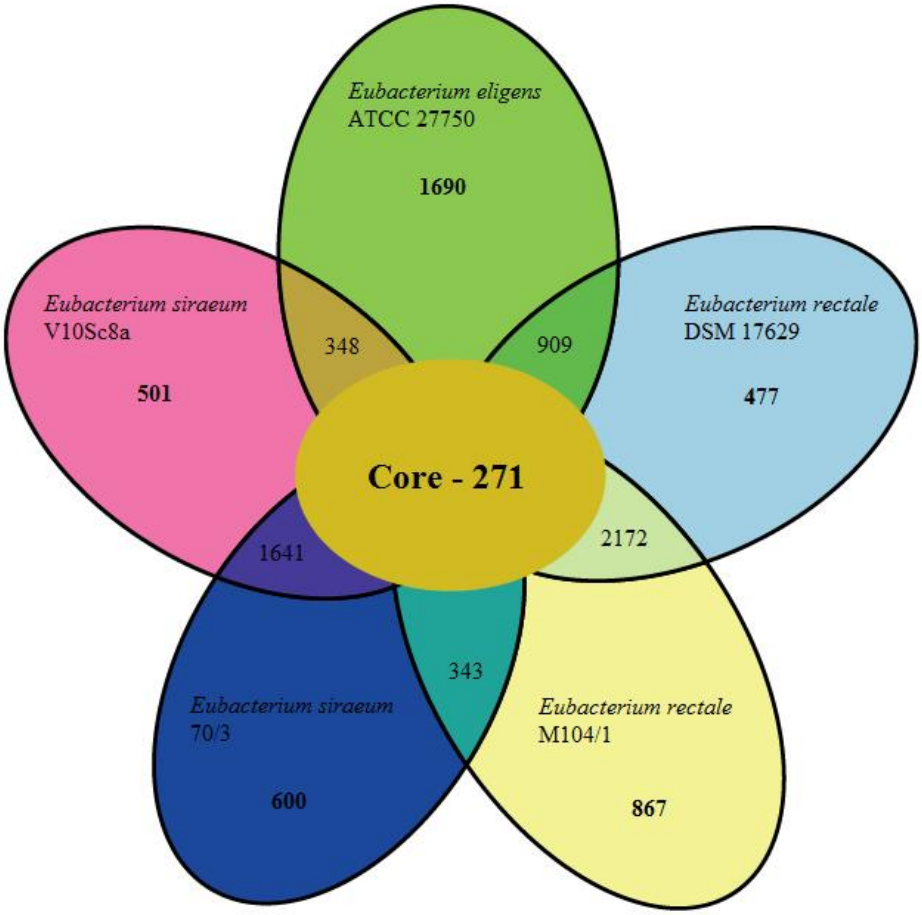

Figure 4.41: Petal plot of *Eubacterium*- Petal plot of *Eubacterium* displaying the frequencies of core gene families and unique gene families

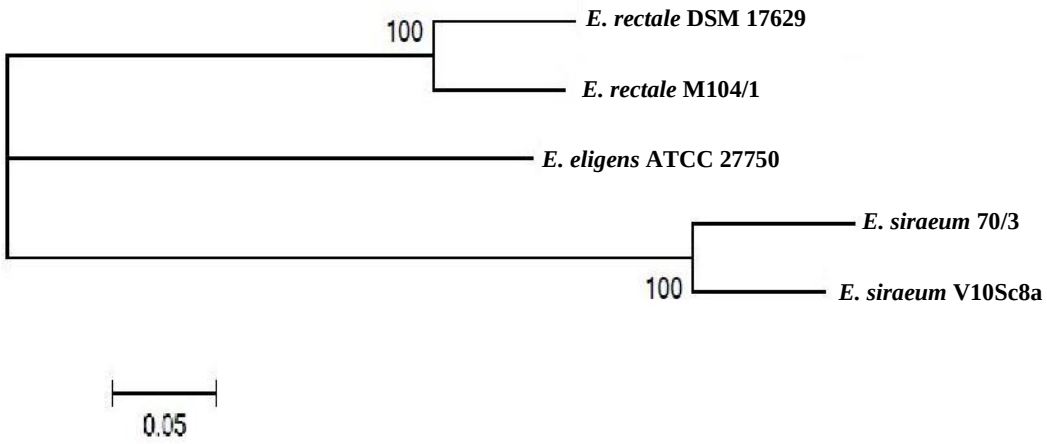

Figure 4.42: MLSA based phylogenetic tree of *Eubacterium*

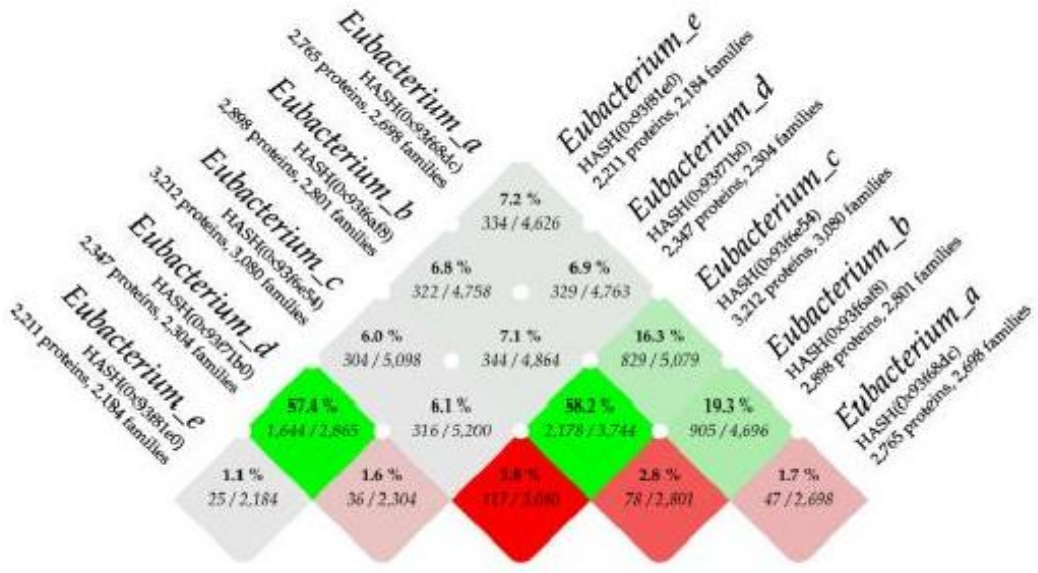

Figure 4.43: BLAST Matrix of *Eubacterium*– To the side, the total number of protein genes and gene families are listed for each genome. In the matrix cells, the shared protein genes are given as a percentage, based on the ratio of the core genome and pan-genome of each pair; Green color represents homology between proteomes; Red color represents homology within proteomes; *Eubacterium_a- E. eligens* ATCC 27750; *Eubacterium_b- E. rectale* DSM 17629; *Eubacterium_c- E. rectale* M104/1; *Eubacterium_d- E. siraeum* 70/3; *Eubacterium_e- E. siraeum* V10Sc8a

eligens ATCC 27750 and *E. rectale* DSM 17629 were found to share 909 gene families. *E. rectale* M104/1 and *E. siraeum* 70/3 had a share of 343 orthologous gene families. There were 348 common gene families between *E. siraeum* V10Sc8a and *E. eligens* ATCC 27750.

4.2.3.3 Multi-locus sequence analysis (MLSA) based phylogeny and BLAST matrix of Eubacterium:

Multi-locus sequence analysis (MLSA) based phylogenetic tree was generated using concatenated multiple housekeeping gene sequences of the concerned members of *Eubacterium*. It was prominent that strains representing a particular species grouped together (Figure 4.42). The phylogenetic tree resulted in two distinct clades- one comprising the members of *E. rectale* species and the other included members of the species *E. siraeum. E. eligens* ATCC 27750 was found to group with the members of *E. rectale* species.

BLAST matrix (Figure 4.43), reflecting the similarity between the concerned proteomes, supported our observations from the phylogenetic tree. Members of the same species displayed higher proteomic resemblance among them. *E.*

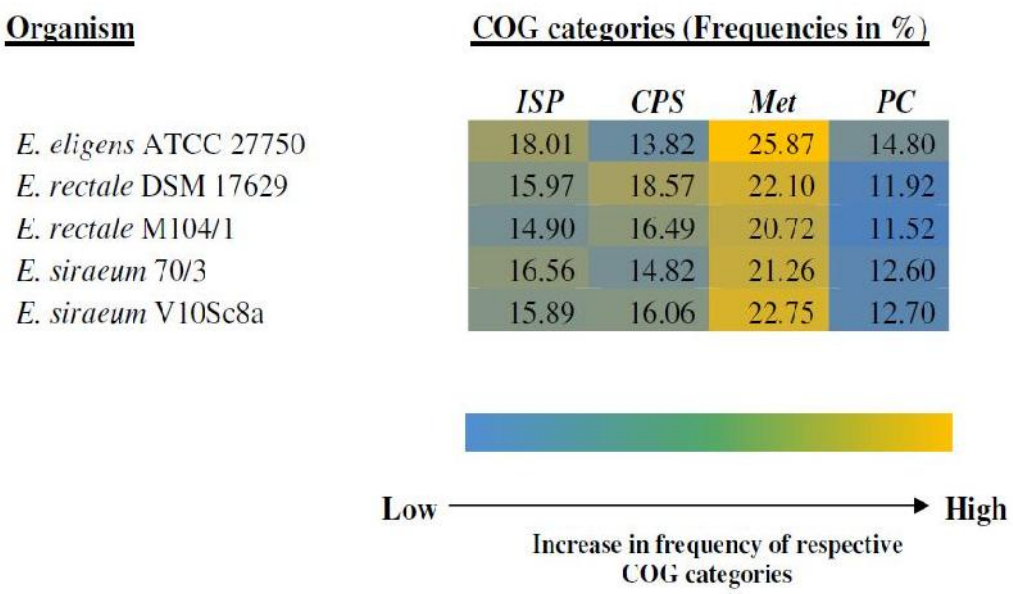

Figure 4.44: COG category heatmap of *Eubacterium*- Heatmap based on the frequencies (in percentage) of various COG categories of proteomes of *Eubacterium*. In the heatmap, *ISP* refers to *Information Storage and Processing*, *CPS* refers to *Cellular Processes and Signaling*, *Met* refers to *Metabolism*, *PC* refers to *Poorly Characterized* COG categories. Change of color intensity from blue to bright yellow refers to increase in frequency of a particular COG category (color scale provided below the heatmap)

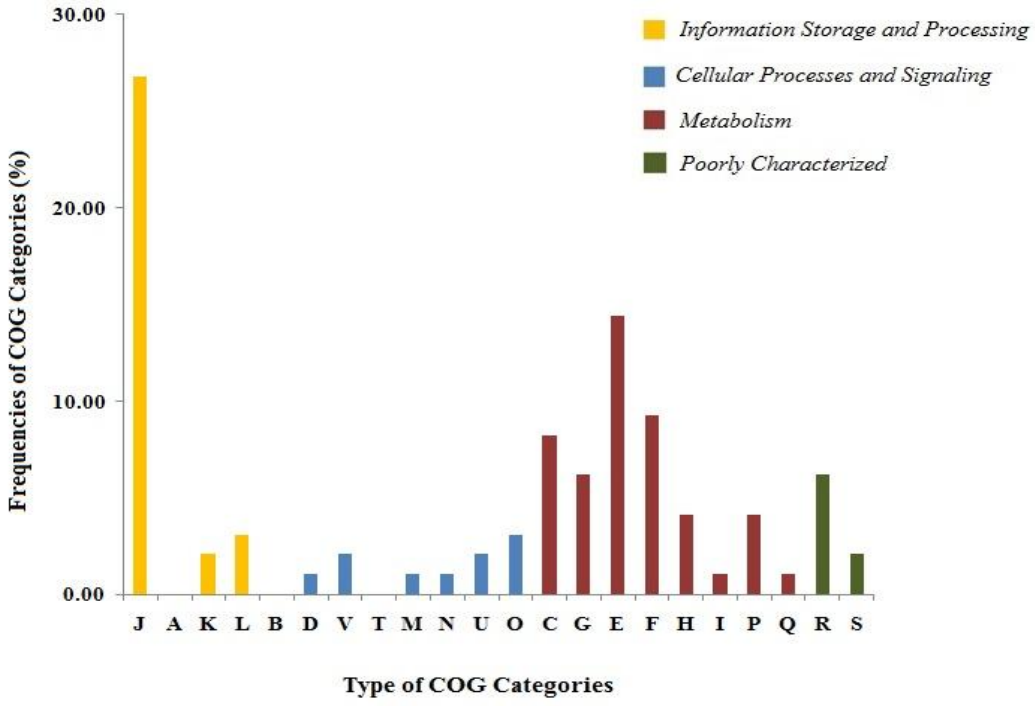

Figure 4.45: Bar diagram displaying COG categories of core proteome of *Eubacterium*

Table 4.34 List of carbohydrate degradation pathways present in *Eubacterium* but absent in *H. sapiens*

Carbohydrate Degradation Pathway	Organism				
	E. eligens ATCC 27750	*E. rectale* DSM 17629	*E. rectale* M104/1	*E. siraeum* 70/3	*E. siraeum* V10Sc8a
Fucose degradation	X	√	√	X	√
Galactose degradation I (Leloir pathway)	X	√	√	X	X
Galactose degradation III	X	X	X	√	X
Gellan degradation	√	√	√	√	√
Glucose and glucose-1-phosphate degradation	X	X	X	√	√
Glycogen degradation I	√	X	X	X	X
Glycogenolysis II	√	√	√	X	X
L-arabinose degradation I	X	X	X	√	X
L-rhamnose degradation I	√	√	√	√	√
Lactose and galactose degradation I	√	√	√	√	√
Lactose degradation III	X	X	X	X	√
Melibiose degradation	√	X	X	X	X
Sucrose degradation	√	X	X	X	X
Sucrose degradation I (sucrose phosphotransferase)	X	√	√	√	√
Sucrose degradation II (sucrose synthase)	X	√	√	X	X
Sucrose degradation III (sucrose invertase)	√	√	√	√	√
Sucrose degradation IV (sucrose phosphorylase)	X	√	√	X	X
Trehalose degradation	X	√	√	X	X
Xylose degradation I	√	√	√	√	√

√: Present; X: Absent

rectale DSM 17629 and *E. rectale* M104/1 were found to exhibit a proteomic similarity of around 58% whereas, *E. siraeum* 70/3 and *E. siraeum* V10Sc8a displayed a similarity of around 57%. *E. eligens* ATCC 27750 was found to have more resemblance with the members *of E. rectale* species than *E. siraeum* species. Such an observation bolstered our findings from the MLSA based phylogenetic tree.

4.2.3.4 Assessment of COG categories in Eubacterium:

COG category analysis was executed for all the members of *Eubacterium* and the numerical account has been provided in the form of a heatmap

(Figure 4.44). There was a gradual change of color intensity from blue to bright yellow with increase in frequency of a particular COG category. It was evident on extensive investigation that proteins representing '*Metabolism'* COG category were prevalent among the members. Considerable share of proteins also reflected '*Information Storage and Processing'* and '*Cellular Processes and Signaling*" COG categories. Some proteins were found to be functionally uncharacterized.

Inspection of COG sub-categories revealed some interesting facts. '*Energy production and conversion* [C]', '*Carbohydrate transport and metabolism* [G]', '*Amino acid transport and metabolism* [E]' and '*Nucleotide transport and metabolism* [F]' sub-categories, representing the '*Metabolism*' COG category, were found to display highest frequencies among the core proteomic set (Figure 4.45). '*Translation, ribosomal structure and biogenesis* [J]' sub-category signifying the '*Information Storage and Processing*' COG category was also predominant among the core proteome of *Eubacterium*.

Presence of a large repository of proteins, associated with various metabolic processes, provides a clue towards successful adaptation of the genus *Eubacterium* in human gut. Various species like *E. rectale* and *E. eligens* have been reported to reside in human intestine where the undigested carbohydrates (components of human diet) serve as major nutrients for the concerned bacterial members (El Kaoutari *et al.*, 2013; Salyers *et al.*, 1977). In return, these bacterial communities systematically regulate metabolism and aid proper digestion of human host by efficient breakdown of the carbohydrate components (Gerritsen *et al.*, 2011).

4.2.3.5 Unique carbohydrate degradation pathways of Eubacterium:

In this present approach, we carried out a comparative investigation of the carbohydrate degrading pathways present in concerned strains of *Eubacterium* and related human host (*Homo sapiens*). The complete list of carbohydrate degradation pathways present in various strains of *Eubacterium* but absent in *H. sapiens* has been provided in Table 4.34. Several crucial pathways, absent in *H. sapiens*, were found to be present in various members of the genus *Eubacterium*. Some pathways like gellan degradation pathway, L-

Table 4.35 Frequencies of various CAZyme families in *Eubacterium*

Organism	GH	GT	PL	CE	CBM
E. eligens ATCC 27750	30	31	7	10	7
E. rectale DSM 17629	47	32	-	5	7
E. rectale M104/1	46	28	-	5	10
E. siraeum 70/3	49	16	-	5	9
E. siraeum V10Sc8a	49	15	-	4	9

GH: Glycoside hydrolases; GT: Glycosyltransferases; PL: Polysaccharide lyases; CE: Carbohydrate esterases; CBM: Carbohydrate-binding modules; -: Not present

rhamnose I degradation pathway, lactose and galactose degradation I pathway, sucrose degradation III (sucrose invertase) pathway and xylose degradation I pathway were found to be present universally among the members of *Eubacterium*, but absent in *H. sapiens*. Sucrose degradation II (sucrose synthase) pathway, sucrose degradation IV (sucrose phosphorylase) pathway and trehalose degradation pathway were only found in members of the species *E. rectale*. Interestingly, glucose and glucose-1-phosphate degradation pathway was present only in the strains of *E. siraeum*. *E. eligens* ATCC 27750 was found to possess some unique pathways like glycogen degradation I pathway, melibiose degradation pathway and sucrose degradation pathway that were absent in other members of *Eubacterium* as well as in human host. Thus, it was noticeable from our analysis that human beings lack a number of essential carbohydrate degradation pathways that are present in the members of *Eubacterium* residing in its gut. It has been suggested that human host heavily depends on various members of its gut bacterial community for proper digestion of complex undigested carbohydrate moieties (White *et al.*, 2014). Our results were in complete agreement with the fact that the genus *Eubacterium* possesses sophisticated carbohydrate degradation machinery which aids its sustenance in human intestinal environment and also facilitates proper digestion in human host (Flint *et al.*, 2012).

4.2.3.6 Carbohydrate-Active enZymes (CAZymes) in the genus Eubacterium:

Members of the genus *Eubacterium* were found to contain a large set of carbohydrate-active enzymes (CAZymes) associated with degradation of complex dietary carbohydrate components in the human

Table 4.36 List and function of CAZyme families constituting core CAZyme set in *Eubacterium*

Core CAZyme Families	Functions
GH2	β-galactosidase
GH5	Endo-β-1,4-glucanase / Cellulase
GH13	α-amylase
GH23	Lysozyme type G
GH43	β-xylosidase
GH77	Amylomaltase or 4-α-glucanotransferase
GH94	Cellobiose phosphorylase
GT2	Cellulose synthase
GT4	Sucrose synthase
GT5	Glycogen glucosyltransferase
GT28	1,2-diacylglycerol 3-β-galactosyltransferase
GT35	Glycogen or starch phosphorylase
GT51	Murein polymerase
CE4	Acetyl xylan esterase
CBM48	Glycogen-binding function

Information pertaining to function has been taken from CAZy database

gut. The major modules of CAZymes that were found to be dominant among the members of *Eubacterium* were Glycoside hydrolases (GH), Glycosyltransferases (GT), Polysaccharide lyases (PL), Carbohydrate esterases (CE) and Carbohydrate-binding modules (CBM). The complete tally of CAZymes has been provided in Table 4.35. Glycoside hydrolases (GH) hydrolyze the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety, such as a protein, or a lipid (Cantarel *et al.*, 2009). Glycosyltransferases (GT) catalyze the transfer of sugar moieties from activated donor to specific glycosyl acceptor molecules and form natural glycosidic linkages (Williams and Thorson, 2009). Polysaccharide lyases have been associated with cleavage of uronic acid-containing polysaccharide chains and generate unsaturated hexenuronic acid residue (Lombard *et al.*, 2010). Carbohydrate esterases (CE) are referred to the class of enzymes that catalyze the de-O or de-N-acylation of esters or amides and other substituted saccharides (Biely, 2012). Carbohydrate-binding modules (CBM)

are defined as the protein domains associated with CAZymes that exhibit carbohydrate-binding activities (Tomme *et al.*, 1988). GH and GT classes were found to display the highest frequencies among the strains of *Eubacterium*. We also filtered the set of core families of CAZymes present universally in all strains of *Eubacterium*. CAZyme families like GH2, GH5, GH13, GH23, GH43, GH77, GH94, GT2, GT4, GT5, GT28, GT35, GT51, CE4 and CBM48 were found to compose the set of core CAZymes. The complete list of core CAZyme families and pertaining functions has been provided in Table 4.36.

Bacterial residents of the human gut are rich sources of CAZymes. Human beings, deficient in CAZyme stock, largely rely on the members of its gut flora for proper breakdown of undigested carbohydrate complexes (Bhattacharya *et al.*, 2015). It was distinct from our analysis that the genus *Eubacterium* harbors an enriched array of diverse families of CAZymes associated with prime biological functions. Besides allowing the bacterial members to sustain in the human intestinal niche, these crucial sets of enzymes also facilitate proper human digestion. There have been previous reports regarding the abundance of CAZymes in various human gut-associated strains of *Eubacterium* (Flint *et al.*, 2012; Bhattacharya *et al.*, 2015). Microbes like *E. rectale*, *E. siraeum* and *E. eligens* have been reported to be associated with proper degradation of complex carbohydrates and cellulose in human intestinal environment (Mahowald *et al.*, 2009; Robert and Bernalier-Donadille, 2003). *E. rectale* plays a vital role in breakdown of undigested starch materials (Cockburn *et al.*, 2015). Thus, a large repertoire of CAZymes in the genus *Eubacterium* build the foundation of proper adaptation in human gut and subsequently, establish mutualistic relationship with human host.

4.3 Secretome profiling of gut-associated bacteria:

Secretomes refer to the complete set of secretory proteins associated with cellular cross-talks and communication (Ranganathan and Garg, 2009). Proteins secreted by the members of human gut microbiota establish the basic platform for successful interaction of the bacterial members with human host. Thus, extracellular protein components of gut bacterial

masses pave way for successful stay in human gut and subsequently, refine and improve host intestinal environment. However, proper identification and comprehensive profiling of gut-bacterial secretomes, on a global scale, still remains unaccomplished. Apart from studying the functional implications, evolutionary significance of secretomes, in comparison to the non-secretome sets, also appears necessary to be explored. In this present approach, we have systematically screened the secretory proteins in 64 bacterial strains, representing 48 bacterial species, dwelling in human gut and executed pertaining functional and evolutionary investigations.

4.3.1 Secretome pool of human gut microflora:

Bacterial members of human gut were found to employ a large repertoire of secretomes for proper execution of cellular cross-talks and interaction with the human host. The frequencies of predicted secretomes in all concerned bacterial members have been enlisted in Table 4.37. Such an observation seemed logical from the aspect of efficient communication with the host. Secretomes are crucial sets of proteins involved in vital biological processes that include cell adhesion, cell migration, cell-cell communication, proliferation, cellular morphogenesis, cell differentiation and detoxification of surrounding environment (Tjalsma *et al.*, 2004). It has been reported that extracellular proteins of several probiotic members of the human gut like *Bifidobacterium* and *Lactobacillus* interact with the mucosal cells of human intestinal lining and bestow several beneficial effects on the host (Sanchez *et al.*, 2010). However, secretomes of pathogenic population of the human gut like *Helicobacter pylori* have been reported to be associated inevitably with pathogenic manifestations (Zanotti and Cendron, 2014). Thus, presence of a large pool of secretomes provides an insight towards the interaction strategies that the gut microbes employ to successfully acclimatize in human gut. Interestingly, it was also evident from our analysis that there was significant correlation between respective proteome size and frequency of predicted secretomes in both Gram-negative ($r = 0.77$, $P < 0.01$) and Gram-positive ($r = 0.75$, $P < 0.01$) bacterial members of human intestinal microflora. A distinct pattern was evident that the frequency of

Table 4.37 Comparative details pertaining to secretomes of human gut-associated microbes

Organism	Abbreviations used	Number of Secretomes	Proteome size	PEC of Secretomes	S	PEC of Non-secretomes
Bacteroides xylanisolvens XB1A	bac1	622	4407	24.07	<	24.44
B. fragilis 638R	bac2	705	4326	24.00	<	24.47
B. fragilis NCTC 9343	bac3	692	4342	24.03	<	24.48
B. fragilis YCH46	bac4	691	4407	24.03	<	24.48
B. thetaiotaomicron VPI-5482	bac5	905	4816	24.10	<	24.48
B. vulgatus ATCC 8482	bac6	673	4192	24.20	<	24.53
Bifidobacterium longum infantis 157F-NC	bif1	60	1999	21.72	<	22.82
B. longum infantis JCM 1222	bif2	78	2552	21.92	<	22.96
B. longum longum F8	bif3	52	1682	21.94	<	22.83
B. longum longum JCM 1217	bif4	68	1924	21.39	<	22.78
B. adolescentis 22L	bif5	93	1745	21.65	<	22.89
B. animalis lactis AD011	bif6	59	1528	21.85	<	22.91
B. bifidum PRL2010	bif7	89	1706	21.64	<	22.66
B. breve UCC2003	bif8	91	1854	21.84	<	22.85
B. longum DJO10A	bif9	87	2003	21.82	<	22.82
B. longum NCC2705	bif10	82	1830	21.79	<	22.79
Butyrivibrio fibrisolvens 16/4	but1	188	2904	22.57	<	24.04
Clostridiales sp. SM4/1	clos1	119	3010	21.95	<	23.63
Clostridiales sp. SS3/4	clos2	165	2993	22.47	<	23.85
Clostridiales sp. SSC/2	clos3	163	2771	23.56	<	24.20
Clostridium saccharolyticum K10	clostr1	168	3073	22.01	<	23.70
Coprococcus catus GD/7	copr1	154	2985	22.53	<	24.02
Coprococcus sp. ART55/1	copr2	109	2128	22.43	<	23.88
Enterobacter cloacae cloacae NCTC 9394	ent1	325	3725	22.62	<	23.50
E. cloacae cloacae ATCC 13047	ent2	514	5518	22.61	<	23.56
Enterococcus sp. 7L76	entc1	132	2295	22.53	<	24.02
Escherichia coli O150:H5 SE15	ecol1	420	4488	22.78	<	23.64
E. coli SE11	ecol2	452	5002	22.73	<	23.66
Faecalitalia cylindroides T2-87	fac1	56	1442	22.85	<	24.32
Eubacterium rectale DSM 17629	eub1	155	2898	22.90	<	24.00
E. rectale M104/1	eub2	169	3212	22.82	<	24.11
E. siraeum 70/3	eub3	161	2347	22.70	<	23.85
E. siraeum V10Sc8a	eub4	147	2211	22.77	<	23.80
Faecalibacterium prausnitzii L2-6	faec1	136	2756	21.98	<	23.42
F. prausnitzii SL3/3	faec2	139	2746	22.12	<	23.48
Finegoldia magna ATCC 29328	fin1	124	1813	23.41	<	24.47
Fusobacterium nucleatum animalis 7_1	fuso1	92	2418	23.54	<	24.84
F. nucleatum vincentii 3_1_27	*fuso2*	140	2017	23.72	<	24.75
Fusobacterium sp. 3_1_36A2	fuso3	144	2153	23.69	<	24.84
Fusobacterium sp. 4_8	fuso4	134	2114	23.69	<	24.81
Gordonibacter pamelaeae DSM 19378	gord1	108	2027	21.78	<	22.87
Helicobacter pylori 35A	hel1	85	1513	22.32	<	24.06
H. pylori 83	hel2	91	1609	22.12	<	24.34
Lactobacillus fermentum IFO 3956	lab1	49	1843	22.10	<	23.52
L. paracasei 8700:2	lab2	153	2952	22.45	<	23.74
L. reuteri F275, JCM 1112	lab3	61	1935	22.57	<	23.94
L. reuteri SD2112, ATCC 55730	lab4	62	2300	22.20	<	24.17
L. rhamnosus ATCC 53103	lab5	148	2905	22.69	<	23.75
L. acidophilus NCFM	lab6	119	1875	23.17	<	24.30
L. brevis ATCC 367	lab7	134	2218	22.93	<	23.68
L. casei ATCC 334	lab8	142	2771	22.48	<	23.76
L. delbrueckii bulgaricus 2038	lab9	83	1792	23.11	<	23.82
L. gasseri ATCC 33323	lab10	60	1755	23.19	<	24.22
L. johnsonii DPC 6026	lab11	72	1772	22.68	<	24.26
L. plantarum WCFS1	lab12	167	3124	22.19	<	23.79
L. salivarius CECT 5713	lab13	54	1552	22.70	<	24.11
Megamonas hypermegale ART12/1	mega	90	2118	23.55	<	24.29
Roseburia intestinalis M50/1	rose1	188	3478	22.56	<	24.08
R. intestinalis XB6B4	rose2	159	3630	22.45	<	24.11
Ruminococcus bromii L2-63	rum1	167	1811	22.95	<	23.89
R. obeum A2-162	rum2	235	3155	22.56	<	24.02
R. champanellensis 18P13	rum3	186	2114	22.54	<	23.60
Ruminococcus sp. SR1/5	rum4	160	3260	22.66	<	24.06
R. torques L2-14	rum5	106	2798	22.45	<	24.03

PEC: Protein Energetic Cost; S: Statistically significant at $P < 0.01$

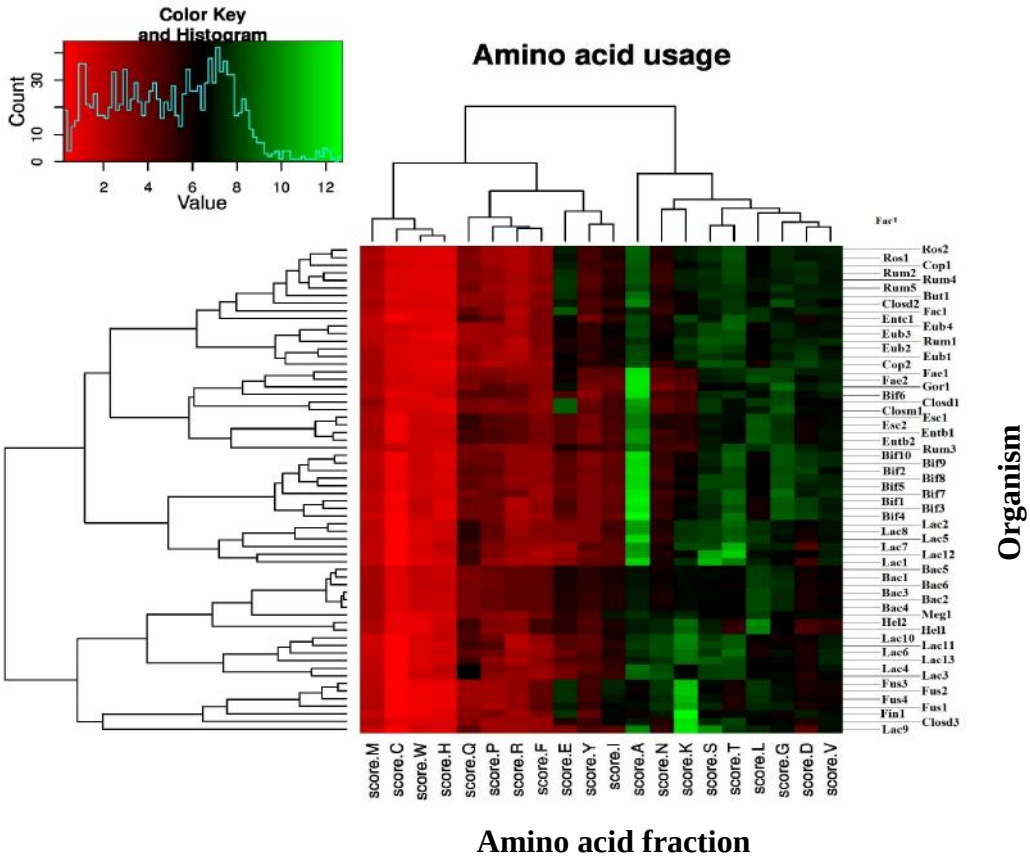

Figure 4.46: Amino acid usage based heatmap of the secretomes in human gut-associated bacterial members; Abbreviations have been used for the concerned bacterial members. The complete list of abbreviation has been depicted in **Table 4.37**

secretomes increased with the increase in proteome size of every concerned bacterial member.

4.3.2 Amino acid usage of bacterial secretomes:

Detailed investigation of amino acid usage was carried out among the bacterial members of human gut. Heatmap based on the relative frequencies of amino acids employed by the secretory proteins was generated and has been shown in Figure 4.46. The intensity of color gradually changed from red to green in the concerned heatmap with increase in frequency of a particular amino acid. It was noticeable from the heatmap that there was a distinct bias towards the usage of specific amino acids like Alanine (A), Serine (S), Glycine (G), Asparagine (N) and Threonine (T). These amino acids were used in high frequencies by the secretomes, unanimously among all concerned bacterial members. Another trend was also well evident that aromatic amino acids like Phenylalanine (F), Tyrosine (Y) and Tryptophan (W) were avoided

Organism	*ISP* (%)	*CPS* (%)	*Met* (%)	*PC* (%)
Bacteroides xylanisolvens XB1A	0.16	6.97	17.91	3.17
B. fragilis 638R	0.42	10.31	16.85	4.32
B. fragilis NCTC 9343	0.43	11.24	18.92	3.98
B. fragilis YCH46	0.43	12.06	18.72	3.26
B. thetaiotaomicron VPI-5482	0.22	8.50	19.83	2.94
B. vulgatus ATCC 8482	0.44	10.76	23.40	3.34
Bifidobacterium longum infantis 157F-NC	0.00	14.93	44.78	1.49
B. longum infantis JCM 1222	0.00	20.22	44.94	2.25
B. longum longum F8	0.00	13.79	48.28	3.45
B. longum longum JCM 1217	0.00	14.47	38.16	5.26
B. adolescentis 22L	0.98	24.51	33.33	1.96
B. animalis lactis AD011	4.69	18.75	35.94	0.00
B. bifidum PRL2010	4.21	15.79	20.00	6.32
B. breve UCC2003	1.01	18.18	41.41	4.04
B. longum DJO10A	1.77	18.58	36.28	1.77
B. longum NCC2705	2.17	20.65	34.78	2.17
Butyrivibrio fibrisolvens 16/4	1.01	14.14	24.24	2.53
Clostridiales sp. SM4/1	0.79	12.70	11.11	3.97
Clostridiales sp. SS3/4	1.18	12.94	24.71	7.06
Clostridiales sp. SSC/2	4.07	14.53	16.86	5.23
Clostridium saccharolyticum K10	1.71	14.29	21.14	5.71
Coprococcus catus GD/7	1.25	15.63	25.63	6.25
Coprococcus sp. ART55/1	2.70	14.41	16.22	4.50
Escherichia coli O150:H5 SE15	0.90	32.13	26.47	7.92
E. coli SE11	0.86	34.69	21.84	8.35
Enterobacter cloacae cloacae NCTC 9394	1.20	33.93	27.63	7.51
E. cloacae cloacae ATCC 13047	1.33	30.29	24.76	7.81
Enterococcus sp. 7L76	0.71	15.71	24.29	5.00
Faecalitalia cylindroides T2-87	1.72	8.62	18.97	8.62
Eubacterium rectale DSM 17629	2.48	16.77	17.39	3.73
E. rectale M104/1	2.30	14.37	17.24	4.02
E. siraeum 70/3	0.60	14.88	14.29	2.98
E. siraeum V10Sc8a	1.32	13.82	16.45	1.32
Faecalibacterium prausnitzii L2-6	0.00	9.09	27.97	4.90
F. prausnitzii SL3/3	0.00	8.97	22.07	6.21
Finegoldia magna ATCC 29328	2.38	17.46	15.08	4.76
Fusobacterium nucleatum animalis 7_1	0.00	25.53	17.02	4.26
F. nucleatum vincentii 3_1_27	0.70	21.13	22.54	4.23
Fusobacterium sp. 3_1_36A2	0.68	21.77	21.77	4.76
Fusobacterium sp. 4_8	0.74	24.26	17.65	5.15
Gordonibacter pamelaeae DSM 19378	2.59	17.24	22.41	7.76
Helicobacter pylori 35A	1.15	19.54	5.75	5.75
H. pylori 83	0.00	20.65	5.43	5.43
Lactobacillus fermentum IFO 3956	3.45	36.21	17.24	6.90
L. paracasei 8700:2	1.92	10.90	26.28	9.62
L. reuteri F275, JCM 1112	4.41	30.88	14.71	4.41
L. reuteri SD2112, ATCC 55730	4.35	31.88	13.04	7.25
L. rhamnosus ATCC 53103	3.23	10.97	27.10	9.68
L. acidophilus NCFM	1.63	16.26	21.95	4.07
L. brevis ATCC 367	2.13	17.73	17.73	4.26
L. casei ATCC 334	2.76	11.03	25.52	10.34
L. delbrueckii bulgaricus 2038	2.25	26.97	21.35	6.74
L. gasseri ATCC 33323	6.15	13.85	21.54	6.15
L. johnsonii DPC 6026	2.63	13.16	25.00	10.53
L. plantarum WCFS1	1.16	14.45	16.76	6.94
L. salivarius CECT 5713	0.00	20.00	20.00	5.00
Megamonas hypermegale ART12/1	2.20	10.99	17.58	2.20
Roseburia intestinalis M50/1	2.04	16.84	21.94	6.63
R. intestinalis XB6B4	2.42	18.18	22.42	7.27
Ruminococcus bromii L2-63	1.74	9.88	11.05	6.98
R. obeum A2-162	1.23	11.89	21.72	4.10
R. champanellensis 18P13	1.59	10.58	14.81	3.17
Ruminococcus sp. SR1/5	1.82	8.48	21.82	4.24
R. torques L2-14	4.55	11.82	25.45	8.18

Figure 4.47: Heatmap based on COG categories of secretomes in bacterial members of human gut; In the heatmap, *ISP* refers to *Information Storage and Processing*, *CPS* refers to *Cellular Processes and Signaling*, *Met* refers to *Metabolism*, *PC* refers *to Poorly Characterized* COG categories; In the concerned heatmap there is a gradual change of color shade from bright yellow to dark orange with increase in frequency of a particular COG category

by the secretomes, universally among the bacterial strains of interest, as was reflected from their relatively lower usage. Bulky and heavy chain amino acids like Arginine (R), Methionine (M) and Histidine (H) were also employed in very low frequencies among the secretome sets.

4.3.3 COG category analysis of bacterial secretomes:

COG categories of secretory proteins in human intestinal bacterial members were estimated and have been displayed in the form of a heatmap (Figure 4.47). There was a gradual change of color shade from bright yellow to dark orange with increase in frequency of a particular COG category in the heatmap. It was evident that the '*Metabolism*' and '*Cellular Processes and Signaling*' COG categories were prevalent, invariably in all concerned bacterial members. Few proteins were poorly categorized. Hypothetical proteins with undetermined COG categories were not taken into consideration. Further COG category analysis revealed that '*Carbohydrate transport and metabolism* [G]', '*Amino acid transport and metabolism* [E]' and '*Inorganic ion transport and metabolism* [P]' sub-categories, representing the '*Metabolism*' COG category, were present in higher frequencies. Among the '*Cellular Processes and Signaling*' COG category, '*Cell wall/membrane/ envelope biogenesis* [M]', '*Signal transduction mechanisms* [T]' and '*Cell motility* [N]' sub-categories were present in significant share. Thus, the secretomes of gut-associated bacteria were found to execute several crucial biological phenomena like transport of carbohydrates and amino acids, cell motility, biogenesis of cell membranes and efficient cell signaling. Such observations can be well explained in the light of the fact that secretory proteins are effectively involved in several 'remote-control' functions that include provision of nutrients, cellular cross-talks, apt cell signaling and morphogenesis (Tjalsma *et al.*, 2004).

4.3.4 Biosynthetic cost of the bacterial secretomes:

Protein energetic cost of the secretomes was assessed for all bacterial members residing in the human intestine and was compared with the non-secretome part of the respective proteomes. The synthetic cost of the secretome and non-secretome sets for all concerned bacterial members have been provided

in Table 4.37. It was apparent that the protein synthetic cost of the secretory proteins was significantly lower than the non-secretory cellular components.

Amino acid usage patterns of the secretomes (Figure 4.46) revealed that the extracellular secretory proteins preferred employing less expensive amino acids like Alanine (A), Serine (S), Glycine (G), Asparagine (N) and Threonine (T). Simultaneously, there was a universal trend among the secretomes to avoid the usage of expensive aromatic amino acids like Phenylalanine (F), Tyrosine (Y) and Tryptophan (W) and bulky amino acid residues like Arginine (R), Methionine (M) and Histidine (H). Enrichment of simple cheaper amino acids acted as the major force driving cost reduction among the secretory proteins. Such an instance of reduced synthetic cost of secretory proteins appeared well justified from the fact that protein secretion is often a 'one-way' street due to the absence of protein import systems (Saier, 1994). Loss of extracellular proteins is likely irreparable and there is least possibility of re-cycling these proteins by the cell's chaperon and protease systems (Reumann *et al.*, 2005; Saier, 1994). Accordingly, it would be economic and pragmatic for the microbial systems to use cheaper amino acids which are metabolically less taxing for the cell to produce, especially in case of secretomes that are lost permanently (Smith and Chapman, 2010). Microbes tend to curtail the cost of the secretory proteins for maintaining proper fitness and growth rate of the cell (Easom and Clarke, 2008; Pintar and Starmer, 2003; Kutsukake and Iino, 1994). Thus, economic pressure seemed to act on the extracellular secretory components to redeem the loss due to secretion, lower the burden on the cell and strictly follow cost-minimization policy among the sets of secretory proteins.

4.3.5 Evolutionary behavior of the secretory proteins:

Study of evolutionary constraints on the secretory proteins in bacterial communities has been a topic of enormous curiosity. The ratio (ω) of rate of non-synonymous substitutions per non-synonymous site (Ka) to rate of synonymous substitutions per synonymous site (Ks) has been suggested as an effective indicator of evolutionary selection acting on genes and genomes (Hurst, 2002). $\omega > 1$ indicates positive (diversifying) Darwinian selection whereas $\omega < 1$

Table 4.38 Comparative statistics of evolutionary rates for different members of human gut microflora

Combination	Ka/Ks of secretome	Significance	Ka/Ks of non-secretome
Bacteroides vulgatus ATCC 8482- *Bacteroides xylanisolvens* XB1A	0.05	N.S	0.05
Bacteroides vulgatus ATCC 8482- *Bacteroides fragilis* 638R	0.05	N.S	0.04
Bacteroides vulgatus ATCC 8482- *Bacteroides fragilis* NCTC 9343	0.05	N.S	0.05
Bifidobacterium longum longum F8- *Bifidobacterium longum longum* JCM 1217	0.22	N.S	0.19
Enterobacter cloacae cloacae NCTC 9394- *Enterobacter cloacae cloacae* ATCC 13047	0.06	N.S	0.05
Escherichia coli O150:H5 SE15- *Escherichia coli* SE11	0.10	N.S	0.10
Eubacterium rectale DSM 17629- *Eubacterium rectale* M104/1	0.21	N.S	0.16
Eubacterium siraeum 70/3- *Eubacterium siraeum* V10Sc8a	0.21	N.S	0.12
Faecalibacterium prausnitzii L2-6- *Faecalibacterium prausnitzii* SL3/3	0.10	N.S	0.07
Fusobacterium nucleatum animalis 7_1- *Fusobacterium nucleatum vincentii* 3_1_27	0.07	N.S	0.06
Fusobacterium sp. 3_1_36A2- *Fusobacterium* sp. 4_8	0.07	N.S	0.06
Helicobacter pylori 35A- *Helicobacter pylori* 83	0.25	N.S	0.24
Lactobacillus paracasei 8700:2- *Lactobacillus rhamnosus* ATCC 53103	0.07	N.S	0.04
Roseburia intestinalis M50/1- *Roseburia intestinalis* XB6B4	0.3	N.S	0.28
Ruminococcus obeum A2-162- *Ruminococcus torques* L2-14	0.10	N.S	0.09

N.S: Not Statistically Significant

signifies purifying (refining) selection. Secretomes, the complete set of secretory proteins, are key components involved in scavenging activities and in proper foundation of mutualistic and antagonistic relationships with host. Our group previously investigated the effect of evolutionary forces of selection acting on the secretory proteins in pathogenic members of Actinobacteria (Thakur *et al.,* 2013). Interestingly, the secretomes were observed to display higher ω values with respect to the non-secretory counterparts which suggested that the evolutionary constraints were much relaxed on the secretory proteins in comparison to the non-secretory ones. Such relaxed evolutionary signatures were suggested as strategies employed by the bacterial masses for apposite adaptation and proficient interaction with concerned host. Nogueira and colleagues found similar results in

selected members of Proteobacteria and Firmicutes (Nogueira *et al.*, 2012).

In this present work, we performed an extensive evolutionary analysis considering representative members from all major bacterial phyla present in human gut. The complete evolutionary details have been provided in Table 4.38. It was evident in our case that the evolutionary rates (ω) of the secretomes were marginally higher than that of the non-secretory proteins however, there was no statistical significance between the respective rates of evolution. Such a trend was consistent in all concerned members of gut-associated bacteria that we studied (Table 4.38). Our observations seemed logical from the viewpoint that microbial members of human gut colonize human gastrointestinal tract right from birth. Similar codon usage patterns between human gut microbes and human host (observed earlier in this thesis) tend to support the theory of co-evolution and co-existence. Members of human gut prefer symbiotic residence in human intestine deriving essential nutrients for their survival and in turn confer various beneficial services pertaining to intestinal homeostasis of human host. Our results indicate that the microbial residents of the gut might adopt such an evolutionary strategy to adapt successfully in human intestinal niche. Further investigations pertaining to evolutionary traits of secretory proteins are required before arriving at any definite conclusion.

4.4 Identification of drug targets in *H. pylori* 35A:

Human gastrointestinal tract is an ideal abode for numerous microbes. Some microorganisms render beneficial effects whereas, others exhibit pathogenic behavior resulting in critical consequences (Sun and Chang, 2014). Several strains of *Helicobacter pylori* constitute a significant part of the human intestinal microflora. However, on commotion of the gut microbial equilibrium and several other compromised conditions of the host, resident *H. pylori* strains exert severe degrees of pathogenesis causing active gastritis, duodenal ulcers, and gastric cancer in human (Kelly, 1998; Kusters *et al.*, 2006; Lehours *et al.*, 2009). *H. pylori* has been a subject of thorough medical research. Current available treatment strategies involve usage of drugs to redeem pain and acidity associated with *H. pylori* mediated infections. Potential therapeutic targets in various *H. pylori* strains like HpB38,

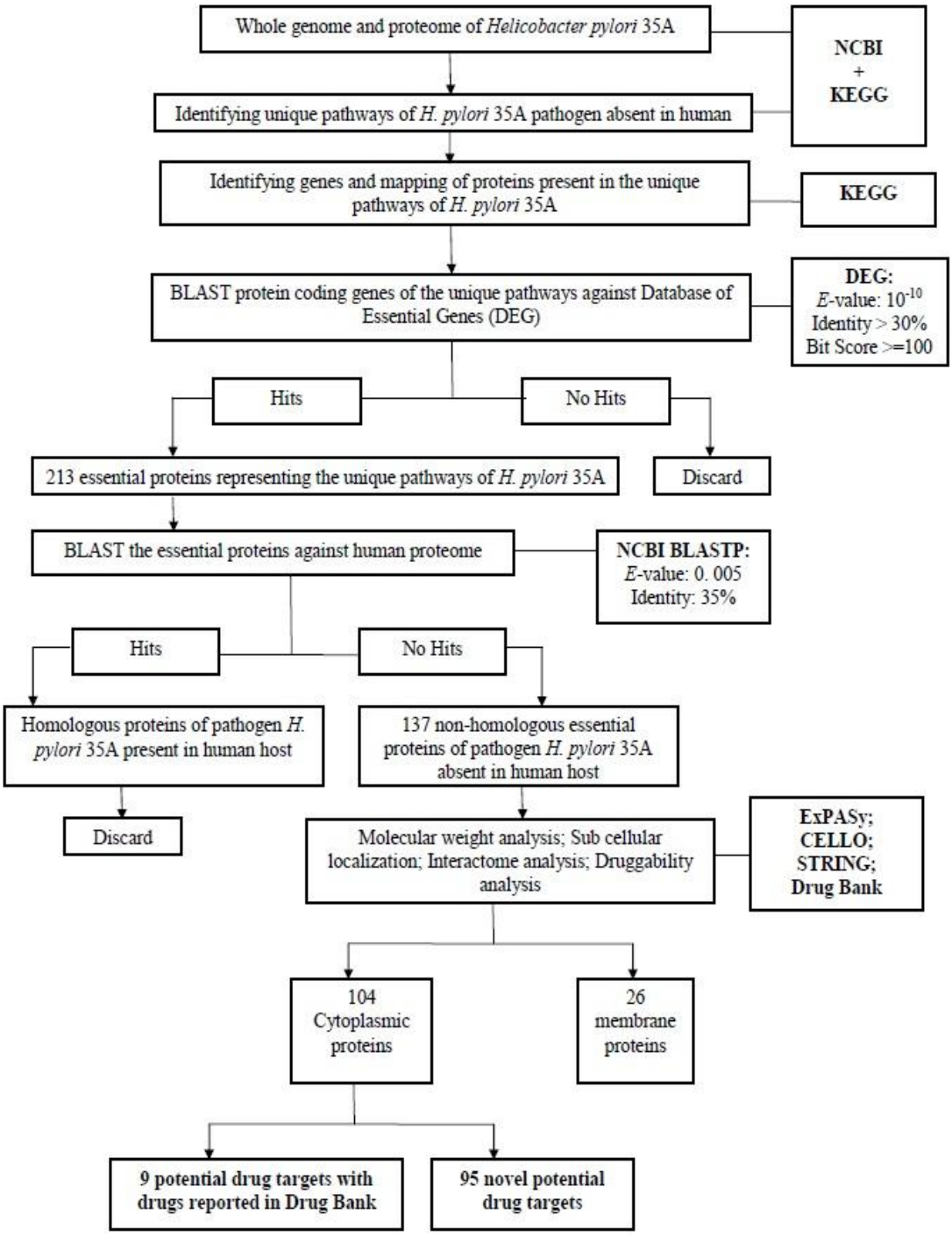

Figure 4.48: Flowchart depicting work scheme for identification of potential drug targets in *H. pylori* 35A

HpP12, HpG27, Hpshi470, and HpSJM180 have recently been identified (Neelapu and Pavani, 2013). However, thorough profiling of putative drug target genes in *H. pylori* 35A strain, a pathogenic member of human gut, and subsequent validation by molecular docking studies is yet to be achieved. We have made a modest effort to fish out the essential proteins of *H. pylori* 35A that are absent in human host and propose novel drug targets for the concerned pathogen. *In silico* subtractive genomics based approach of screening the targets has been supported with molecular docking studies and investigations. Pertaining results promise to establish a platform for successful and effective drug development against *H. pylori* 35A.

4.4.1 Non-host essential proteins of H. pylori 35A:

Bacterial systems possess metabolic pathways like monobactam biosynthesis pathway, benzoate degradation pathway, vancomycin resistance pathway, etc., that are absent in human beings. The preliminary strategy we opted was to screen these unique pathways present in *H. pylori* 35A and fish out genes associated with the pathways. Presently, KEGG contains information about 91 different metabolic pathways of *H. pylori* 35A and 297 metabolic pathways of *Homo sapiens*. Out of 91 metabolic pathways in *H. pylori* 35A, 18 were found to be unique i.e., absent in *H. sapiens*. A total set of 376 unique proteins represented the unique pathways of *H. pylori* 35A. Among the 376 unique proteins, 213 were found to be essential as observed from Database of Essential Genes (DEG). Proper BLASTP of these proteins against the human proteome resulted in a set of 137 proteins that had no significant similarity with human host.

The step of filtering non-host (non-human) essential proteins of *H. pylori* 35A was a decisive one. We intended to ensure the inclusion of the essential genes of the organism *H. pylori* 35A associated crucially with its existence (Duffield *et al.*, 2010; Freiberg *et al.*, 2001) and simultaneously, prevent host toxicity by avoiding the incorporation of any gene of the pathogen displaying significant similarity with human host (Wang *et al.*, 2004). Natural knockout-rate prediction states that essential genes are more evolutionarily conserved than the non-essential ones (Jordan *et al.*, 2002). The reason behind such an observation is the fact that purifying selection tends to exert

more constraint on the essential genes than on the non-essential ones (Jordan *et al.*, 2002). Keeping this theory in mind Alan Wilson and his colleagues proposed that essential genes should evolve slower than the non-essential genes (Wilson *et al.*, 1977). This is the basis of classic knockout-rate prediction. The non-host essential proteins representing the unique pathways of *H. pylori* 35A served as our final set of data for the next phases of analysis. Proper disruption of functions of any of the non-host essential proteins of *H. pylori* 35A would imply efficient choking of the concerned organism as the essential proteins are vitally associated with the physiological robustness and sustenance of the organism in the human intestinal niche.

4.4.2 Analysis of molecular weight and subcellular localization:

A comprehensive set of 137 non-host essential proteins of *H. pylori* 35A were obtained after preliminary screening. One hundred and thirty five putative drug candidate proteins were found to possess molecular weight less than 100 kDa, assumed to be ideal drug targets (Duffield *et al.*, 2010). Subcellular localization analysis of a protein reveals whether the protein is suitable as a drug or a vaccine target. Cytoplasmic proteins tend to work better as drug targets while membrane proteins have been suggested to be effective as targets for vaccine development (Barh *et al.*, 2011). Out of 135 proteins, 26 were predicted to be membrane proteins and 104 proteins were found to be cytoplasmic. The cytoplasmic proteins were considered as plausible drug target proteins and were considered further for drug target analysis. The complete flowchart of identification of potential therapeutic targets in *H. pylori* 35A has been depicted in Figure 4.48.

4.4.3 Interactome survey of the putative drug targets:

STRING database aims to provide a critical evaluation and integration of protein–protein interactions, including direct (physical) as well as indirect (functional) associations. The 104 proteins of our interest were checked for interaction among them. The concerned proteins displayed multiple interactions among themselves (Supplementary Figure 1 [SF1]). Dihydrodipicolinate reductase (KEGG ID: HMPREF4655_21095) was found to display a maximum number of 26 interaction partners.

Interactome study reveals the

interaction between proteins and gives a confidence score according to various analyzed parameters. It is believed that if a protein having several interacting partners is blocked or cut off, the related proteins which are in interaction or association with the concerned protein also become vulnerable (Jensen *et al.*, 2009) (Franceschini *et al.*, 2013). Due to mammoth scale of interactions between the biological entities in an organism it is difficult to keep track of all the cellular processes. Interaction study plays a vital role to determine systematic characterization (Szklarczyk *et al.*, 2011) that illustrates how proteins work in concert to synthesize different compounds or to transmit information. In functional study analysis, arguably the most useful networks are those that integrate all types of interactions: stable physical associations, transient binding, substrate chaining, information relay and others. These functional associations give a vivid idea to resolve the chokepoint reactions for the concerned organism (Hasan *et al.*, 2006; Nwaka and Hudson, 2006). Many proteins present in our dataset displayed interaction score > 0.90 which indicated their ability to interact with large number of proteins. Thus, blocking these proteins and electing them as drug targets might cause a greater damage to *H. pylori* 35A.

4.4.4 Strand specific asymmetry among the potential drug target genes:

Methodical inspection of strand specificity revealed that there was an even share of drug candidate genes in the complimentary strands of replication. It was observed that 54 out of 104 protein coding genes were found to represent the leading strand whereas, 50 out of 104 protein coding genes were found to lie on the lagging strand. Generally genes coding for essential proteins are placed on the leading strand (Samal *et al.*, 2015). This is a general molecular strategy to avoid the expected head-on collision between DNA polymerase and RNA polymerase in lagging strand. The occurrence of a significant share of essential drug candidate genes in the lagging strand can be well explained in the light of the fact that microbial genomes can afford to keep a higher proportion of protein coding genes and even the essential ones in the lagging strand to attain proper mutation-selection balance (Chen and Zhang, 2013). Thus, it was evident that there was no strand bias among the tentative

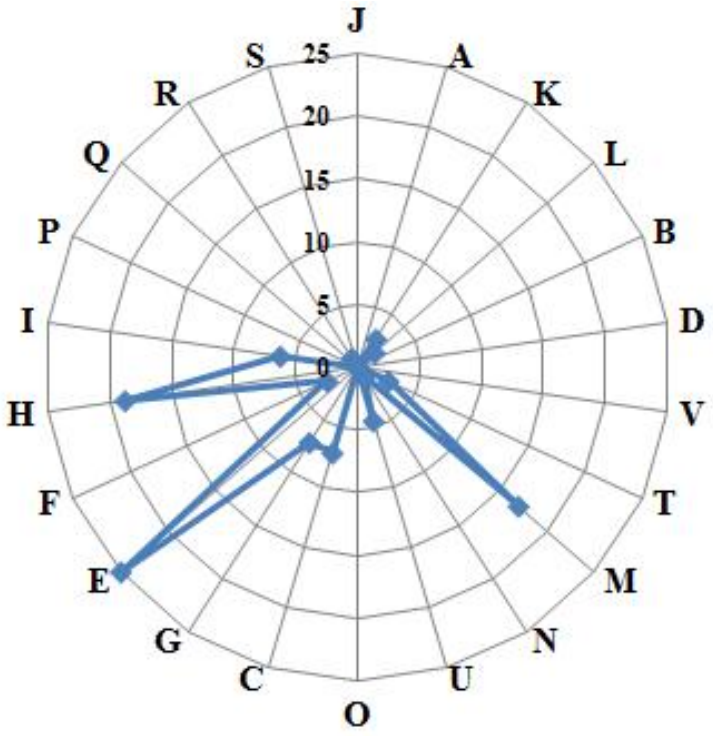

Figure 4.49: Radial plot depicting COG categories of potential therapeutic targets in *H. pylori* 35A- [J]- Translation, ribosomal structure and biogenesis; [A]- RNA processing and modification; [K]- Transcription; [L]- Replication, recombination and repair; [B]- Chromatin structure and dynamics; [D]- Cell cycle control, cell division, chromosome partitioning; [V]- Defense mechanisms; [T]- Signal transduction mechanisms; [M]- Cell wall/membrane/envelope biogenesis; [N]- Cell motility; [U]- Intracellular trafficking, secretion, and vesicular transport; [O]- Posttranslational modification, protein turnover, chaperones; [C]- Energy production and conversion; [G]- Carbohydrate transport and metabolism; [E]- Amino acid transport and metabolism; [F]- Nucleotide transport and metabolism; [H]- Coenzyme transport and metabolism; [I]- Lipid transport and metabolism; [P]- Inorganic ion transport and metabolism; [Q]- Secondary metabolites biosynthesis, transport and catabolism; [R]- General function prediction only; [S]- Function unknown

therapeutic target genes.

4.4.5 Clusters of Orthologous Groups (COG) category analysis:

COG categories of the putative drug candidate proteins in *H. pylori* 35A were estimated and displayed in the form of a radial plot (Figure 4.49). It was apparent that majority of the putative drug target proteins reflected '*Amino acid transport and metabolism* [E]' and '*Coenzyme transport and metabolism* [H]' sub-categories representing '*Metabolism*' COG category and '*Cell wall/membrane/ envelope biogenesis* [M]' sub-category representing the '*Cellular Processes and Signaling*' COG category. '[E]' and '[H]' COG sub-categories represent crucial functional components associated with proper amino acid and coenzyme metabolism respectively and are vital for the fitness of an organism. Proteins associated with '[M]' sub-category execute vital

functions like biogenesis of cell membranes and efficient cell signaling and thus, are imperative for proper growth and survival of an organism. Therefore, targeting such critical components seems rational for proper drug development against *H. pylori* 35A.

4.4.6 Enzyme classification of the potential drug targets:

It could be inferred from systematic enzyme classification that among the potential drug targets, majority of the proteins were transferases followed by oxidoreductases and lyases. Transferases enact the transfer of specific functional groups from one molecule to another. They are involved in countless reactions in a cell. It has already been reported that transferases are excellent drug targets (Dekker and Haisma, 2009). Furthermore, transferases have also been reported as ideal drug target candidates for a wide range of other diseases encompassing gastrointestinal disorders and cancer (Townsend and Tew, 2003). Hence, targeting the proteins with transferase enzymatic activities would mean effective hindrance to growth and survival of *H. pylori* 35A.

4.4.7 Druggability of the potential therapeutic targets:

DrugBank has a database that contains unique cheminformatics and bioinformatics data with comprehensive drug target information. Among our targets, only 9 out of 104 proteins were found to have DrugBank entries (Table 4.39) with approved or experimental drugs reported against them. Ninety-five proteins present in our dataset had no DrugBank entries and might be considered as 'novel' drug discovery targets (Supplementary Table 5 [ST5]). 6,7-dimethyl-8-ribityllumazine synthase (KEGG ID: HMPREF4655_20208) was found to have the highest number of drugs reported in DrugBank.

4.4.8 Ranking the putative therapeutic targets:

It is useful to rank the putative therapeutic targets in order to decide which target has a higher probability of being effective for laboratory experiments pertaining to drug development. We ranked the putative therapeutic targets in *H. pylori* 35A based on Drug Target Score (DTS). 6,7 -dimethyl-8-ribityllumazine synthase was found to display the highest score and was inferred to possess the maximum potential of being an effective therapeutic target. The detailed rank tally has been provided in

Table 4.39 Potential 'druggable' targets identified in *H. pylori* 35A by *in silico* subtractive genomics based approach

KEGG ID	Gene Name	STRING Interaction	DrugBank	DEG Result	Drug Score	Rank
HMPREF4655_20208	ribH	17	10	14	23.8	1
HMPREF4655_20212	panC	15	4	13	7.8	2
HMPREF4655_21095	dapB	26	1	21	5.46	3
HMPREF4655_20365	mtnN	15	5	6	4.5	4
HMPREF4655_20654	aroQ	18	2	10	3.6	5
HMPREF4655_21173	alr	21	1	15	3.15	6
HMPREF4655_21474	trpG	16	2	8	2.56	7
HMPREF4655_21027	rpiB	13	2	7	1.82	8
HMPREF4655_21300	porB	23	1	1	0.23	9

Table 4.39. Pantoate-beta-alanine ligase (KEGG ID: HMPREF4655_20212) and dihydrodipicolinate reductase (KEGG ID: HMPREF4655_21095) were found to rank second and third respectively based on their DTS scores.

4.4.9 Molecular Docking Analysis of the 'druggable' and 'novel' drug targets:

Druggability analysis of the putative therapeutic targets in *H. pylori* 35A produced a set of 9 proteins with reported DrugBank entries ('druggable') whereas, 95 proteins were found to have no DrugBank entries ('novel' drug discovery targets). Protein 3-dehydroquinate dehydratase (aroQ) (KEGG ID: HMPREF4655_20654) was chosen from the 'druggable' set of putative drug target proteins for molecular docking analysis. It is an essential protein of *H. pylori* 35A that catalyzes trans-dehydration (Chalker *et al.*, 2001) and choosing such an essential protein as probable drug target would mean effective choking of the organism. The aroQ protein was found to have two experimental drugs, DB02801 (2,3-Anhydro-Quinic Acid) and DB03739 (3-Hydroxyimino Quinic Acid), reported in DrugBank to probably act effectively against it. We also considered the four approved drugs- DB00130 (L-Glutamine), DB00173 (Adenine), DB00260 (Cycloserine) and DB00698 (Nitrofurantoin) that are reported against other proteins of the 'druggable' set in DrugBank in order to

Table 4.40 Docking scores of 'druggable' and 'novel' drug targets with experimental and approved drugs and phytochemical D-Limonene

Drug Type	'Druggable' aroQ protein		'Novel' VirB11 protein	
Experimental Drugs	DrugBank ID	Docking Scores (kcal/ mol)	DrugBank ID	Docking Scores (kcal/ mol)
	DB02801	-4.2	DB02801	-6.2
	DB03739	-4.0	DB03739	-7.1
Phytochemical	D-Limonene	-4.4	D-Limonene	-6.1
Approved Drugs	DrugBank ID	Docking Scores (kcal/ mol)	DrugBank ID	Docking Scores (kcal/ mol)
	DB00130	-3.0	DB00130	-4.3
	DB00173	-3.8	DB00173	-5.7
	DB00260	-3.5	DB00260	-5.7
	DB00698	-4.3	DB00698	-7.4

perform an exhaustive molecular docking analysis. Phytochemical D-Limonene was also considered as a probable ligand as the concerned chemical has been reported to be associated with anti-cancer (gastric) and anti-ulcer (gastric) activities in Dr Duke's Phytochemical and Ethnobotanical Database (https:// phytochem.nal.usda.gov/phytochem). P -type DNA transfer ATPase VirB11 (VirB11) (KEGG ID: HMPREF4655_21080) protein was considered from the 'novel' set of tentative drug targets for docking investigations. VirB11 protein is essential one associated with ATP binding activities. It has also been reported to be associated with type IV secretion pathway (Backert *et al.*, 2000). The same set of ligands was employed for VirB11 protein. The comprehensive tally of binding affinities has been provided in Table 4.40. It was evident from Table 4.40 that the 'druggable' protein aroQ was found to display moderate binding affinities of -4.2 kcal/mol and -4.0 kcal/mol against the two experimental drugs DB02801 and DB03739 respectively. Interestingly, aroQ protein was found to display better binding affinities of -4.3 kcal/mol and -4.4 kcal/mol with approved drug DB00698 (Nitrofurantoin) and phytochemical D-Limonene respectively. The docked complexes of aroQ with Nitrofurantoin and D-Limonene respectively have been displayed in Figure 4.50.

Surprisingly, in case of VirB11 protein the experimental drugs DB02801 and DB03739 were found to exhibit considerable high binding affinities of -

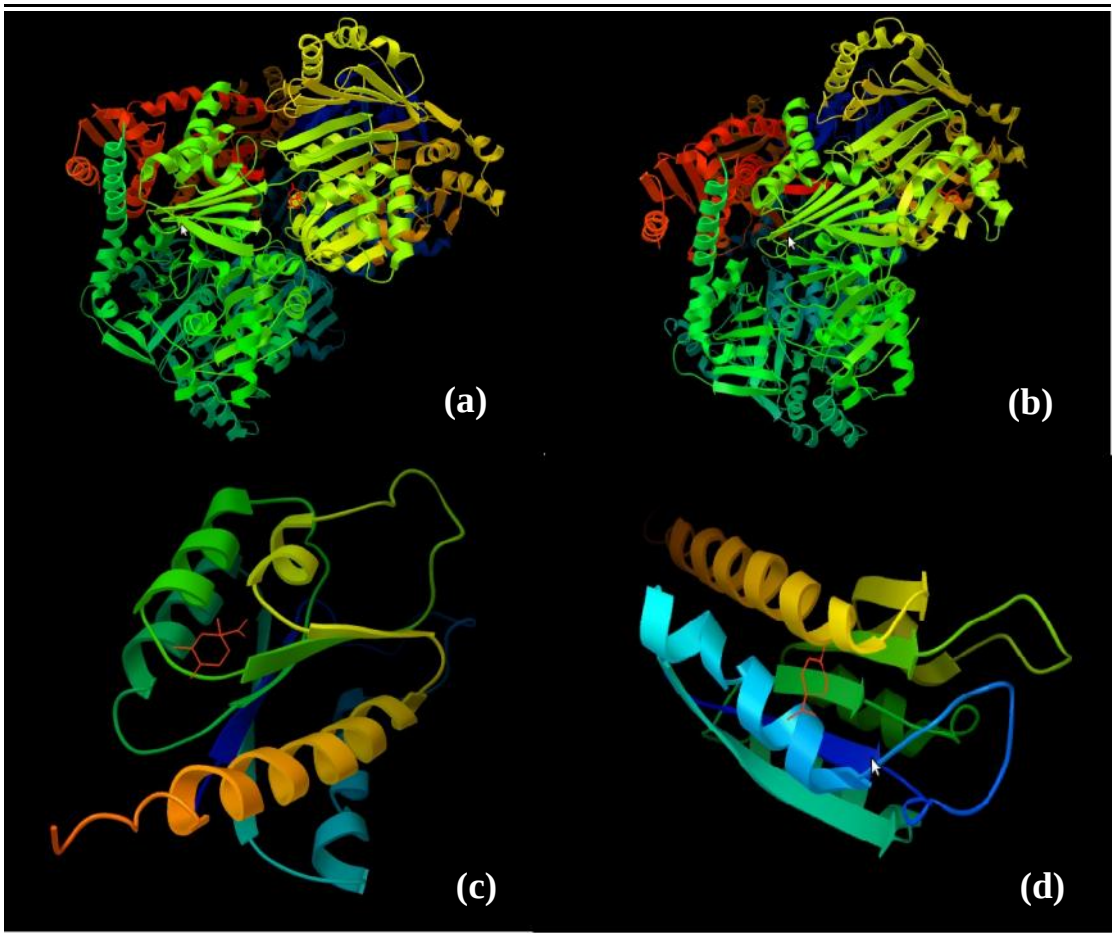

Figure 4.50: Docked complexes of **(a)** aroQ protein with Nitrofurantoin (DB00698) **(b)** aroQ protein with D-Limonene **(c)** VirB11 protein with Nitrofurantoin (DB00698) **(d)** VirB11 protein with D-Limonene

6.2 kcal/mol and -7.1 kcal/mol respectively (more than the binding affinities with aroQ protein). However, the approved drug DB00698 (Nitrofurantoin) was found to display the highest binding affinity of -7.4 kcal/mol. Phytochemical D-Limonene also exhibited significant high binding affinity of -6.1 kcal/mol with VirB11 protein. The docked complexes of VirB11 with Nitrofurantoin and D-Limonene respectively have been displayed in Figure 4.50.

Thus, it was evident from molecular docking analysis that the 'novel' target VirB11 protein might act as a better drug target than aroQ. Furthermore, it could be suggested from extensive docking analysis and know-how of binding affinities that the chemical Nitrofurantoin (DB00698), known to exhibit bacteriostatic or bactericidal effects against a wide range of Gram-negative bacteria (D'Arcy, 1985), might be a better lead molecule than the other tentative drug candidates for both the 'druggable' aroQ and 'novel' VirB11 protein targets. Phytochemical

D- Limonene also promises to be a potential lead molecule as was evident from docking analysis. However, the effectiveness of the targets predicted in this approach and lead molecules proposed against them can only be confirmed once all the reported targets are tried with the respective lead molecules and further validated with wet lab experiments for serving the wholesome purpose of killing the menacing pathogen *H. pylori* 35A.