Codon usage bias of the overlapping genes in microbial genomes

Master's Thesis
Preethy Sasidharan Nair
Master's Degree Programme in Bioinformatics
Institute of Medical Technology
University of Tampere
October 2010

Acknowledgements

My sincere gratitude goes to

Professor Mauno Vihinen for giving me an opportunity to work with him in the Bioinformatics Research Group at Institute of Medical Technology, University of Tampere, for the wonderful topic given to me for doing my Masters Thesis work, for the guidance provided and for the corrections to the thesis.

Docent Csaba Ortutay for the patient guidance provided by him all throughout this work and for the corrections to the thesis. Without his timely help and suggestions, it would not have been possible to finish the thesis work this fast. Thanks really a lot Csaba.

Martti Tolvanen, Ph. Lic., my teacher, who has helped me all around the Masters Programme, the teachers from Turku, Dr. Philip Ginter, Dr. Pentti Riikkonen and Dr. Hanna Suominen. Special thanks to Philip Ginter for the very interesting classes which laid a solid foundation for programming.

Professor Esa Uusipaikka, Professor Tapio Nummi, Dr. Klaus Nordhausen and Dr. Heidi Kainulainen for the Statistics and R classes, Dr. Tommi Aho for the R programming classes, Hannu Korhonen and Toni Vormisto for help with all the computer related problems, Ayodeji Olatubosun for the moral support, advice and suggestions throughout the Masters Degree Programme and all my friends at IMT and outside who have emotionally supported me.

Venki, our son, who has suffered the most due to my studies, my husband, Sivaram for his patient and steady support, my mother and father, who have encouraged me constantly throughout my studies.

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE Institute of Medical Technology Faculty of Medicine Tampere, Finland
Author:	Preethy Sasidharan Nair
Title:	Codon usage bias of the overlapping genes in microbial genomes
Pages:	90 pp. + Appendix 10 pp.
Supervisors:	Professor Mauno Vihinen, Ph. D. and Docent Csaba Ortutay, Ph. D.
Reviewers:	Professor Mauno Vihinen, Ph. D. and Docent Csaba Ortutay, Ph. D
Time:	October, 2010.

Abstract

Background and aims: Overlapping genes are adjacent genes whose coding sequences overlap partially or completely. They are abundant in the viruses and also present in archaea, prokaryotes and eukaryotes. Former studies have showed that the overlapping genes play substantial roles in the genome size minimization, gene expression regulation and slowing down of evolution as modifications of the overlap part may cause deleterious effect on both of the genes. The present study was aimed to find whether there is any difference in the codon usage by the overlapping part of the genomes compared to the other parts and if any, to investigate trends in the codon usage bias patterns in the phylum level of taxonomy.

Methods: GenBank identifiers for the completely sequenced microbial genomes from the NCBI ftp site were collected and filtered to contain those for which the codon usage tables were available from the CUTG database. Perl and R scripts were created for data generation, result formatting and analyses. Genomic overlaps were identified, codon usage of the overlapping parts were estimated and compared to that of the normal parts. As the codon usage by the genomic overlaps in the analysed genomes showed a significant bias, trends of the codon usage bias were investigated in the phylum level of taxonomy by principal component analysis, self-organizing maps, correspondence analysis and heat map visualisations.

Results: Overlapping genes are present in all of the analysed microbes. Overlaps of length greater than nine base pairs are same strand overlaps. Codon usage by the codons in the overlapping part of all of the examined genomes showed a significant bias (p-vale <2.2e-16 from the χ^2 goodness of fit tests) compared to the overall codon usage for the respective species in the CUTG database. The most overrepresented codon was TGA (73% of the analysed species). Analysis of the codon usage bias patterns using self-organizing map clustering followed by heat map visualisations shows that the clusters (codon usage bias patterns) identified with self-organizing map are very well associated with most of the Phyla analysed.

Conclusions: Genomic overlaps are present in all of the analysed species. Same strand overlaps are more abundant than the opposite strand overlaps. Codon usage of the genomic overlaps differs significantly from the normal part of the genomes in the analysed species. A well-defined trend in the stop codon usage bias pattern by the overlap parts in the phylum level of taxonomy is evident from the self-organizing map and heat map visualizations. This finding gives the evidence for the role of genomic overlaps in translational efficiency and gene expression regulation through transcriptional and translational coupling.

Contents

Abbreviations

ASN.1	Abstract syntax notation 1
BPhyOG	Bacterial phylogenies based on overlapping genes, a web server for reconstruction of whole-genome bacterial phylogenies based on overlapping genes
CA	Correspondence analysis
CAI	Codon adaptation index
CDS	Coding sequence
CUTG	Codon Usage Tabulated from GenBank
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
dsRNA	Double stranded ribonucleic acid
EMBL	European Molecular Biology Laboratory
ftp	File transfer protocol
GI No	GenBank identifier number
GC	Guanine-cytosine
INSDC	International Nucleotide Sequence Database Collaboration
mRNA	Messenger ribonucleic acid
NAT	Natural antisense transcript
N_c	Effective number of codons
NCBI	National Center for Biotechnology Information
OG	Overlapping gene
PCA	Principal component analysis
RNA	Ribonucleic acid
RSCU	Relative synonymous codon usage
SD	Standard deviation
SOM	Self-organizing map
tRNA	Transfer ribonucleic acid
UTR	Untranslated region

1 Introduction

Overlapping genes are adjacent genes in the genome of an organism whose coding sequences overlap partially or completely and get translated to different proteins. They are abundant in viruses, and are common in prokaryotes and eukaryotes (Barell *et al*., 1976; Normark *et al*., 1983; Wagner and Simons, 1994; Veeramachaneni *et al*., 2004; Kim *et al*., 2009). Different studies show that genomic overlaps have originated as a result of the stop codon deletion or mutation or a near-end frame shift extending the protein translation till the next in-frame stop codon (Fukuda *et al*., 1999, 2000; Sakharkar *et al*., 2005) or due to the acquisition of an upstream start codon by the downstream gene (Cock and Whitworth, 2010). Genomic overlaps in microbes have significant role in genome size minimization (Lillo and Krakauer, 2007) and have regulatory properties in the gene expression through translational coupling of functionally related polypeptides (Normark *et al*., 1983; Cooper *et al*., 1998; Inokuchi *et al*., 2000; Zheng *et al*., 2002). As a mutation in the overlapping part affects both genes, changes are less likely to occur in the overlapping part, hence highly conserved (Miyata and Yasunaga, 1978; Krakauer, 2000), and can be used as rare genomic markers for inferring phylogeny (Luo *et al*., 2006).

Studies on gene overlaps in viruses and eukaryotes showed that the sequence composition of overlapping proteins have bias towards disorder-promoting amino acids (Rancurel *et al*., 2009). The niche and life style of the microbes influence the numbers of overlapping genes present (Sakharkar *et al*., 2005; Fukuda *et al*., 2003; Johnson and Chisholm, 2004) and same strand overlaps are more abundant in microbes (Fukuda *et al*., 1999; Johnson and Chisholm, 2004; Cock and Whitworth, 2007).

The present study was aimed to find the occurrence of overlaps in the microbial species for which the complete genomes were available, calculate the codon usage by the overlap part, determine whether the codon usage by the overlapping parts exhibit any bias when compared to the overall codon usage of the genomes and find the patterns, if any, in the codon usage bias in the overlapping parts in different levels of taxonomy and to investigate the biological significance of the codon usage bias patterns.

2 Literature review

2.1 Overlapping genes

2.1.1 Discovery and prevalence

Overlapping genes have been discovered to occur in the genomes since the mid 1970's and was invented first from the single stranded DNA-phage, ΦX174, while sequencing its genome (Barell *et al*., 1976; Sanger *et al*., 1977). The discovery of overlapping genes chipped in to the crisis of the gene concept, which was widely accepted to be having a linear model (Portin, 1993, 2009). The first overlapping gene pair discovered was the genes D and E in the bacteriophage ΦX174 which overlap each other and were translated in two different reading frames from a common DNA sequence. Overlapping genes were found later in the genomes of DNA viruses, RNA viruses, prokaryotes and eukaryotes (Normark *et al*., 1983; Samuel 1989).

Extensive studies on the overlapping genes were initially done in viral genomes, where the genomic size constraints made it a common instance to maximise the information content (Dillon, 1987; Pavesi, 2000; McGirr and Buehuring, 2006; Pavesi, 2006; Bofkin and Goldman, 2007). Bacterial and archaeal genomes have very high gene density with more than 90% of their genomic DNA coding for proteins and there are many pairs of genes in their genomes, whose coding regions overlap. In *Escherichia coli*, promoter of the *ampC* β-lactamase gene found to be located within the last gene of the fumarate reductase (f*rd*) operon, acts as the transcription terminator of its preceding operon (Grundström *et al*., 1982).

The first overlapping gene in eukaryote was discovered in the *Drosophila melanogaster* (Henikoff *et al*., 1986), where a pupal cuticle protein (*Pcp*) gene was found within the first intron of the *Gart* locus, which encodes three enzymes involved in the purine biosynthesis. Further incidences of the overlapping gene in eukaryotes were found in Drosophila's dopa decarboxylase (*Ddc*) region (Spencer *et al*., 1986) and in mouse (Williams *et al*., 1986). The instances of overlapping genes were found in humans since the finding of the overlap between the last exon of the *P450c21*, the gene that encodes human adrenal steroid 21-hydroxylase and the transcript from its

opposite strand (Morel *et al*., 1989). Occurrence of overlapping gene pairs in plants have also been exposed in several studies (Terryn and Rouzé, 2000).

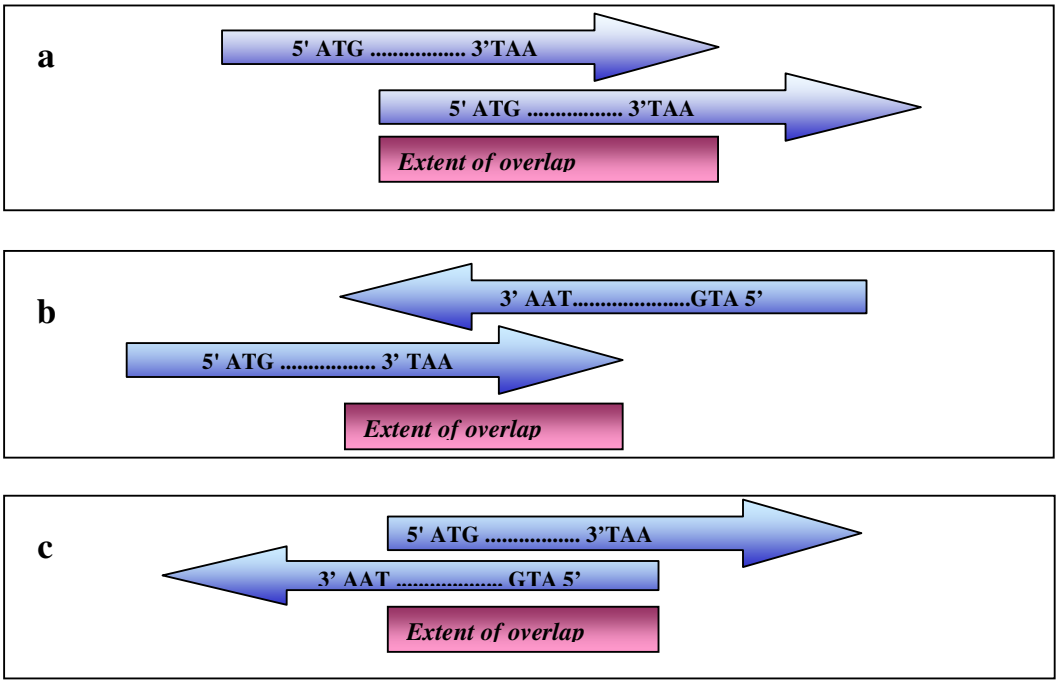
Bioinformatics tools have been used for discovering and analysing overlapping genes since the draft sequences of human, mouse and fruit fly became available in the public databases (Shendure and Church, 2002; Fahey *et al*., 2002; Yelin *et al*., 2003; Kiyosawa *et al*., 2003). But, there can be changes in the inferred number of overlapping genes as the analyses of the overlapping genes are hindered by the poor annotation, sequencing errors, limitations of the gene-finding algorithms etc. (Burge & Karlin, 1998).

2.1.2 Origin of genomic overlaps

A variety of reasons have been proposed about the origin of the overlaps in the genomes. Overlapping genes were proved to be evolved for packing more amount of genetic information to the genomes (Sakharkar *et al*., 2005) and due to the mutational bias for deletion (Clark *et al*., 2001). A comparative study conducted on the genomic overlaps in the two different Mycoplasma species viz., *Mycoplasma genitalium* and *Mycoplasma pneumoniae* by Fukuda *et al*., (Fukuda *et al.,* 2003) along with other studies (Fukuda *et al*., 1999, 2000; Sakharkar *et al*., 2005) show that the genomic overlaps in the analysed species were created by the 3' end elongation of the upstream gene due to a stop codon loss as a result of deletion, point mutation or frame shift in the end of the coding part.

Overlapping genes were suspected to also have arisen due to overprinting (Keese and Gibbs, 1992; Sander and Schulz, 1979) generating different coding sequences from an existing nucleotide sequence by translating it *de novo* in a different reading frame or from noncoding open reading frames, evident from the original gene function retention. Translation of multiple reading frames can also happen by internal *de novo* initiation in an alternative reading frame without any need for ribosomal frame shifting (Atkins *et al*., 1979; Chang *et al*., 1989). Another study shows that gene overlaps arise due to the N-terminal or 5' extension of the downstream gene by adopting new start codons, creating a novel amino acid sequence at the N-terminus of the encoded protein in prokaryotes (Cock and Whitworth, 2007, 2010).

2.1.3 Classification of the overlapping genes



d

Type Of Overlap	Direction Of Transcription	
Partial	Convergent Divergent	Parallel
Complete	Nested Antiparallel Embedded Antiparallel	Nested Parallel Embedded Parallel

Figure 2.1. Extent of overlap and classification of overlaps. The extent of overlap as used in this study for (a) parallel overlaps, (b) anti-parallel convergent overlaps and (c) anti-parallel divergent overlaps. Classification of overlaps from Solda *et al*. (Solda *et al*., 2008) is given in (d).

Different criteria had been used for the classification of overlapping gene pairs including the gene reciprocal orientation, overlap extent and the regions involved in the overlap (Kiyosawa *et al*., 2003; Boi *et al*., 2004; Solda *et al*., 2008; Cock and Whitworth, 2007, 2010) etc. Depending on the reciprocal direction of the transcription, the overlaps can be classified into parallel, where both of the genes are transcribed in the same direction and antiparallel where the genes involved in the overlap are transcribed in the opposite direction. Antiparallel overlaps can further be divided into convergent, where the genomic overlap is formed by the 3' termini of both of the genes and divergent, where the genomic overlap is formed by the 5' termini of the genes (Rogozin *et al*., 2002; Fukuda *et al*., 2003).

According to the overlap extent, the overlapping genes can be classified into complete, when the sequence of one of the gene involved in the overlap occurs entirely within the sequence of the other gene and partial where the only a part of the sequences of both genes overlap (Boi *et al*., 2004; Solda *et al*., 2008). In eukaryotes, complete overlaps can further be divided into nested overlaps, where the entire sequence of one of the gene lies within an intron of the other and embedded overlaps, where more than one intron or exon is shared. Another system classifies the overlaps according to the regions involved in the overlap viz., 5'untranslated region (UTR), 3' untranslated region (UTR), coding sequence (CDS) and introns (Boi *et al*., 2004; Kiyosawa *et al*., 2003). The classification of the overlaps as given in the article published by Solda *et al*. (Solda *et al*., 2008) and the extent of overlap in the different kinds of overlaps are given in Figure 2.1.

2.1.4 Strand and phase bias by overlapping genes

Majority (~84%) of known overlapping genes in eukaryotes lie on opposite strands with an antiparallel arrangement (Boi *et al*., 2004). Studies on overlapping genes in pairs of human–mouse ortholog genes also shows that majority (~90%) of the overlaps are different strand overlaps and the evolutionary transition between overlapping and non-overlapping genes are mainly caused due to higher rates of evolution in the untranslated regions, mainly in the 3' untranslated regions (Sanna *et al*., 2008). Evidence from eukaryotes shows that antiparallel overlaps, by the means of double stranded RNA formation, exerts roles in different biological processes, including

transcription, RNA editing, mRNA splicing and stability, and translation (Vanhee-Brossollet and Vaquero, 1998; Kumar and Carmichael, 1998).

Korbel *et al.* states that if the genes involved in the overlap are transcribed divergently with conserved gene orientation, they must be strongly co-regulated (Korbel *et al.*, 2004). On the other hand, the pattern is opposite in prokaryotes in which the same strand overlaps form the majority of the overlaps. Unidirectional overlaps are the most conserved in prokaryotic genomes and as the conserved operons strongly indicate functional associations, links can be predicted between all the conserved overlapping gene pairs (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999; Johnson and Chisholm, 2004). More opposite strand overlaps in eukaryotes might be due to the presence of complex gene structures with introns and because of the role the genomic overlaps have in the regulation of gene expression. Phase describes the shift in the reading frame between the adjacent genes (Rogozin *et al.*, 2002). Depending on which codon positions face each other in an overlap, the effects of DNA mutations on the two participating genes can be different. For each type of overlap, there can be three distinct phases, except for unidirectional overlaps. In prokaryotes, same strand overlaps occur mostly in the +1 and +2 reading frames, whereas different-strand overlaps are evenly distributed in the three reading frames (Johnson and Chisholm; Lillo and Krakauer, 2007; Cock and Whitworth, 2007, 2010)

2.1.5 Databases of overlapping genes in prokaryotes

There are a variety of databases that contains information about the overlapping genes in prokaryotes. The interactive database server called BPhyOG (Luo *et al.*, 2007) is used for reconstructing the phylogenies of completely sequenced prokaryotic genomes based on their shared overlapping genes. OGtree (Jiang *et al.*, 2008), a web-based tool, can also be used for the genome tree reconstruction of some prokaryotes based on the distance between the overlapping genes, viz., (1) overlapping gene content, the normalized number of shared orthologous OG pairs and (2) overlapping gene order, the normalized OG breakpoint distance (Jiang *et al.*, 2008). OGtree2, a newer version of OGtree, uses gene order distance to account for the genomic rearrangements and regulatory regions for defining the overlapping gene (Cheng *et al.*, 2010) and can reconstruct genome trees more precisely. PairWise Neighbours (Pallejà et al., 2009) is

another database which contains information about the spacers and overlapping genes among bacterial and archaeal genomes and their conservation across species. It permits the reliability analysis of the overlapping genes by taking into account the presence and location of the regulatory signal, Shine-Dalgarno (SD) sequence, and the ribosomal binding site in mRNA located 8 base pairs upstream of the start codon, among the adjacent prokaryotic genes, responsible for the efficient translation.

2.1.6 Significance of genomic overlaps

Studies about the overlapping genes in the genomes lead to different conclusions about their existence in the genomes. Overlapping genes play a major role in the genome size minimisation (Sakharkar *et al*., 2005), gene expression regulation through translational coupling (Normark *et al*., 1983; Chen *et al*., 1990; Inokuchi *et al*., 2000; Johnson and Chisholm, 2004), slowing down of evolution of the genes as a change in the overlapping part changes both of the proteins translated (Miyata and Yasunaga, 1978; Keese and Gibbs, 1992; Krakauer and Plotkin, 2002) etc. Considerable numbers of shorter unidirectional overlapping genes occur in the same transcriptional unit as genes within operons have shorter intergenic regions depicting the role overlapping genes play in the transcriptional coupling (Moreno-Hagelsieb and Collado-Vides, 2002). In *E. coli* and *Bacillus subtilus*, same strand overlaps occur within operons with twice the frequency than in the whole genome (Fukuda *et al*., 2003). The longer overlaps in the complementary strands can reduce the amount of nucleotides for coding two or more proteins (Lillo and Krakauer, 2007). Novel genes formed and their products may show structural similarities that may reflect common origin or evolutionary convergence (Keese and Gibbs, 1992).

As a result of the study conducted on 13 γ-Proteobacteria genomes, it has been proved that overlapping genes can be used as rare genomic markers to get insight into the phylogeny of the completely sequenced microbial genomes (Luo *et al*., 2006). The studies on the natural anti-sense transcripts (NATs) in prokaryotic cells (Lacatena and Cesareni, 1981; Itoh and Tomizawa, 1980) have aided to understand and interpret the role genomic overlaps play in the cell regulation. Natural anti-sense transcripts are endogenous transcripts showing complementarity to the sense transcripts with a known function. The role of complementary transcripts in regulatory processes

including transposition, plasmid replication and transcription etc. had been demonstrated in bacteria (Wagner *et al*., 2002). Most antisense RNAs are posttranscriptionally acting inhibitors of target genes, but a few examples of activator antisense RNAs are known (Wagner *et al*., 2002).

Studies on antisense transcripts in the human genome found that most of the genomic overlaps occur in the 5' or 3' exons which contain untranslated regulatory regions of mRNAs and hence sense antisense overlap may be associated with gene regulation and these areas are highly conserved (Yelin *et al*., 2003; Lapidot and Pilpel, 2006). The information about the antisense transcription helps in the study of the RNA interference (Bosher and Labouesse, 2000) that causes gene silencing, in selecting synthetic antisense oligonucleotides in functional studies and drug design (Delihas *et al*., 1997; Yelin *et al*., 2003) etc. In yeast, out of the newly discovered 137 open reading frames, 79 non-annotated and expressed open reading frames were found to occur opposite the previously annotated genes (Kumar *et al*., 2002). Non-coding antisense transcripts have roles epigenetic mechanism including RNA editing, DNA methylation, histone modification, genomic imprinting etc and studies on these natural antisense transcripts can help to explore therapies for diseases including cancer (Su *et al*., 2010).

The analyses of the proteins coded by the manually curated overlapping genes from 43 genera of unspliced RNA viruses infecting eukaryotes shows that the protein composition is biased towards disorder-promoting amino acids with more structural disorder than nonoverlapping proteins (Rancurel *et al*., 2009). In these viruses, some of the overlapping proteins were created *de novo* and were present only in certain species or genus with a role in the viral pathogenity or spread and no role in the viral replication or structure. Some of the novel proteins predicted to have ordered structures had novel folds (Rancurel *et al*., 2009). Comparison of amino acid composition revealed an increased frequency of amino acid residues with a high level of degeneracy (arginine, leucine, and serine) in the proteins encoded by overlapping genes which can be viewed as a way to expand their coding ability and gain new specialized functions (Pavesi *et al*., 1997).

Overlapping genes are present in almost all of the sequenced microbial genomes and it has been estimated that a third of the microbial genes are overlapping (Fukuda *et al*., 2003; Johnson *et al*., 2004). A strong correlation has been found between the total number of genes and the total number of overlapping genes that are present in the genome of an organism (Fukuda *et al*., 2003; Johnson and Chisholm, 2004). On the other hand, another study by Sakharkar *et al*. on prokaryotes showed that the numbers of genomic overlap are influenced by the niche and life style of the microbes (Sakharkar *et al*., 2005). Obligatory intracellular parasites and facultative intracellular parasites show a higher proportion of overlapping genes compared to the free-living ones causing to genome reduction and having larger amount of pseudogenes.

Overlapping genes are more conserved between the species than the non-overlapping genes as the mutation in the overlapping parts causes changes in both of the genes involved in the overlap (Krakauer, 2000; Sakhar *et al*., 2005). Hence selective forces against such mutations will be stronger and therefore the overlapping genes can be used as phylogenetic markers or characters for the evolutionary tree reconstruction among bacterial genomes (Luo *et al*., 2006; Luo *et al*., 2007) by using the normalized number of shared orthologous overlapping gene pairs as the distance measure.

2.3 Codon usage

2.3.1 The genetic code and its origin

The genetic code triplets are made up of the four nucleotides, viz., adenine, guanine, cytosine, uracil, which were envisioned to account for the 20 amino acids. The genetic code is sustained in all the living things with a few reassignments (Knight *et al*., 2001) and is hence assumed to be nearly universal. There are different concepts on the origin and evolution of the code which are not mutually exclusive. The frozen accident hypothesis (Crick *et al*., 1961) states that the standard genetic code was fixed as all living things sharing a common ancestor, with the codons changing subsequently without adding deleterious effects of codon reassignment. The major theories on the origin and evolution of genetic codes are the stereochemical theory (Dunnill, 1966; Han *et al*., 2010), coevolution theory (Wong, 1975, 2005) and the error minimisation theory (Alff-Steinberger, 1969; Shah and Gilchrist, 2010). Codons are assigned by the physicochemical affinity between amino acids and the anticodons according to the

stereo chemical theory, codons are coevolved along with the amino acid biosynthesis pathways according to the coevolution theory and codons are evolved to minimize the adverse effect of point mutations and translation errors according to the error minimization theory. Further studies on the evolution of the genetic code and structural analyses shows that the genetic code is robust to translational misreading and might have arisen based on the combination of frozen accident hypothesis and selection that minimizes translation error (Higgs, 2009; Koonin and Novozhilov, 2009; Sammet *et al*., 2010).

2.3.2 Variations in the standard genetic code

The variations in the standard genetic code were discovered since an alternative genetic code used by the human mitochondrial genes was found. Now, there are 16 additional alternative genetic codes available in the NCBI's taxonomy database (Sayers *et al*., 2010). The different genetic codes are given in the Table 2.1. The translation table number 11 is used by plastids of the bacteria, archaea and plant and is given in the Figure 2.2. Although with the translation table 11, translation initiation is most efficient at AUG, two additional codons GUG and UUG, can also serve as start in archaea and bacteria (Kozak, 1983; Golderer *et al*., 1995; Sazuka and Ohara, 1996; Wang *et al*., 2003). The codon UUG can act as initiator codon for around 3% of the bacterium's proteins in *E. coli* (Blattner *et al*., 1997) and the codon CUG can function as an initiator for one plasmid-encoded protein, RepA (Spiers and Bergquist, 1992). Exceptional cases are there where the bacteria use AUU, in addition to NUG as the translation initiation codon (Polard *et al*., 1991; Binns and Masters, 2002). UGA codes less efficiently for tryptophan in *Bacillus subtilis* and, presumably, in *E. coli* (Hatfield and Diamond, 1993), but the internal assignments are the same as the standard code.

The translation table number 4 differs from the standard genetic code in that the termination code UGA in the standard code is used to code tryptophan. There are alternative initiation codons including UUA, UUG, CUG for Trypanosoma; AUU, AUA for Leishmania; AUU, AUA, AUG for Tertrahymena and AUU, AUA, AUG, AUC, GUG and possibly GUA for Paramecium. The table number 4 is used for bacteria, fungi, some eukaryotes and metazoan and is given in Figure 2.3. In bacteria, the code is used by *Entomoplasmatales* and *Mycoplasmatales.* The codon

reassignment of UGA from stop codon to tryptophan is found in an α-proteobacterial symbiont of cicadas: *Candidatus hodgkinia cicadicola* (McCutcheon *et al*., 2009).

Table 2.1. The standard genetic code and the alternative genetic codes along with their respective translation table numbers.

Type of genetic code	***Translation table number***
Standard code	1
Vertebrate mitochondrial code	2
Yeast mitochondrial code	3
Mold, Protozoan, and Coelenterate mitochondrial code and the Mycoplasma or Spiroplasma code	4
Invertebrate mitochondrial code	5
Ciliate, Dasycladacean and Hexamita nuclear code	6
Echinoderm and flatworm mitochondrial code	9
Euplotid nuclear code	10
Bacterial, Archaeal, and Plant plastid code	11
Yeast alternative nuclear code	12
Ascidian mitochondrial code	13
Alternative flatworm mitochondrial code	14
Blepharisma nuclear code	15
Chlorophycean mitochondrial code	16
Trematode mitochondrial code	21
Scenedesmus obliquus mitochondrial code	22
Thraustochytrium mitochondrial code	23

TTT	F	Phe		TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe		TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu		TCA	S	Ser	TAA	*	Ter	TGA	*	Ter
TTG	L	Leu	i	TCG	S	Ser	TAG	*	Ter	TGG	W	Trp
CTT	L	Leu		CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu		CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu		CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu	i	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile	i	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile	I	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile	I	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met	I	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val		GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val		GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val		GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val	i	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

Figure 2.2. The bacterial, archaeal and plant plastid code (translation table number 11).

TTT	F	Phe	TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe	TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu i	TCA	S	Ser	TAA	*	Ter	TGA	W	Trp
TTG	L	Leu i	TCG	S	Ser	TAG	*	Ter	TGG	W	Trp
CTT	L	Leu	CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu	CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu	CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu i	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile i	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile i	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile i	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met i	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val	GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val	GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val	GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val i	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

Figure 2.3. The mold, protozoan, and coelenterate mitochondrial code and the mycoplasma/spiroplasma code (translation table number 4).

2.3.3 Codon usage bias

All of the 20 different amino acids, except methionine and tryptophan, are encoded by more than one codon. The codons that produce same amino acids are called synonymous codons and they are not used with equal frequency leading to codon usage bias. The patterns in the codon usage have been analysed since the collation efforts of the first molecular sequence databases (Grantham *et al*., 1981). The studies on the patterns of codon usage by the synonymous codons demonstrate that the genes within a species show similar patterns of codon usage as stated by the genome hypothesis (Grantham *et al*., 1980; Grantham *et al*., 1981; Gouy and Gautier, 1982; Ikemura, 1985; Sharp and Li, 1986; Aota and Ikemura, 1986). Hence, summing up the patterns of all the genes in an organism to get the codon usage of the organism may conceal the underlying heterogeneity (Aota *et al*., 1988) and it is better to specify the codon usage trends among the genes in a species and between closely related species.

2.3.4 Causes of codon usage bias

A variety of causes and consequences of the variability in the codon usage have been identified (Sharp and Cowe, 1991). Variation in codon usage occurs due to one or a combination of the several factors including the translational selection (Grantham *et al*., 1981; Sammet *et al*., 2010), replication-transcriptional selection (McInerney, 1998), mutational bias (Levin and Whittome, 2000; Gupta et al., 2004) etc. Other

biological phenomena associated with the codon usage bias are the gene expression level (Gouy and Gautier, 1982; Roymondal *et al*., 2009), gene length (Bains, 1987), translation initiation signal of the gene (Ma *et al*., 2002), amino acid composition of the protein ((Lobry and Gautier, 1994), structure of the protein (D'Onofrio *et al*., 2002; Gu at al., 2004), abundance of the trna (Ikemura, 1981) and the GC compositions (Bernadi and Bernadi, 1986; Karlin and Mrazek, 1996; Knight *et al*., 2001; Wan *et al*., 2004).

2.3.5 Calculation of codon usage bias

Codon usage bias can be evaluated using methods based on the statistical distributions and methods based on comparing the codon usage to that of the optimal codons (Bennetzen and Hall, 1982). Methods based on the statistical distributions include codon-usage preference bias measure based on the χ^2 and scaled χ^2 analyses (McLachlan *et al*., 1984; Shields and Sharp, 1987). Another method based on the Shannon informational theory, called synonymous codon usage order, measures the synonymous codon usage bias within a genome and across the genomes (McLachlan *et al*., 1984; Bernadi and Bernadi, 1986). Codon usage bias can also be analysed using the multivariate analysis, correspondence analysis followed by the cluster analysis (Angellotti *et al*., 2007; Sharp *et al*., 1986). Correspondence analysis finds trends in the data and distributes the genes in a genome or genomes of different species according to the trends along the axes and cluster analysis partitions the data base on the trends within the data. PCA, another multivariate analysis method was also used to investigate the heterogeneous codon usage (Grantham *et al*., 1980; Kanaya *et al*., 1996a; Kunst *et al*., 1997; Kanaya *et al*., 1999). These multivariate methods are applied usually on relative codon usage frequencies instead of the absolute codon usage frequencies to ward off the biases in gene length and amino acid usage, which make the variation in synonymous codon usage unrecognizable. Another method that has been used to measure the codon usage bias patterns is the unsupervised artificial neural network, SOM (Wang *et al*., 2001; Supek and Vlahovicek, 2004).

2.3.6 Codon usage database

Codon usage database (Nakamura *et al*., 2000) is an extended web version of the codon usage tabulated from GenBank (Nakamura *et al*., 1997; Benson *et al*., 2010),

denoted by CUTG. The database provides information about the frequencies of the codon usage in different organisms, searchable through the web, complied from the taxonomical divisions of the GenBank sequence database. The database has been updated with the nucleotide sequences obtained from the NCBI-GenBank Flat File Release 160.0 and presently houses codon usage tables for 35,779 organisms obtained from completely sequenced protein coding genes (coding sequences), avoiding codons with ambiguous bases. Codon usage for the organisms and for each of the genes in an organism is available through the web site http://www.kazusa.or.jp/codon/ either in the codon frequency compatible format or in the traditional table format, helping to analyse variations in codon usage among different genomes.

2.3.7 GenBank

GenBank (Benson *et al.*, 2010) is a comprehensive and publicly available nucleotide sequence database maintained by the National Center for Biotechnology Information, NCBI, at the National Institute of Health and was formed in 1988. GenBank receives nucleotide sequences primarily from the scientific community and from the daily data exchange through International Nucleotide Sequence Database Collaboration (INSDC), a combined effort by DNA Databank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) and GenBank which ensure uniform and comprehensive collection of sequence information worldwide. GenBank can be accessed through the NCBI Entrez retrieval system, the bi-monthly releases and daily updates. The updates of the GenBank database and the complete genomes can be accessed via anonymous ftp from NCBI at ftp.ncbi.nih.gov/genbank. GenBank releases are distributed in the flat file format as well as in the Abstract Syntax Notation 1 (ASN.1) version for internal maintenance at the NCBI's anonymous ftp server.

2.3.8 Taxonomy database at NCBI

Taxonomy database at NCBI attempts to incorporate phylogenetic and taxonomic knowledge from a wide range of sources viz., information from the published literature, web-based databases, information from the sequence submitters, taxonomic experts outside NCBI etc. The consistent taxonomy that is provided in the NCBI's sequence databases (Sayers *et al.*, 2010) is taken from the Taxonomy database. It comprises the names and lineages of all the organisms for which at least one

nucleotide or protein sequence is available in the NCBI's sequence databases. NCBI taxonomy database serves as the standard reference for the INSDC. The database provides links to all data for each taxonomic node from super kingdom to subspecies. The taxonomy browser can be used to view the taxonomic position or retrieve data from any of the Entrez databases for a particular organism or group. The database is also available as files from the ftp site (ftp://ftp.ncbi.nih.gov/pub/taxonomy/) and as a domain of Entrez, (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy), which will be updated every two hours.

2.4 Multivariate analyses for detecting codon usage bias

2.4.1 Principal component analysis

Principal component analysis (PCA) is an unsupervised statistical learning method used for transforming high dimensional data to lower dimensional space (Morrison, 1976; Jolliffe, 2002), preferably two dimensional. It can be used for exploring and generating consistent patterns within the data. It has applications in wide range of research areas including image processing, genomic analysis, information retrieval, codon usage bias analysis (Kanaya *et al.*, 1999) etc. It describes the structure of the high dimensional data by reducing its dimensionality to the uncorrelated principal components, which can explain the variation in the data (Morrison, 1976). The first and second principal components account for as much variability in the data as possible. Further components account for smaller amounts of the residual variation. The plot of the PCA with the first principle component versus the second, shows distribution of the points on a principal plane where x-axis corresponds to point projection on the first principal component and y-axis to the projection on the second principal component.

2.4.2 Correspondence analysis

Correspondence analysis (Greenacre, 1984) is a method to quantify categorical data by assigning numerical scale values to the response categories of discrete variables, with certain optimal properties. These scale values have been shown to have interesting geometric properties and provide maps of the relationships between variables. Total variance is measured by the inertia, the Pearson χ^2 statistic calculated on the cross-tabulation divided by the total sample size. The working principle is similar to the

principal components analysis (PCA), the total variance of the data is decomposed optimally along the principal axes. Each principal axis accounts for a certain amount of the total inertia and the first and the second principal axes account for a large percentage of total variance which allows the data to visualise in two dimensions. The correspondence analysis is based on three basic concepts viz., (1) the point in multidimensional space, (2) the weight associated with the point and (3) the χ^2 distance, which is the distance function between the points. The correspondence analysis projects the points into a low dimensional subspace, usually two-dimensional plane that optimally fits the points by χ^2 distance and hence causes dimension reduction. The two different ways of mapping the columns along with the rows are the asymmetric map and the symmetric map. In asymmetric map, the row profiles are depicted by principal coordinates, and the column points by the projections of unit profile vectors onto the same space, with the problem being that the column points are spread out more than the row points. Symmetric map represents the row points and column points in principal coordinates and each of them are projected in different spaces.

2.4.3 Self-organizing maps

The self-organizing map is a form of the artificial neural networks that was first proposed in the 1970's (von der Malsburg, 1973; Kohonen, 1982) and it works on the principle of unsupervised learning. It can be used for clustering and visualising the trends inherent to the problem. It creates a map of a set of high dimensional input vectors to a low-dimensional (one or two dimensional) grid through vector quantization (Kohonen *et al*., 1996). Self-organizing map uses a neighborhood function and preserve the topological properties of the input space (Kohonen *et al*., 1996). Self-organizing maps operate by training, when it constructs the map using input examples by vector quantization, and by mapping, when it classifies the new input vectors. It has higher resolving powers than the multivariate analysis method, PCA, when analysing genes from a large number of species simultaneously. Self-organizing maps were proved to be efficient for characterising horizontally transferred genes (Kanaya *et al*., 2001) and for analysing codon usage bias patterns (Wang *et al*., 2001; Supek and Vlahovicek, 2004).

2.4.4 Heat maps

The clustered heat map is one of the popular graphical representations that can display large amounts of data compactly to a smaller space visualising the coherent patterns in the data. Heat maps can reveal both the row and column hierarchical cluster structure from a data matrix. It is made up of rectangular tiling and each tile is shaded with a color scale according to the value of the corresponding element in the data matrix. They have been used since 1997 to display the expression patterns of microarray data, microRNA, protein, DNA copy number, DNA methylation, metabolite concentration, drug activity etc. two dimensionally (Eisen *et al*., 1998; Weinstein *et al*., 1997; Wang *et al*., 2006; Brauer *et al*., 2006). In the heat maps displaying the gene expression data, the color on the rectangular tiling is proportional to the expression of the RNA or protein. But, the heat maps may not be able to show the complex patterns of nonlinear relationship in the samples and the bifurcation of the cluster tree should be specified. The functional ordering of the axes, the coherent patterns generated and the meaning revealed by the clustered heat map depends on the choice of the preprocessing algorithm (data normalization, filtering etc.), which minimizes noise while keeping the meaningful signal; clustering algorithm, that decide the grouping of the data; the distance metric (Euclidean or correlation), that defines the measure of similarity; and the color scheme (linear, logarithmic, quantile, two-color, three-color etc.), that emphasizes the patterns to be visualised (Weinstein, 2008). Also the patterns differ according to the use of relative or absolute data to create a "difference" heat map. Hence different kinds of heat maps can be generated from the same experiment, each having its own visual meaning and therefore it is important to specify the parameters for interpreting the heat maps (Weinstein, 2008).

2.5 χ^2 test

Pearson's χ^2 test is a univariate test used for performing tests of goodness of fit and tests of independence (Agresti, 2002; Thompson, 2009). χ^2 test for the goodness of fit is used for evaluating whether an observed frequency distribution differs from a theoretical frequency distribution. The test of independence assesses whether a pair of observations on two variables, expressed in a contingency table, are independent of each other.

2.5.1 Test of goodness of fit

χ^2 test of goodness of fit is used for testing the validity of a distribution and evaluates the null hypothesis, H_0, which states that the data belongs to an assumed distribution against the alternative hypothesis, H_a, which states that the data does not come from the assumed distribution. In order to provide proofs for the hypotheses testing, the χ^2 test statistic, χ^2, is calculated. The χ^2 test statistic is calculated by using the equation,

$$\boldsymbol{\chi^2 = \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{E_i}}$$

Where,

χ^2 = Test statistic that asymptotically approaches a χ^2 distribution.

O_i = Observed frequency;

E_i = Expected frequency, given by the null hypothesis;

n =Number of possible outcomes of each event.

The χ^2 test statistic value obtained from the calculation can be compared to the χ^2 distribution for the p-value calculation. P-value is the relative standard used for determining whether the null hypothesis is to be rejected or not. It represents the probability that the deviation of the observed frequencies from the expected frequencies is due to chance alone. A p-value of 0.05 or less makes the null hypothesis to be rejected indicating that the data are not fitting to each other.

2.5.2 Test of independence

The χ^2 test of independence is used to test whether two outcomes of a single observation are statistically independent. Here, the null hypothesis, H_0, assumes that the two outcomes are statistically independent. The outcomes will be arranged in a two way contingency table comprising of r rows and c columns. Then, the value of the test-statistic is calculated by,

$$\boldsymbol{\chi^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{i,j} - E_{i,j})^2}{E_{i,j}}}$$

The χ^2 test of independence evaluates whether the variables within a contingency table are independent or not associated. The χ^2 statistic is calculated by summing up the squared difference between observed and expected data and dividing it by the

expected data in all possible categories. The expected frequencies are calculated using the multiplication rule of the probability and the degree of freedom is the number of independent variables in the data set. The value of the χ^2 statistic is compared with the appropriate χ^2 distribution. As the degrees of freedom increases, the χ^2 value required to reject the null hypothesis increases. A chi-square probability of less than or equal to 0.05 is commonly interpreted as justification for rejecting the null hypothesis (rows and columns are independent) and for accepting the alternative hypothesis, both the row and column variables are associated.

2.5.3 Residual analysis of the χ^2 test

The null hypothesis for the Pearson's χ^2 test of independence, if rejected, indicates an association between the row and the column variables. Residuals of the Pearson's χ^2 test can be followed up by residual analysis to assess where association lies. Residual analyses are usually performed with standardized residuals, which is the difference between observed and expected values in a cell divided by a standard error for the difference. Standardized residuals help to find the direction and strength of the association. A large standardized residual provides evidence of association in that cell, i.e., the observed number was greater for this category than it was expected. Standardized residuals with negative value indicate that the cells were underrepresented in the actual sample, i.e., the number of subjects in this category was fewer than the expected number.

3 Objectives

- Find the microbial species for which the completed genomes are available from the NCBI ftp site and for which the codon usage tables are available from the codon usage database.
- Investigate the prevalence of overlapping gene pairs in the selected prokaryotic genomes using annotations from the GenBank files
- Calculate the number of overlapping genes in each species.
- Get the codon usage frequency of the normal part of the genome from the CUTG (codon usage database), the codon usage tabulated from GenBank.
- Calculate the codon usage of the overlapping part of the genome.
- Compare the codon usage patterns and analyse whether there is any bias in the codon usage in the overlapping part compared to the normal parts.
- Find the association between the codon usage bias patterns and the different taxonomy levels of classification.

4 Materials and methods

4.1 Data Source

4.1.1 Selection of Prokaryotes

Prokaryotic species needed for the analyses were chosen based on two factors viz., the availability of the completed genomes in the NCBI's (Sayers *et al*., 2010) ftp site and the availability of the codon usage table from the codon usage database (CUTG) (Nakamura *et al*., 2000). The GenBank identifiers of all the prokaryotic species whose completed genomes were available from http://www.ncbi.nlm.nih.gov/Ftp/, the NCBI ftp site, were collected and the selected only those for which the codon usage tables were available from the codon usage database.

4.1.2 GenBank

GenBank (Benson *et al*., 2010) is a comprehensive database of genetic sequences and it provides annotated collection of all the publicly available DNA sequences. It was created and is maintained and distributed by the National Center for Biotechnology Information, a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), formed in 1988. Sequence data are submitted to GenBank by individual scientists around the world, by the larger sequencing centers involved in major sequencing projects and by sharing of the data by the International Nucleotide Sequence Database Collaboration. Genomic GenBank files of the various completed prokaryotic genomes from the NCBI ftp were used for finding the overlapping genes.

4.1.3 Codon usage database

Codon usage bias by the overlapping part of the coding sequences in the genomes of the species under study is calculated by comparing the frequencies of the codons from the overlapping part to the frequencies of the codons from the normal regions. Codon frequencies of the normal part of the genomes were collected from the database named the codon usage database, CUTG, http://www.kazusa.or.jp/codon/, (Nakamura *et al*., 2000), which gives the codon usage frequencies for different species tabulated from the GenBank. Codon usage tables from the CUTG database for all the species chosen for the study were extracted using Perl scripts.

4.1.4 NCBI Taxonomy database

The taxonomy database at NCBI is one of the databases supported and distributed by the NCBI for the medical and scientific communities. The database incorporates phylogenetic and taxonomic knowledge from the published literature, web-based databases, and advice of sequence submitters and outside taxonomy experts. The taxonomic position can be visualised or the data from any of the principal Entrez databases for a particular organism or group can be retrieved using the taxonomy browser.

4.2 Methods

4.2.1 Data generation using Perl scripts

The GenBank identifiers of the microbial species for which the completed genomes were available from the NCBI ftp were collected manually. A total of 6 different Perl scripts that extensively uses the BioPerl modules (Stajich *et al*., 2002) were designed for generating the data that was needed to analyse the codon usage bias by the codons of the overlapping part of the genomes in the selected species. All of the scripts were designed to work with the GenBank identifiers of the species as the input. *E. coli* was used as the test organism for running the scripts.

4.2.1.1 Download.pl

This Perl script took GenBank identifier as input and extracted the scientific name of the species for the identifier using the BioPerl module Bio::DB::Taxonomy and searched for the existence of the codon usage table for that species from the CUTG using the BioPerl interface to the CUTG, Bio::CodonUsage::Table. If the codon usage table was available for a species, the script created a directory named after the organism's GenBank identifier. Two subdirectories were made inside this main directory, one with the name of the GenBank identifier and the other one with the name, `target`. The first subdirectory was for writing (1) the genomic file of the species, (2) codon usage table of the species named `CodonUsageTable` and (3) `Results.txt`, the final result file for the species with information about their overlapping parts needed for further analyses. The second subdirectory was for writing

the GenBank formatted files, one for each of the overlapping part in the genome of that particular species. The codon usage tables were downloaded and written locally using the BioPerl modules Bio::CodonUsage::Table and Bio::CodonUsage::IO respectively. GenBank files were downloaded using the BioPerl's GenBank interface, Bio::DB::GenBank. The algorithm for the Perl script is given below and the workflow is given in the Figure 4.1.

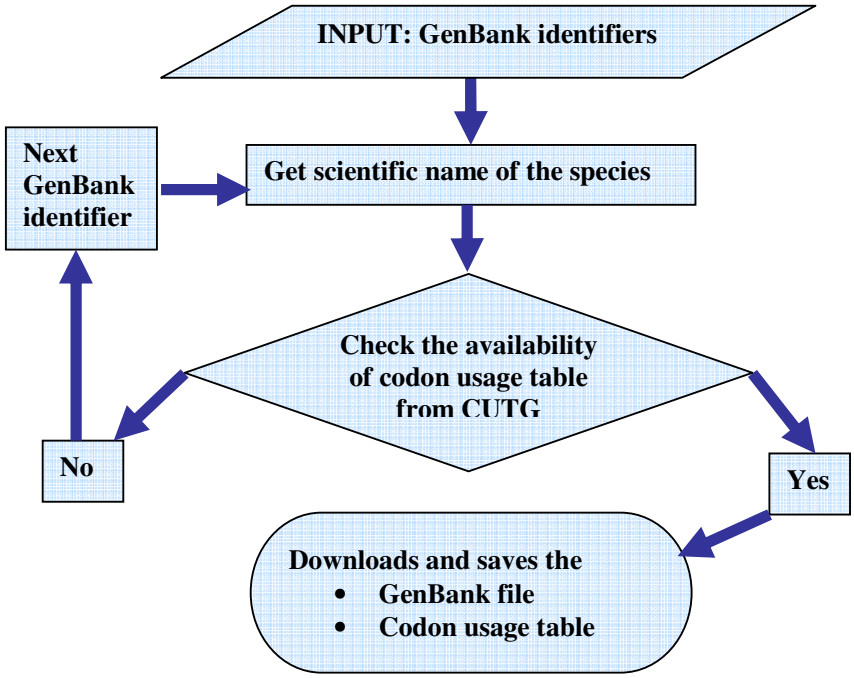


Figure 4.1. The workflow of the Perl script Download.pl mentioned in the section 4.2.1.1.

The algorithm

- Get the species name for the GenBank identifier from NCBI.
- Check for the availability of codon usage table from CUTG.
- If the codon usage table is available, create a folder for the species and name it using GenBank identifier.
- Create two subfolders in the above folder. The first one named after GenBank identifier is for writing the genomic GenBank file, codon usage table and the final result file for each of the species. The other one, named `target`, is for writing the GenBank files for the overlap part in the genome.

- Download the codon usage table from CUTG and the genomic GenBank file from NCBI GenBank and write them to the subfolder named after the GenBank identifier.

4.2.1.2 OverlapFind.pl

GenBank identifiers were given as input for the script. The script finds overlapping gene pairs in the genomes. Coding sequences with introns that are rare in prokaryotes, evident from the split location of the coding sequences, were not taken into account. The `FEATURE` section of the GenBank formatted genomic files contains information on the genes, its products, other biologically relevant regions of significance, as well as a number of other features. The details of a feature were extracted from the `Feature key`, the name of feature, under which the details of the feature are stored. One of the features under this section is `CDS`, the part of the genomic sequence that codes for proteins. The base span of the `CDS` is specified from the position of the start codon to that of the stop codon and can be complete (1..206), partial on the 5' end (<1..206) and partial on the 3' end (4821..5028>). The base span is described in different ways in different situations. It can be

- single base e.g. (p)
- continuous span of bases that occurs in prokaryotes where there is not any intron indicated by e.g. (a..c)
- joining a continuous span of bases when introns are there - join (a..b, c..d, e..f).

The coding sequences located on the complementary strand are indicated by prefix `complement` before the base span. The feature data in the genomic files were converted to `Bio::SeqFeature::Generic` objects using BioPerl. Location and other details of the `CDS` were extracted from these objects. Genomic files were parsed, and the location and the strand information of consecutive `CDS`s were extracted. If the start of the second coding sequence was before the end of the first coding sequence, a GenBank formatted file was written for the overlapping part to the subfolder `target` with the name `OverlapStart_OverlapEnd.gb`. The sequence of the overlapping part was converted to a `Bio::Seq` object, to which three primary tags and their features were added. The primary tags created for the overlapping part were

OVERLAP_1, CDS1 and CDS2. The features that were added to the primary tag, OVERLAP_1, were its location and the start and the end position of the overlap. The features that were added to the primary tag CDS1 and CDS2 were

- location
- strand information
- feature tags and their values available for the coding sequences from the genomic GenBank file
- sequences of the coding sequences from the original genomic sequence according to the location.

Length of overlap is the base span from the start position of the second CDS to the end of the first CDS, when the second CDS is longer than the first CDS as given in the Figure 4.2. If the second CDS is entirely within the first, the overlap part will span the entire second CDS as in the Figure 4.3. The script used two subroutines as mentioned below and the workflow for the Perl script is given in the Figure 4.4.

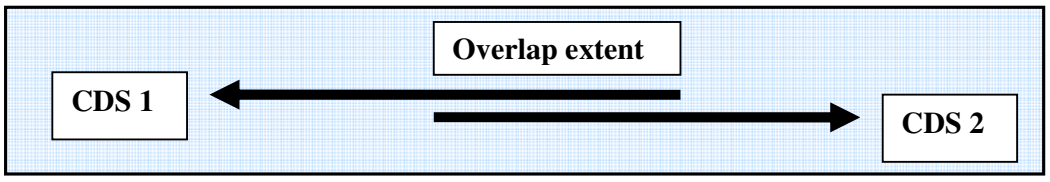


Figure 4.2. Extent of the overlap in partial overlap.

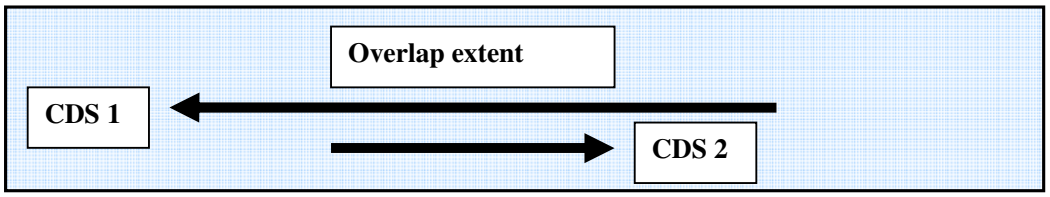


Figure 4.3. Extent of overlap when one of the coding sequences occurs entirely within the other.

- **File_parse**

  It converts the genomic GenBank file of the species to a Bio::Seq object, the sequence along with its features from the original genomic GenBank file.

- **Write_file**

  Writes a GenBank formatted file for the overlaps. It creates a new Bio::SeqIO object from the sequence of the overlap part. In order to add

features, its primary tags, tags and their respective values to the Bio::SeqIO object, `Bio::SeqFeature::Generic` objects are created. A new file is created for writing the `Bio::SeqIO` object along with its features in the GenBank format with the file name `OverlapStart_OverlapEnd.gb`.

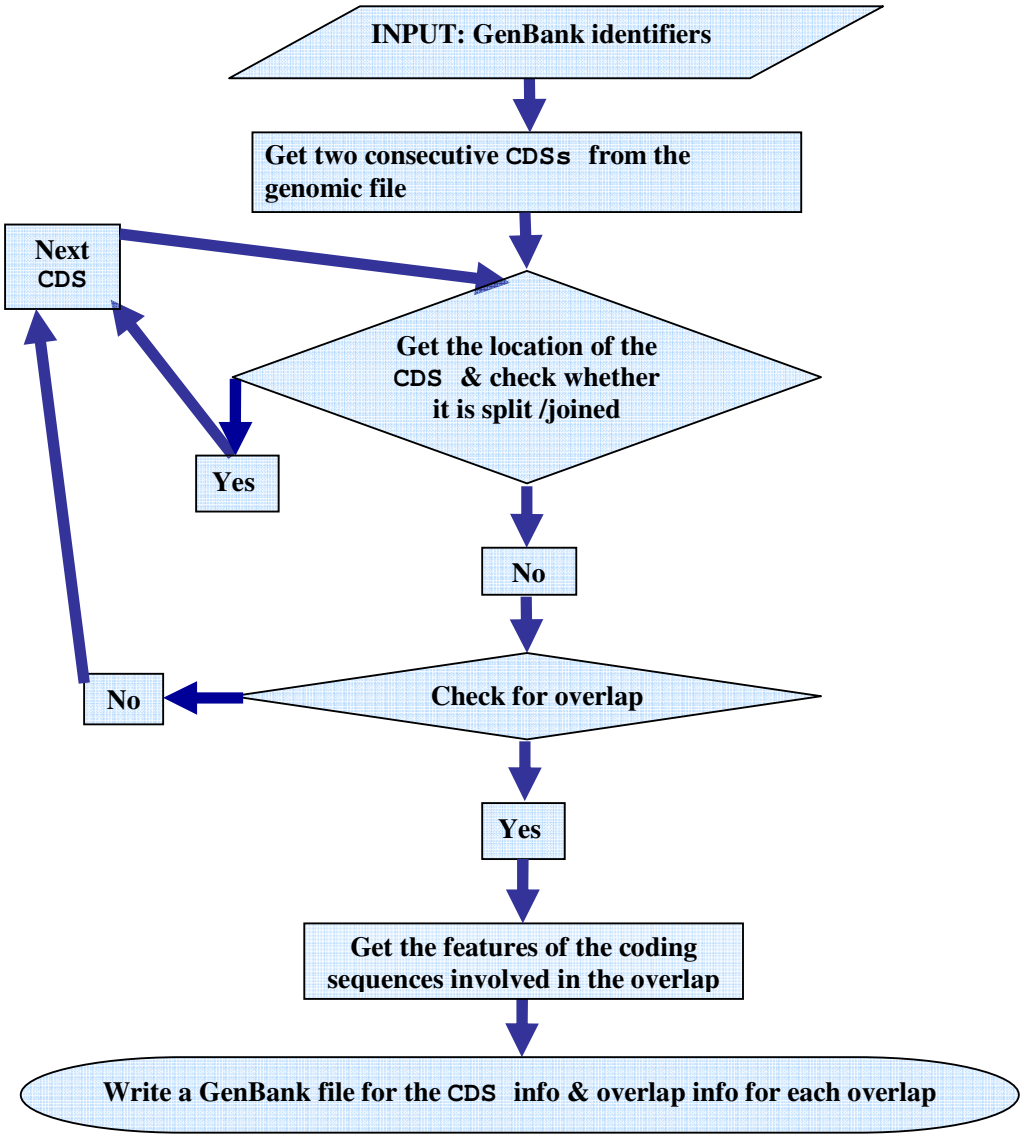


Figure 4.4. Workflow for the second script OverlapFind.pl which is described in the section 4.2.1.2.

The algorithm

- Parse the GenBank files, get consecutive `CDSs` and exclude those with split location.

- If there is overlap is between adjacent CDSs, write a GenBank formatted file, OverlapStart_OverlapEndposition.gb for the overlap part.
- The sequence of the overlapping part is used to create a sequence object to which the primary tags viz. OVERLAP_1, CDS1 and CDS2 with their features are added.
- For the primary tag OVERLAP_1, location is added as the feature.
- For the primary tags CDS1 and CDS2, feature tags and their respective values originally present in the GenBank file, strand information, and their sequences from the original sequence according to the location are added

4.2.1.3 AnalyseOverlap.pl

The Perl script took GenBank identifiers as input. A part of the analyses was done and a result file was generated which was further modified to serve as input for analyses using R (R Development Core Team, 2009). The program invoked R to do the χ^2 test. It parsed the GenBank formatted files for overlapping parts of the genome previously saved to the subfolder target. The threshold length was kept as nine nucleotides and only those overlaps with length greater than nine nucleotides were taken for the further analyses. The main program is assisted by five subroutines, viz., feat_obj, table, translation, phase and codon_count_func. For overlaps passing the threshold check, feature objects of the primary tags, CDS1 and CDS2 from the GenBank file of the overlapping part, were obtained by the subroutine feat_obj. The translation table numbers of the overlapping CDSs were obtained using the subroutine, table, and the translations of the CDSs using the subroutine translation. The reading frames of the coding sequences were obtained by the subroutine phase. These subroutines take as input

- strand information of the coding sequences
- seqob, a new Bio::PrimarySeq object created from the overlap sequence
- translation of the coding sequences got using the subroutine translation
- translation table number from the output of the subroutine table
- GenBank file for the overlap

The coding sequences from the plus-strand and those from the minus-strand after reverse complementing were translated using the Perl method `translate()` in three different frames with the retrieved translation table numbers. Stop codons were not taken into account. The translations obtained for the `CDS`s were checked for matching, using Perl method `index()`, to the translation already provided for the respective `CDS`s in the genomic file and the phase was calculated. The translation matching were done after taking into account of the fact that in translation table 11, the Bacterial, Archaeal and Plant Plastid Code, the codon triplets GUG and UUG can serve as the alternate start codons. The fifth subroutine, `codon_count_func`, counts the codons from the overlapping part using the method `count_codons()` from the BioPerl module `Bio::Tools::SeqStats`. Only the substrings of the overlapping parts corresponding to their phase were taken to count the codons. The codon counts were stored in a hash, `codon_count`, with codons as the keys and their counts as the values. The codons were counted from the overlapping part for both the coding sequences.

The other part of the script generated statistics for the overlap length distribution, determined the observed number of the codons from overlapping part, expected number of the codons from the codon usage table, did the χ^2 test to detect the difference in the observed and expected numbers of the codons and finally wrote a result file. The statistics were generated using the Perl functions `max()` and `min()` and methods from the `Statistics::Descriptive` module. The codons (keys) from the hash, `codon_count`, were sorted alphabetically and the number of codons (values of the keys) and codon usage frequency for these codons from the codon usage table were stored in the same order to the arrays `@prob1` and `@exp_count` respectively. The expected probabilities of the codons were estimated by dividing the expected counts of each codon by the total sum of the expected codon counts, both extracted from the codon usage table. R was invoked in Perl for doing the χ^2 test of goodness of fit to test the null hypothesis, "There is no significant difference between the expected frequency of codons from the codon usage table and observed frequency of the codons in the overlap part". Observed codon numbers stored to the array, `@prob1`, and the expected probability calculated were used for the χ^2 test. Finally a result file `Result.txt` was written for each species to the subfolder named after the

GenBank identifier with the following information, each on a new line, viz., GenBank identifier, species name, total number of codons from the overlapping part, overlaps with length above nine, minimum overlap length, maximum overlap length, median of the overlap lengths, mean of the overlap lengths, standard deviation of the overlap lengths, χ^2 statistic from the results of the χ^2 test, p-value from the results of the χ^2 test, codons along with their expected and observed numbers from the χ^2 test, observed numbers for the codons, observed relative percentages for the codons, divergence in percentage, divergence, five most biased codons along with their respective relative percentages keeping the expected as 100%, five least biased codons along with their respective relative percentages, keeping the expected as 100% and ambiguous codons along with their respective relative percentages, keeping the expected as 100%.

Subroutines:

- `feat_obj` : Get feature objects of for the coding sequences, `CDS1 & CDS2`
- `table` : Get the codon usage table id of `CDS1 & CDS2`
- `translation` : Get the value of the tag translation from `CDS1 & CDS2`
- `phase`: Calculate phase of `CDS1 & CDS2`
- `codon_count_func`: Count the codons from the overlap part

The algorithm

- Read the GenBank formatted files of the overlap part of the species.
- Check overlap length, select overlaps with length greater than 9 bases.
- Get the strand information
- Translate the `CDSs` in different phases and match it with the translations provided in the genomic file and get the frames of the `CDSs` .
- Count codons from the substring of the overlapping part corresponding to their phase (observed codon numbers) and store the counts.
- Read the codon usage table for the species, extract each codon's frequencies (expected codon numbers), store in an array and divide the array members by the total codon counts (expected probabilty).
- Do χ^2 test using R to test whether there is any significant difference between the expected and observed frequencies of the codons.

- Create a result file named `Result.txt` for each of the species with the details including the species name, GenBank identifiers, length distribution of the overlaps, results of the χ^2 tests, the bias showed by the codons etc.

The workflow for the Perl script is given in the Figure 4.5.

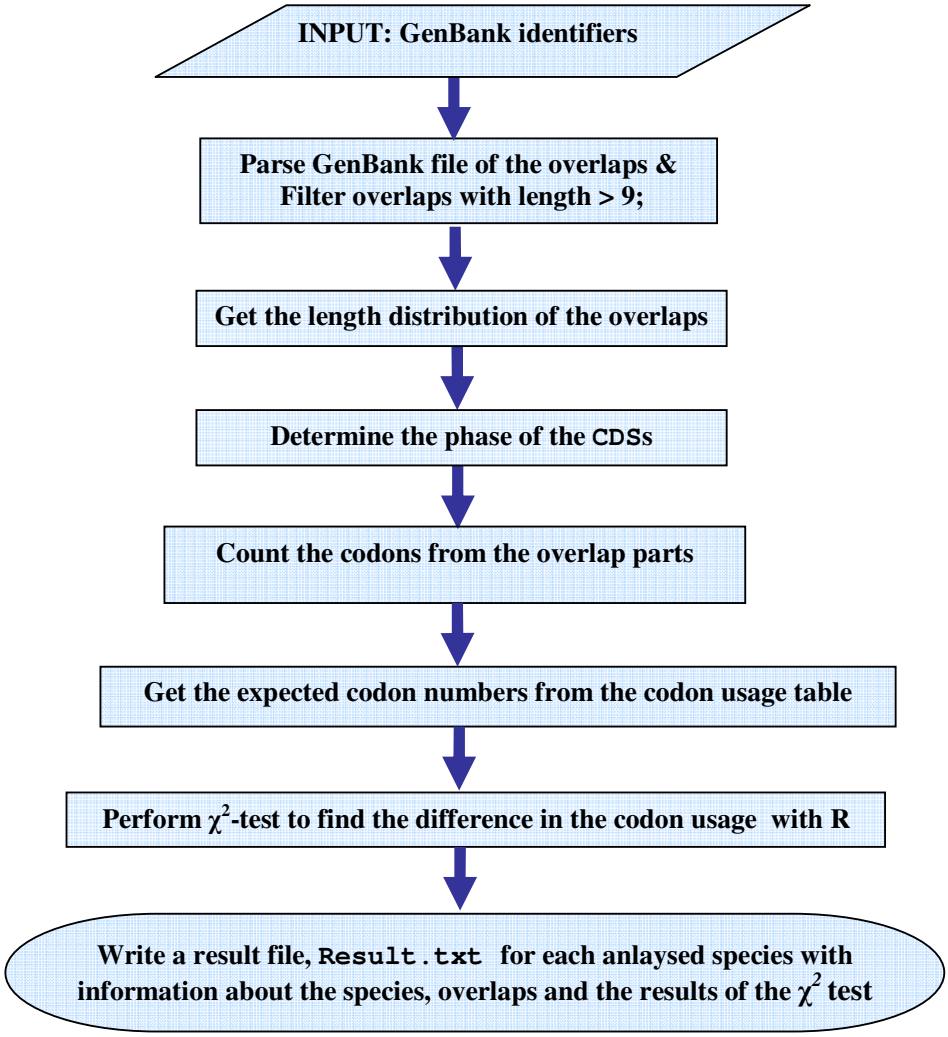


Figure 4.5. Workflow for the Perl script AnalyseOverlap.pl. It analyses the overlaps for all the genomes, does the χ^2 test to find whether the codon usage of the overlapping parts differs from that of the normal parts and writes a result file containing with all the information from the analyses for the analysed genomes.

4.2.1.4 CreateFinalText.pl

CreateFinalText.pl took GenBank identifiers as input and made a final text formatted tab separated file, `Final.txt`, with the information for all the analysed species on

length distributions of the overlaps, p-value from the χ^2 tests, biased codons and their percentages, ambiguous codons and the codons not used. The information needed for creating the `Final.txt` were extracted from the information already present in the `Result.txt` file created for each organism under study. `Final.txt` contains the following items as the header, each separated by a tab, viz., GenBank identifier, species name, minimum overlap length, maximum overlap length, median overlap length, mean overlap length, standard deviation of the overlap lengths, p-values as given by the χ^2 test, five least deviated codons and their comparative percentage, five most deviated codons and their comparative percentage and ambiguous codons and their numbers. The header line is followed by those details from all of the analysed microbes, information for each new species on a new line. The workflow for this script is given in the Figure 4.6.

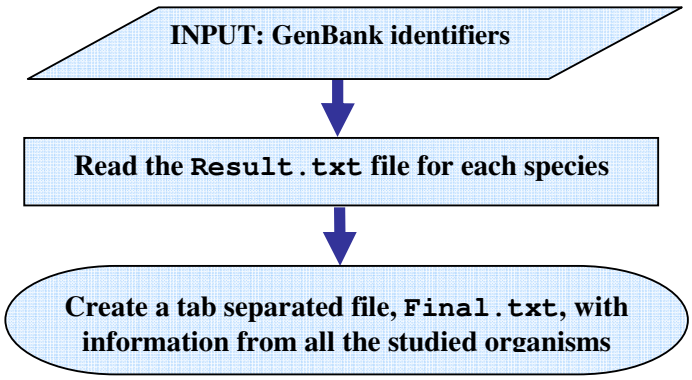


Figure 4.6. Workflow for the Perl script CreateFinal.pl. It creates a tab separated file, `Final.txt`, with all the information for the entire microbes to do further analyses.

4.2.1.5 FormatResult.pl

This script was made to produce a tab separated file with some chosen information from the `Result.txt` for each species to be given as input to the R for doing further analyses. GenBank identifiers of the studied microbes were given as input. `Result.txt` file for each species was read and observed numbers of the different codons along with their numbers in the order specified in the file were captured using Perl regular expression matching and were written to a tab separated text file named `Observed_Numbers.txt`. The observed relative percentages of codons for each species were also extracted from the `Result.txt` in the same order as the observed number of codons using the Perl regular expression match and were written to another

tab separated file, `Observed_Relative_Percent.txt`. The information about each new species was written to a new line. The workflow is given in the Figure 4.7.

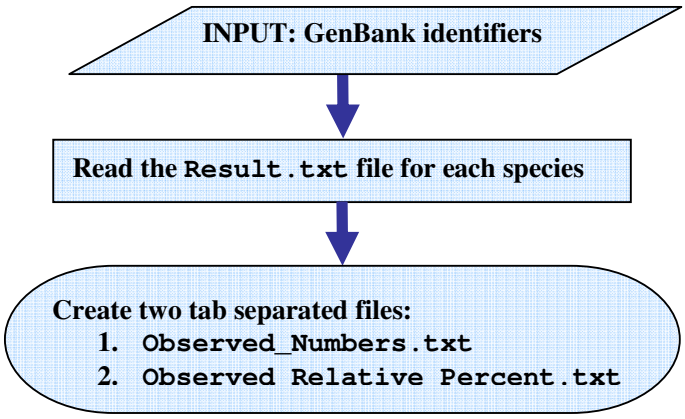


Figure 4.7. Workflow for the Perl script FormatResult.pl. Two formatted result files were created, `Observed_Numbers.txt` with observed numbers for the 64 codons and (2) `Observed_Relative_Percent.txt` with the relative observed percentages of the 64 codons, both for all the studied organisms.

4.2.1.6 Taxonomy.pl

The script Taxonomy.pl was created to retrieve the taxonomical classification at different levels for each of the species and the GenBank identifiers were given as input. The script parsed the genomic GenBank files for each species and extracted the taxon identifier. `Bio::DB::Taxonomy`, the BioPerl's interface to NCBI taxonomy database, was used to retrieve the rank names and scientific names of ancestor nodes in the taxonomy lineage for a given taxon id. These information were then written to a tab separated file, `Taxonomy.txt`, to collect the taxonomy information of all the species under study in an order specified by the script viz., Species, Genus, Family, Suborder, Order, Subclass, Class, Phylum, Super kingdom and no rank. Information about each species was written to a new line. The file `Taxonomy.txt` was used for further analyses using R. The workflow for this Perl script is given in the Figure 4.8. The details of all of the species that are used in the analyses viz., their species names and their respective GenBank identifiers, are organised into a table in the Appendix.

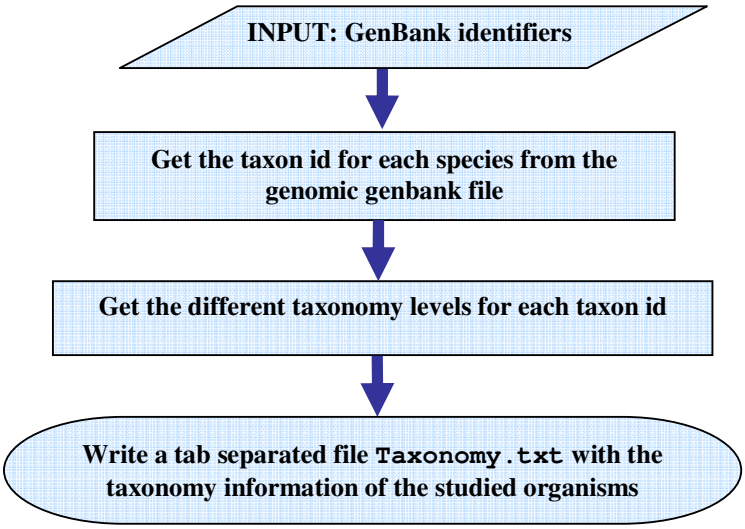


Figure 4.8. Workflow for the Perl script Taxonomy.pl. This Perl script extracted the taxonomy information at different levels for all of the organisms under study and wrote it to a tab separated file, `Taxonomy.txt`, which was used as the input to R.

4.2.2 Data analysis using R scripts

R (R Development Core Team, 2009), was used for generating plots to visualise the results of principal component analysis, correspondence analysis, self-organizing maps etc. Plots have also been generated using R to study the length distribution of the overlaps and for generating the statistics about the overlap length in the different genomes studied.

4.2.2.1 Descriptive statistics for the overlaps

R script was created to generate and visualise the descriptive statistics for the overlaps from a total of 435 different species from 16 different Phyla used in the study. The data needed for the statistical analyses were generated by the different Perl scripts mentioned in the methods section. The information saved to the formatted files generated earlier viz., `Final.txt` and `Taxonomy.txt`, were combined to generate table with the descriptive statistics for the overlap length and overlap numbers in the phylum level of taxonomy. Also plots visualising the distribution of the overlap numbers in the studied species, distribution of the overlap numbers in the different phyla and strand wise distribution of the overlaps and were generated.

4.2.2.2 Principal component analysis

R script was created for doing the principal component analysis using the method `prcomp()` from the package `stats`. The file, `Observed_percent.log`, a tab separated file with observed relative percentages of codons where, columns represent the codons and rows represent the different species used in the analysis, was used as input for the R script. The expected percentages from the codon usage table were kept as 100 % and the observed relative percentages of codons were calculated relative to this. The codons for which the expected number was zero and the observed number greater than zero were indicated by a -1. The standard deviations for the 64 different codons from the different species were calculated using the R function `sd()` and were plotted. PCA was performed by scaling as the codons TAA, TAG, TCA and TGC showed a large deviation. There were a total of 16 different phyla, to which the species used in this study belonged to. First principal component was plotted against second principal component. Three dimensional visualisation of the PCA plot was generated using methods from the library `scatterplot3d`. A PCA plot colored according to the phylum and separate PCA plots for each phylum with the same scales for the x and y axes were also generated to find trends, if any, in the codon usage bias of the overlapping part by the different phyla.

4.2.2.3 Correspondence analysis

R script was created using the function `ca()` from the package `stats` to do simple correspondence analysis and to generate symmetric and asymmetric maps from the correspondence analysis based on the singular value decomposition. The input was `Observed_percent_som.txt`, a two way table summarizing the distribution of the 64 codons in the 413 species. Simple correspondence analysis was applied to this 413×64 matrix and maps of the given table, comprising of the relative codon frequencies of each of the species, were constructed where each row and column was represented by a point.

4.2.2.4 Self-organizing maps

The R package `wccsom` that uses weighted cross correlation (WCC) similarity measure was used to generate the self-organizing maps for classification. The file,

`Observed_percent_som.txt`, was given as input to R. The observed relative percentages of codons were calculated relative to the expected percentages from the codon usage table, which was kept as 100 %. Only 413 out of the 435 species were included, as the overlapping part of the rest of the organisms showed codons that were not present in their codon usage tables. SOM grids with x and y ranging from 2 to 8 were analysed for the optimal correlation for classification according to taxonomy using the χ^2 test of independence and the best grid was selected for doing the SOM classication which was colored according to phylum.

4.2.2.4.1 χ^2 test

Table 4.1. The p-values obtained from the χ^2 test of independence of the groupings generated by the selected grid and the Phyla. The grid combination with the lowest P-value is high-lighted in cyan.

X	*Y*	*P-value*
2	2	1.872849e-28
2	3	6.116133e-21
2	4	5.212978e-35
2	5	4.205769e-3
2	6	3.1□3977e-36
2	7	1.02381e-29
2	8	1.850073e-35
3	2	7.509055e-21
3	3	7.914824e-23
3	4	1.502644e-31
3	5	1.065453e-38
3	6	6.288628e-29
3	7	7.295743e-35
3	8	1.210746e-41
4	2	1.386803e-32
4	3	7.491554e-38
4	4	5.483592e-34
4	5	2.265415e-30
4	6	4.179611e-32
4	7	8.452169e-33
4	8	7.92672e-36
5	2	1.482315e-39
5	3	1.955443e-41
5	4	1.212172e-42
5	5	1.478913e-37
5	6	5.233457e-25
5	7	1.515037e-30
5	8	1.509897e-31
6	2	6.028986e-36
6	3	5.115547e-39
6	4	1.400757e-45
6	5	2.361710e-38
6	6	8.683324e-24
6	7	1.970256e-26
6	8	2.346110e-52
7	2	5.72812e-40
7	3	4.160587e-33
7	4	1.417446e-33
7	5	5.36238e-33
7	6	6.014189e-23
7	7	7.42164e-50
7	8	1.182855e-28
8	2	2.409672e-37
8	3	3.045161e-30
8	4	1.222534e-39
8	5	2.962932e-37
8	6	3.642299e-23
8	7	1.365134e-38
8	8	8.555289e-32

The aim of the χ^2 test was to find the SOM grid that was best correlated for classification according to the phylum level of taxonomy. The grids with x and y

ranging from 2 to 8 were evaluated using the χ^2 test of independence of the rows (groups) and columns (Phyla). The resulting p-values from the different combinations of the grid are given in the Table 4.1. The table indicates that the optimal grid for the SOM classification is 6×8 from the p-values.

4.2.2.4.2 Residual analysis

A 2×2 contingency table was created with the 48 groups from the SOM optimal grid as the rows and the 16 different Phyla as the columns making it a 48×16 matrix. This contingency table was given as the input to the R function `chisq.test()` for doing the χ^2 test of independence of rows and columns for the contingency tables. The residuals of the χ^2 test for groups (rows (48 in total as the grid is 6×8) versus phyla (column (16 in total as there are 16 different phyla in the analysis)) were used to create a table with rows containing the residuals of χ^2 test for the SOM groups and columns representing the residuals of the χ^2 test for the phyla. The table cells with the absolute value of the residuals greater than equal to 2 were highlighted.

4.2.2.5 Heat maps

A 48×16 matrix created with the 48 different groups generated by the optimal SOM grid making the rows and the 16 different Phyla making the columns was used as input for creating the heat map. A heat map with dendrograms was created with R using function `heatmap()` from the package `stats`. Color scheme used was `redgreen`. A second heat map was created with default parameters using the residuals from the χ^2 test used for testing the independence of the grouping created by the grid, 6×8, and the 16 different Phyla. A third heat map was generated by using the codebook vectors, generated by the self-organizing map with the chosen grid viz., 6×8, as the matrix to the R function `heatmap`. The rows represent the 48 different groups generated by the optimal grid 6×8 and the columns shows the 64 different codons.

4.2.3 Comparison of the overlap numbers with other databases

Two other databases viz., BPhyOG (Luo *et al*., 2007) and PairWise Neighbours ((Pallejà *et al*., 2009), also provide the overlap numbers for the prokaryotic species. These two databases were searched for the presence of the microbial species used in

this study to compare the overlap numbers obtained from this study to those available from these databases. A table was prepared to contain the species name, the accession number from the GenBank, the GenBank identifier and the overlap numbers from these three studies. Plots were also generated to visualise the comparison of the overlap numbers from these different studies.

5 Results

5.1 GenBank formatted file for the gene overlaps

The GenBank formatted file written for one of the overlaps in *E. coli* 536 is given in the Figure 5.1. The different sections of the written GenBank formatted file are

1. **LOCUS:** The section contains information about (1) locus name, `Overlap_1`, (2) length of the overlap in nucleotides – here 17 bp, (3) molecule type, DNA, (4) type of the molecule, linear in this case and (5) GenBank division to which the record belongs to as a three letter abbreviation – here UNK is used which stands for the unannotated sequences.

2. **ACCESSION:** This section contains the specific identifier of a sequence record. As this was a custom made GenBank formatted file created to include the information of the overlap part and both of the coding sequences involved in the overlap, the accession number is `unknown` in this case.

3. **FEATURE:** The features in this section are the overlap part, `OVERLAP_1` and both the coding sequences, `CDS1` and `CDS2`. For each feature, feature keys and its values were generated to include the information needed for further data generation for the analyses. The features and the details given under each of them that were written to the GenBank file for the overlap part are given below.
    - ***OVERLAP_1:*** Location of the overlap
    - ***CDS1 and CDS2:*** (1) Location of the coding sequences. If the coding sequence is located on the complementary strand, the term `complement` is written before the base span. (2) Feature keys and values for the coding sequence originally present in the genomic GenBank file including its `translation`, `translation table`. (3) DNA sequence of the coding part from the original genomic GenBank file corresponding to the location mentioned in the feature part. (4) Strand information of the coding sequence as got from the genomic GenBank file.

4. **ORIGIN:** The sequence data for the overlap part was written under this section.

GenBank formatted files for the overlap part were created to provide the information needed for translation checking of the CDSs involved in the overlap, phase calculation of the CDSs and their overlapping parts. These files for the overlaps can help to generate the statistics for the overlap parts, to calculate the codon usage of the overlap part etc. without looking into the original genomic GenBank file.

```
LOCUS Overlap_1   17 bp dna linear UNK
ACCESSION unknown
FEATURES Location/Qualifiers
 OVERLAP_1 1023783..1023799
 CDS1 1023629..1023799
   /locus_tag="ECP_0963"
   /sequence="TTGTATAAGGTACGTTTAATCTTTTTTGTCAGCGACAATTTACAGAA
GAAAATCGCGGAAACCGCTTCAGACAAGCCTCCGCAAGGAAAATTAGTCACGACTGAA
AGCATTGGCTGGGCGACAAAAAAAGTTCCAGGATTAATCCTAAATTTACTTAATGATA
   CAAATTAG"
   /protein_id="YP_668877.1"
   /transl_table=11
   /db_xref="GI:110641147"
   /db_xref="GeneID:4190177"
   /codon_start=1
   /strand="1"
   /translation="MYKVRLIFFVSDNLQKKIAETASDKPPQGKLVTTESIGWATKKV
   PGLILNLLNDTN"
   /product="hypothetical protein"
 CDS2 complement(1023783..1024292)
   /locus_tag="ECP_0964"
   /sequence="TTAATGATACAAATTAGAGTGAATTTTTAGCCCGGAAAGTTGTCTCG
TTGCGTGAGAGGATGCGCTTACCGGACGCATAATAAACCCCATAGCGTTACCTTCATT
TGCCGCATCAACAAGTTCAGCATGCTCTTCTGCAGTCAAATCATCTGCCAACCAACCG
ATCACCACGCTGTAATTGCCCGTGCGTAAAGCGCGAACCATTGACTCCACGGTGTGGC
AAGGGGAGAGCTGGCTAATCTGCATTACTTTCGTTAAGGGTAGCCCAGATGCCTGAAC
CCATTCGCGACTCAGTTTTTGTTGCGGTGTTAACCAGAGTTGCCAGCGCGATTGCTGA
CCGAGTTGCTGTAACAATGGCAACAGTAGAAGTTGCGTCATCATGGGCTGATCTTCGC
GATAGACAACTTCACTGATAAGCCCGGCTGTAGTGTTTTCCGTAGAGACACGCGCAAT
TTTACTTGCTGCGGATGAGAACGACGAAGAACGATGTGCATAGCCTGAAGTGTACAT"
   /protein_id="YP_668878.1"
   /transl_table=11
   /db_xref="GI:110641148"
   /db_xref="GeneID:4190178"
   /codon_start=1
   /strand="-1"
   /translation="MYTSGYAHRSSSFSSAASKIARVSTENTTAGLISEVVYREDQPM
   MTQLLLLPLLQQLGQQSRWQLWLTPQQKLSREWVQASGLPLTKVMQISQLSPCHTVES
   MVRALRTGNYSVVIGWLADDLTAEEHAELVDAANEGNAMGFIMRPVSASSHATRQLSG
   LKIHSNLYH"
   /product="SOS cell division inhibitor"
ORIGIN
 1 ttaatgatac aaattag
//
```

Figure 5.1. GenBank formatted file created for an overlap from the species *E. coli* 536.

5.2 Result files for each organism

The result file, `Results.txt`, for *E.coli* 536 written to the subfolder named after its GenBank identifier is in the Figure 5.2.

```
GI number : 110640213
Species Name : Escherichia coli 536
Total No: of codons of from the overlapping part : 4897
Overlaps_with_length_above_nine:238
Min_Ovrlp_len:2
Max_Ovrlp_len:250
Median_Ovrlp_len:4
Mean_Ovrlp_len:15.2101105845182
StdDev_Ovrlp_len:22.8373810574615
X-squared = 9469.226
p-value < 2.2e-16
Codon  Expected      Observed
AAA    163.662211    94
……..
TTT    110.832006    110
Observed_numbers:94 …..110
Observed_relative_percents:57.4353721764152…… 99.2493089045054
DIVERGENCE IN PERCENTAGE
GGT    42.1809562526994……
TAG    3433.3349355563
DIVERGENCE:42.1809562526994       …… 3433.3349355563
LeastDeviated:GGT=42.1809562526994%.........   GAG=56.3639721214132%
MostDeviated:TAG=3433.3349355563% ……... AGA=440.696545950089%
```

Figure 5.2. The formatted `Results.txt` file created for *E. coli* 536.

In the `Results.txt` file, information was written in a format similar to that of `FEATURE` section of the GenBank formatted file for the overlaps with feature keys and values. The details given in the feature section in the `Results.txt` file are the GenBank identifier of the species, name of the species, total number of codons from all of the overlapping parts in that particular species, total number of overlaps whose lengths were greater than nine nucleotides in the given species, minimum overlap length for the species, maximum overlap length for the species, median overlap length for the species, mean overlap length for the species, standard deviation of the overlap lengths, χ^2 statistic from the results of the χ^2 goodness-of-fit test for evaluating difference in the codon usage by the overlapping part and the normal parts, p-value from the results of the χ^2 test for evaluating difference in the codon usage of the overlapping part and the normal parts, 64 different codons where each codon is given in a separate line along with their expected and observed numbers from the results of the χ^2 goodness-of-fit test, p-value from the χ^2 test, observed numbers for the codons from the overlapping part, observed relative percentages for the codons from the overlap part, codons from the overlapping part and their respective divergence from

the expected numbers in percentage, divergence in percentages written in the ascending order without the codons, five least biased codons and their relative percentages, five most biased codons and their relative percentages, ambiguous codons and their relative percentages.

5.3 Final results for all the microbes in the analysis

The final result file, `Final.txt`, contains the information of the overlaps for all the species under study. A part of it is given in the Figure 5.3, with the details of two out of the 435 species included in the study.

```
GI No: SpeciesName  MinOvrlpLength      MaxOvrlpLength      MedianOvrlpLength
       MeanOvrlpLength     StdDev_Ovrlp_len    P-value      LeastDeviated
       MostDeviated Ambiguous_and_Codons_Not_Present

110640213     Escherichia coli 536      2      250    4      15.2101105845182
       22.8373810574615    <2.2e-16            GGT=42.1809562526994%,
GAA=49.6185096554146%, CTG=51.7337911888818%, ATC=54.8513897584427%,
GAG=56.3639721214132%            TAG=3433.3349355563%, TGA=3252.39752715752%,
TAA=1437.5899385833%, AGG=688.58838279785%, AGA=440.696545950089%


117622295     Escherichia coli APEC O1  2      609    8      24.3671428571428
       44.4495606233358    <2.2e-16     CTG=43.550350768567%,
ATC=48.4981274894063%, ATT=49.5733227529503%, GGT=51.3388980518769%,
ACC=55.1347591251911%      TGA=2861.93595391624%, TAG=2832.57231557122%,
TAA=1256.21455925246%, AGG=508.752914023635%, CGA=389.457986162274%
```

Figure 5.3. A part of the formatted `Final.txt` file. This figure includes information on two out of the 435 species used in the study with headers and information of the overlaps.

The `Final.txt` was used as input to R for further analyses and it is a tab separated file with a header line followed by the information about the headers for each species in a new line. The headers and the details given under them are

- **GI No:** GenBank identifiers of the species.
- **SpeciesName:** Name of the species.
- **MinOvrlpLength:** Minimum overlap length for the species.
- **MaxOvrlpLength:** Maximum overlap length for the species.
- **MedianOvrlpLength:** Median of the overlap lengths for the species.
- **MeanOvrlpLength:** Mean of the overlap lengths for the species.
- **StdDev_Ovrlplen:** Standard deviation of the overlap lengths for the species.
- **P-value:** P-value of the χ^2 goodness of fit test to find whether there is any difference between the observed and the expected numbers of codons.

- **LeastDeviated:** Five least biased codons and their divergence in percentage.
- **MostDeviated:** Five most biased codons and their divergence in percentage.
- **Ambiguous_and_Codons_Not_Present:** Information about the number of ambiguous codons and codons which were not present in the codon usage table with their respective numbers.

5.4 Taxonomical classification of the studied microbes

`Taxonomy.txt` is a tab separated file with the taxonomical classification of the studied microbes at the different levels extracted from the NCBI's Taxonomy database. It consists of a header line and the details as indicated by the header line for each species in a separate line. The headers in the file are `GI_no` for the GenBank identifier of the species, `Species_name` for species name, `Tax_id` for taxonomy identifier of the species from the NCBI's taxonomy database followed by `Species, Genus, Family, Suborder, Order, Subclass, Class, Phylum, Super kingdom, No_rank`. The file was given as input to R for further analyses. A small portion of the file `Taxonomy.txt` with taxonomic information at different levels mentioned above is illustrated in the Figure 5.4.

```
GI_no  Species_name  Tax_id Species       Genus  Family Suborder      Order
       Subclass      Class  Phylum Superkingdom  No_rank

110640213     Escherichia coli 536       362663 Escherichia coli
       Escherichia   Enterobacteriaceae         Enterobacteriales
γ-proteobacteria     Proteobacteria      Bacteria      cellular organisms


117622295     Escherichia coli APEC O1   405955 Escherichia coli
       Escherichia   Enterobacteriaceae         Enterobacteriales
γ -proteobacteria    Proteobacteria      Bacteria      cellular organisms
```

Figure 5.4. A section of the `Taxonomy.txt` file, with the headings, created to contain the taxonomy information at different levels for the microbial species.

5.5 Descriptive statistics of the overlaps

5.5.1 Distribution of the overlap numbers

The information from the files, `Taxonomy.txt` and `Final.txt` were used as inputs to R to generate the overlap number distribution for the 16 different phyla viz., Actinobacteria, Aquificae, Bacteroidetes, Chlamydiae, Chlorobi, Chloroflexi, Crenarchaeota, Cyanobacteria, Deinococcus-Thermus, Euryarchaeota, Firmicutes,

Nanoarchaeota, Proteobacteria, Spirochaetes, Tenericutes, Thermotogae, to which the microbes used in this study belonged to. The Figure 5.5 shows the box plots visualising the distribution of the overlap numbers in the different phyla mentioned above.

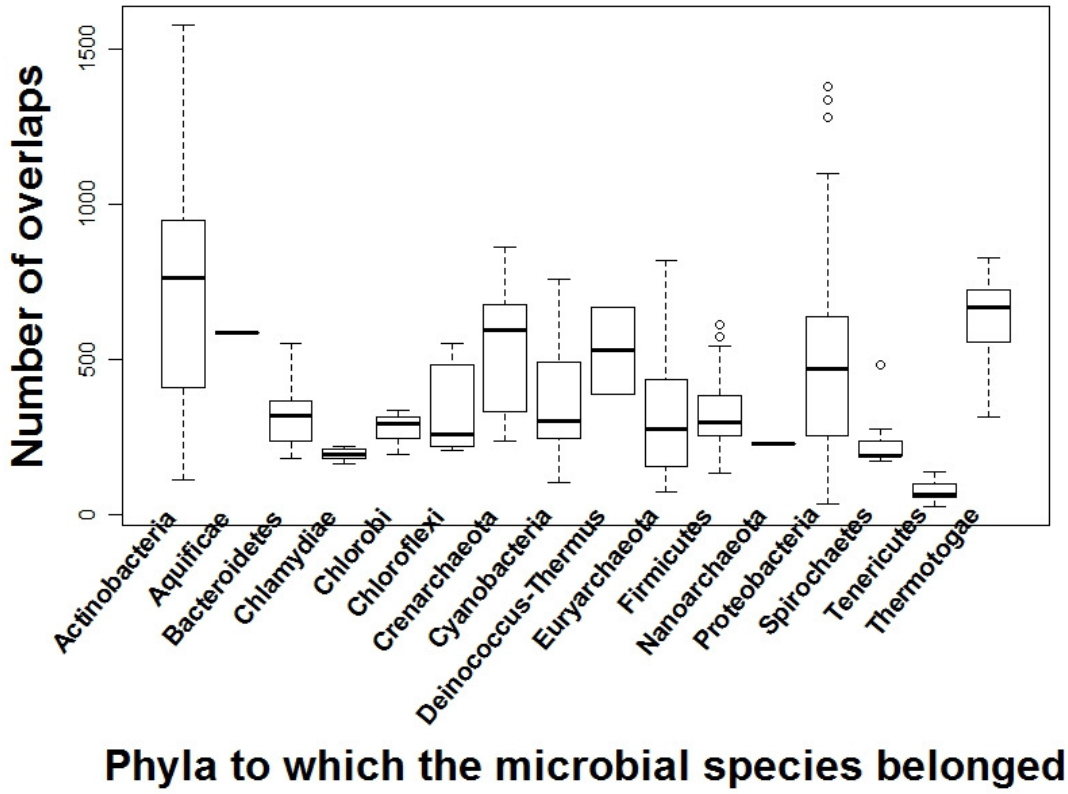


Figure 5.5. Phylum-wise distribution of the overlap numbers. X-axis represents the 16 different phyla to which the microbial species used in the study belonged to and y–axis, the overlap numbers exhibited by the different microbial species in each phylum. Colors used for each phylum is explained in the legend box.

Phylum Actinobacteria has the greatest number of overlaps and phylum Tenericutes has the least. Due to the presence of only one member species for the phyla Aquificae and Nanoarcheota, they were represented only by the median line in the box plot. The overlap numbers from the 435 different species used in the study were arranged in the ascending order of their overlap numbers and were plotted against their respective species in Figure 5.6. Number of same strand overlaps and opposite strand overlaps were calculated for all the microbes used in the analyses and plotted along with their respective total overlap numbers arranged in the ascending order in Figure 5.7. The figure shows that same strand overlaps are present in greater amounts than the opposite strand overlaps in the analysed microbes.

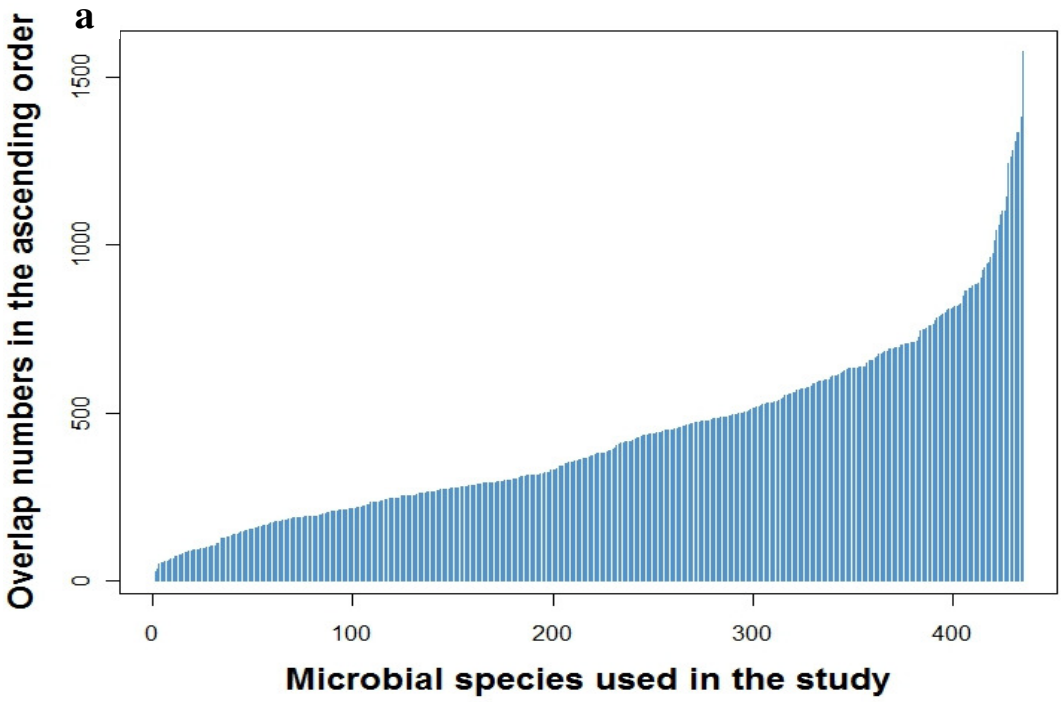


Figure 5.6. Distribution of the overlap numbers in the studied species with the overlaps arranged in the ascending order of the overlap numbers

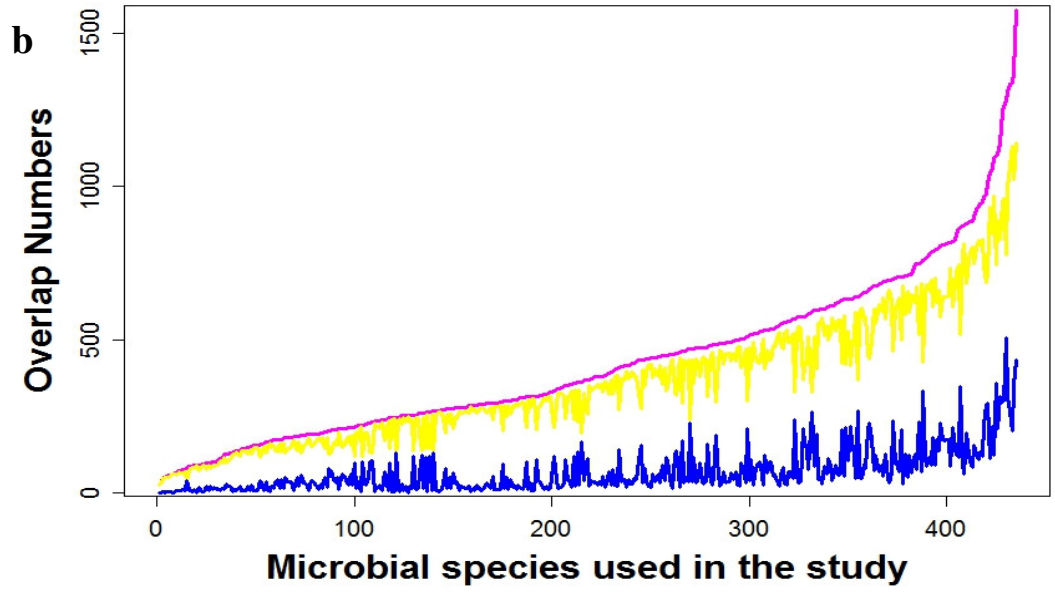


Figure 5.7. Strand-wise distribution of the overlaps. Overlap numbers are plotted in the ascending order. Magenta line is for the total overlap numbers, yellow line for the same strand overlaps and the blue line for the opposite strand overlaps in each species.

5.5.2 Statistics for the overlap lengths and numbers

Table 5.1 depicts the statistics for the overlap numbers and overlap lengths for the 16 phyla. It shows that phylum Proteobacteria has the highest number of species for the analyses. Average number of overlaps was highest in the phylum Actinobacteria, with 707 overlaps and least in the phylum Tenericutes with 75.5. Average maximum overlap length is greatest for the phylum Actinobacteria with 525.5 base pairs and least for Chlorobi with 90.33 base pairs.

Table 5.1. Phylum–wise statistics for the overlap number and length of the studied species. Highest values for each of the categories are highlighted in red.

Phyla	No of species analysed	Average number of overlaps	Average no: of overlaps with length>9	Average minimum of overlap length	Average maximum of overlap length	Average median of overlap length	Average mean of the overlap length	Average SD of overlap length
Actinobacteria	41	707	170.12	1.68	525.20	3.49	14.66	38.67
Aquificae	1	587	285	3	434	9	18.85	35.50
Bacteroidetes	9	327.22	148.56	1.44	209.56	8	15.55	23.34
Chlamydiae	10	194.20	103.70	1.60	188.60	10.90	20.08	26.47
Chlorobi	3	273.33	105.00	1.67	90.33	4	11.12	13.33
Chloroflexi	5	343.60	117.80	2.20	139.20	5.40	10.49	13.92
Crenarchaeota	11	538.27	259.55	2.64	230.82	9.05	18.05	27.51
Cyanobacteria	27	371.59	165.48	1.37	283.41	6.48	17.14	27.75
Deinococcus-Thermus	2	527.50	172.50	1	430.50	5	11.81	26.28
Euryarchaeota	24	307.29	112.08	2.33	151.50	5.67	11.75	16.97
Firmicutes	95	323.55	145.35	2.33	272.04	8.07	15.07	24.76
Nanoarchaeota	1	227	137	3.00	187.00	13	18.20	21.53
Proteobacteria	181	476.85	147.12	1.86	403.92	5.27	14.71	31.16
Spirochaetes	8	234.88	119	2.50	220.50	9.44	17.91	26.70
Tenericutes	11	75.45	47.82	3.09	147.45	11.86	19.13	24.89
Thermotogae	6	625	242.33	2.33	430.33	4.83	15.25	35.81

5.6 Principal component analysis

Standard deviations of the observed relative percentages of the 64 different codons in the alphabetical order were plotted and are given in Figure 5.8. As this figure showed very much deviation for the codons TAA, TAG, TCA, TGA, PCA was performed after scaling using the R function `prcomp()` from the package `stats` to find patterns of codon usage bias in the phyla level of taxonomy. A 435 × 64 data matrix was given as input, which contained the observed relative percentages of the 64 different codons from the overlapping part of the coding sequences for 435 species.

The information about the phylum for the different microbes will be taken into account after the principal component analysis.

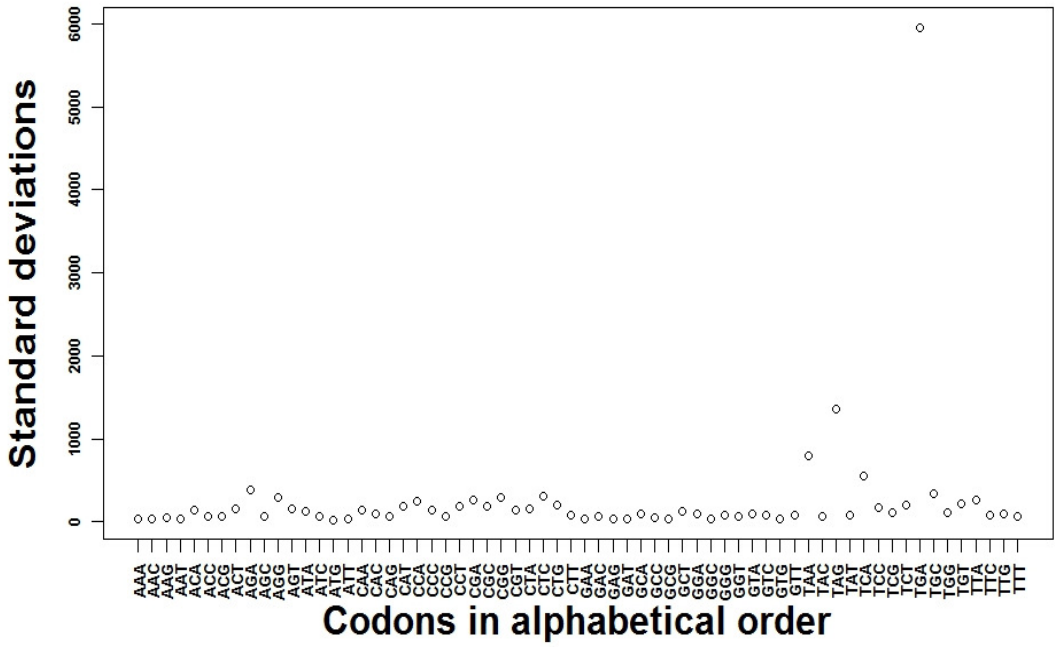


Figure 5.8. Plot of the 64 different codons in their alphabetical order versus standard deviations of their observed relative percentages of codons.

PCA results show that the first principal component accounts for 29.1 % and the second principal component accounts for 9.5 % of the total variance. The first principal component was plotted against the second principal component using the plotting character `bullet` in the Figure 5.9(a) and a smoothed color density plot corresponding to the scatter plot in (Figure 5.9(b)). As the Figure 5.9 cannot visualise any particular grouping of the codon usage bias patterns, the first and second principal component was plotted against each other and was colored according to the phyla in Figure 5.10(a). Individual principal component plots for each of the 16 different phyla were created, to explore patterns of codon usage bias by the codons of the overlapping part in each individual phylum (Figure 5.10 (b, c)). All of them were plotted using the same measurements for the x and y axes. From the plots that visualise the principal component analysis, it is evident that codon usage bias patterns within the phyla level cannot be inferred.

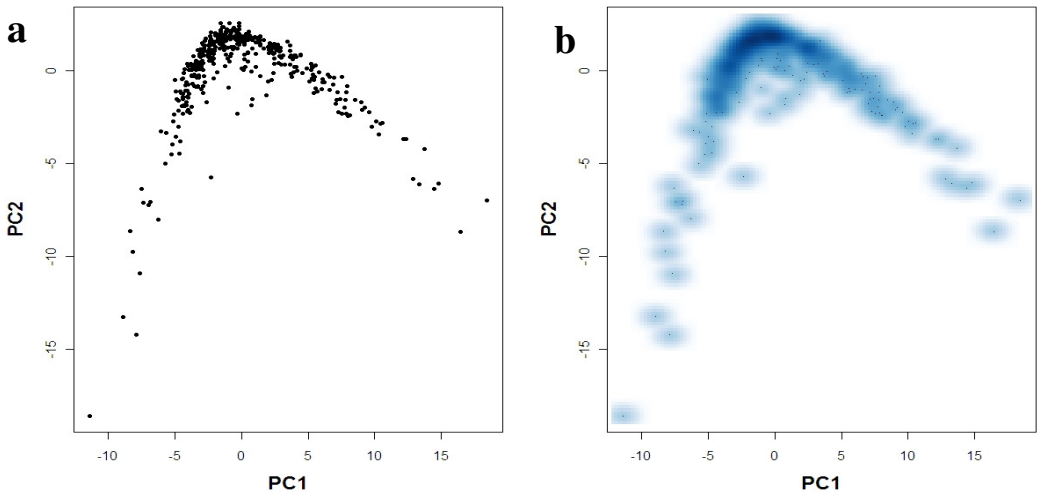


Figure 5.9. Scatter plot (a) and smooth scatter plot (b) of the first two principal components against each other, from the principal component analysis of observed relative percentages of the codons intended for finding patterns in the codon usage bias.

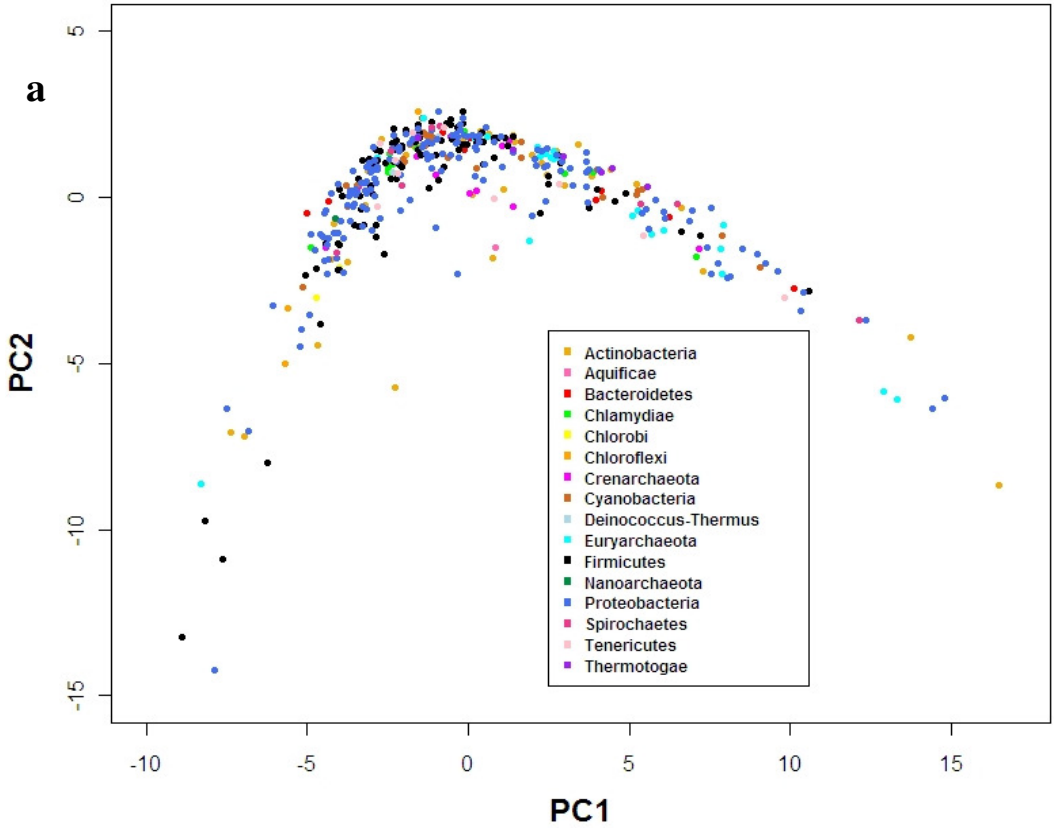
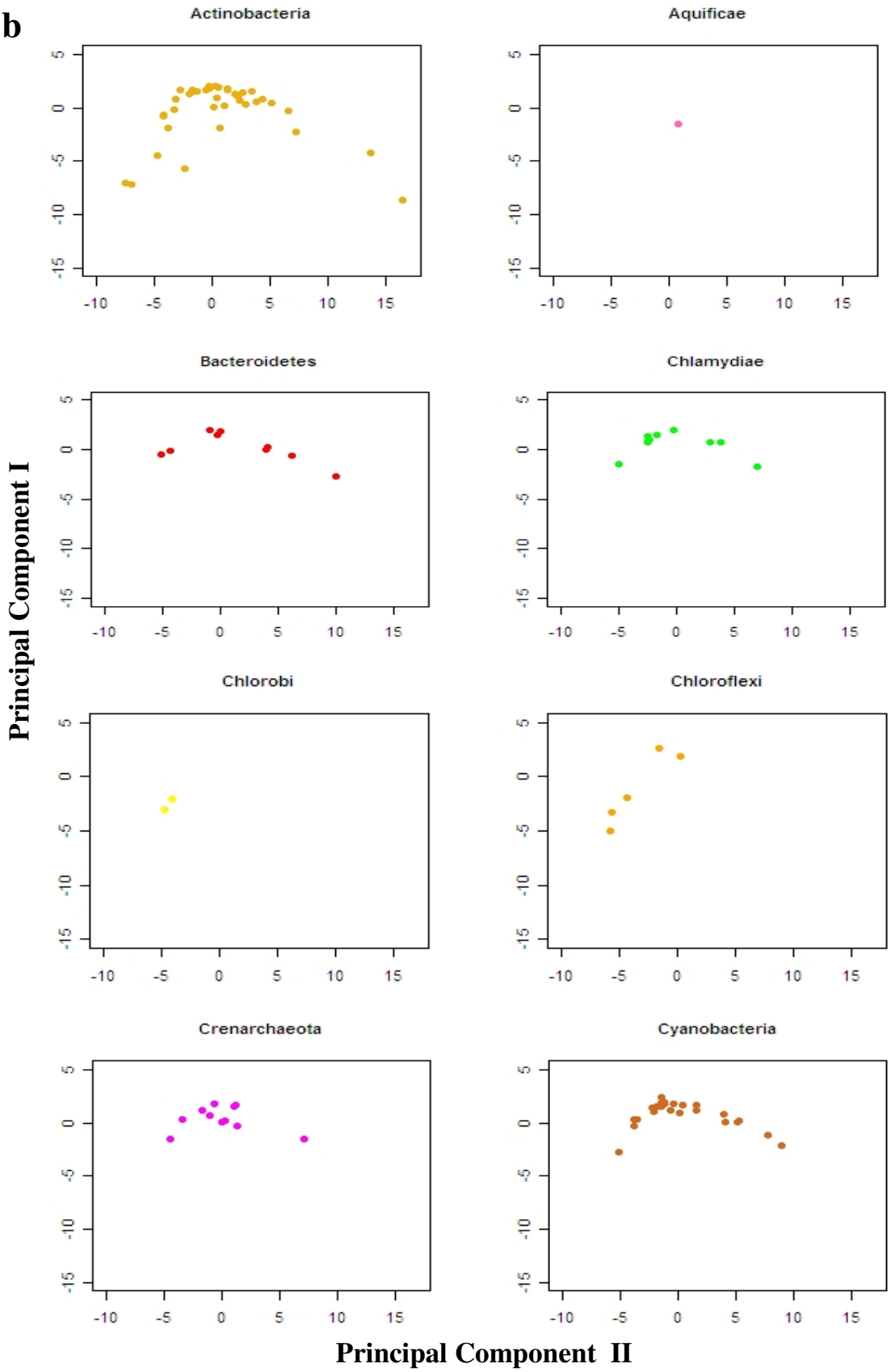


Figure 5.10(a). Scatter plot of the first and second principal components against each other for detecting the codon usage bias patterns. The scatter plot is colored according to the phylum and the colors used for each phylum are given in the legend box.


b
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Crenarchaeota
Cyanobacteria
Principal Component I
Principal Component II

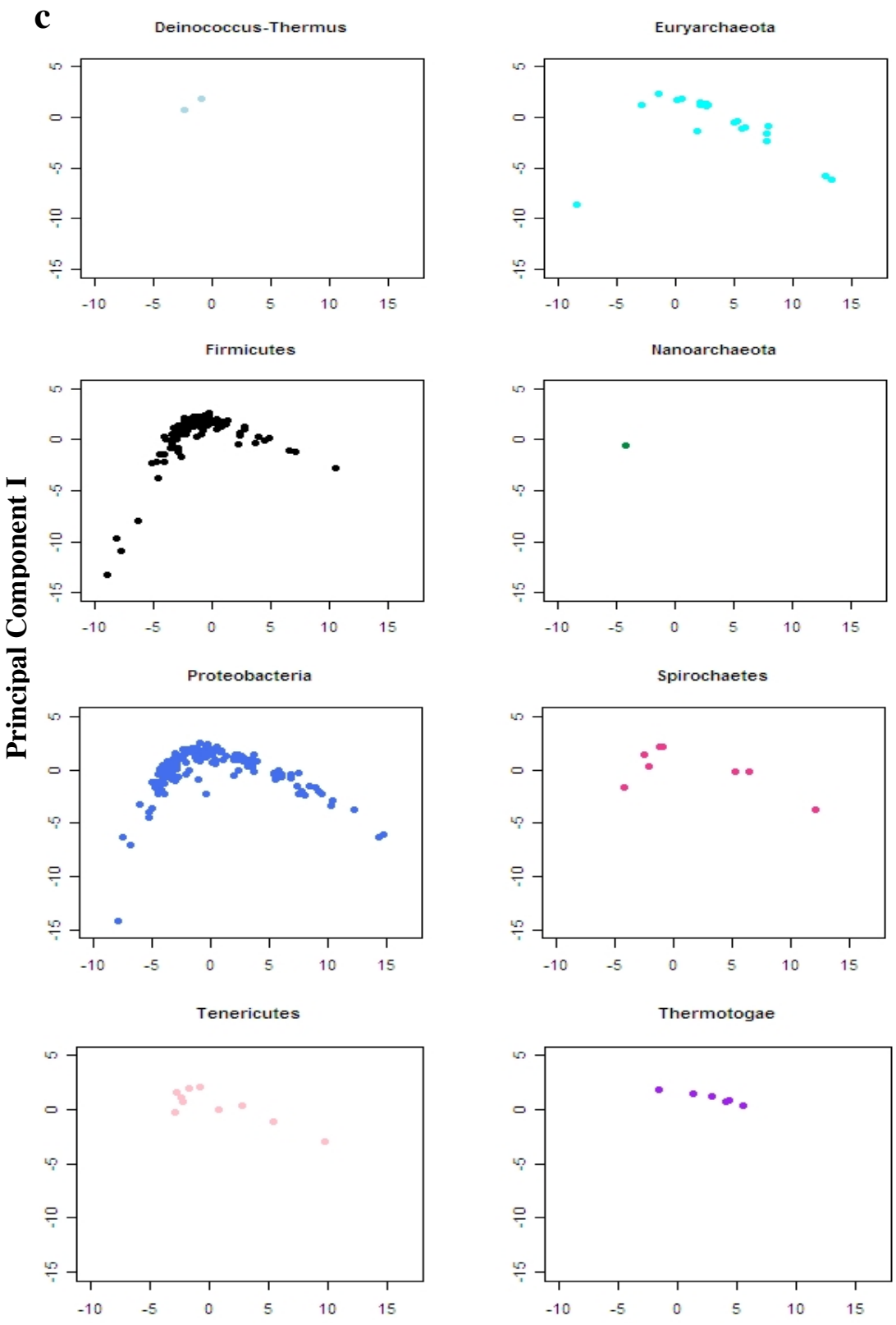


Figure 5.10(b, c). Scatter plot of the first two principal components for each Phyla

5.7 Correspondence analysis

Two different kinds of maps were generated using the correspondence analysis, the symmetric map and asymmetric maps given in Figure 5.13and Figure 5.14. Point intensity (shading) corresponds to the absolute contributions of the points to the planar display and the point size is proportional to the relative frequency (mass) of each point.

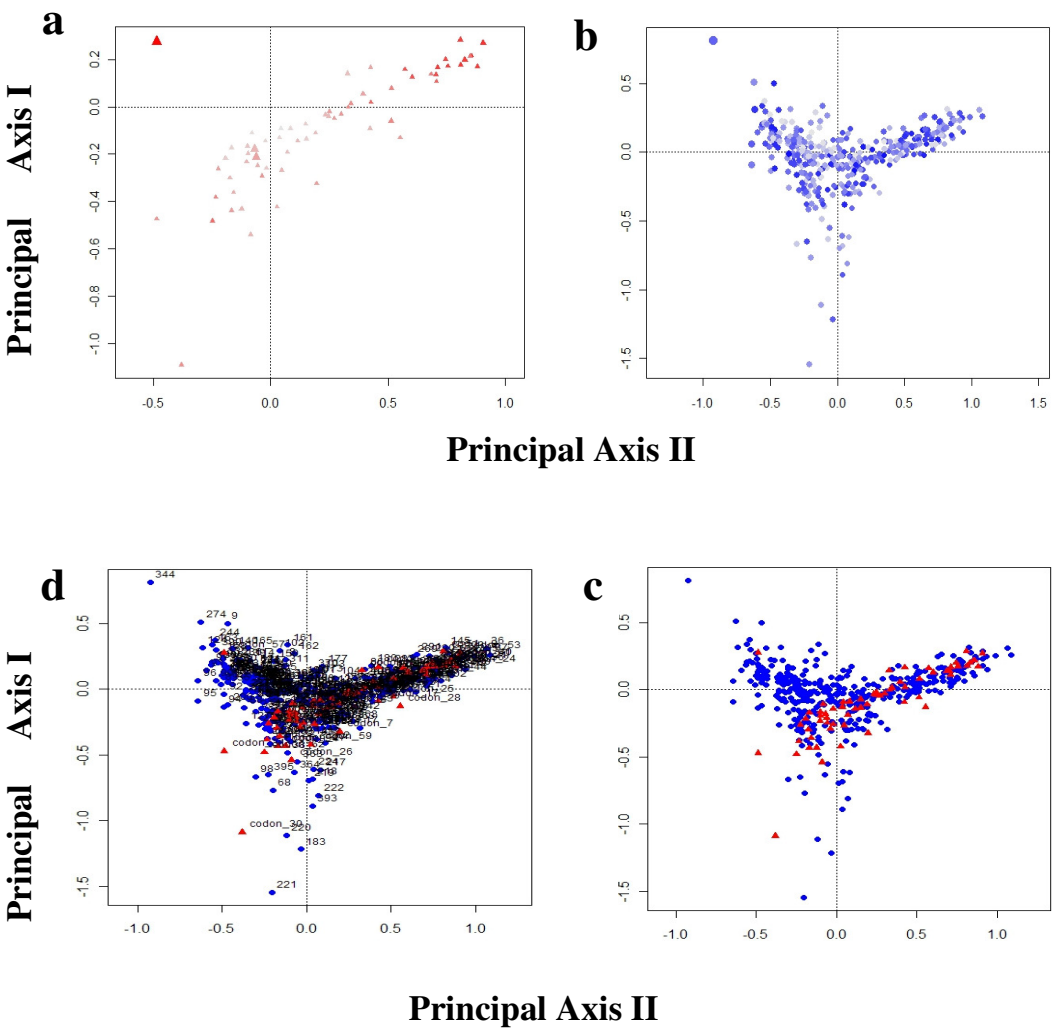


Figure 5.13. Symmetric map from the correspondence analysis of the 2 way contingency table. Blue dots represent the 413 different species (rows) and the red triangles, the 64 different codons (columns). (a) is the symmetric map from the correspondence analysis with labels of species and codons added onto the plot. (b) is without labels. (c) and (d) are symmetric maps without labels. (c) consists of only row points (species) and (d) consists of only column points (codons).

The symmetric map, Figure 5.13, shows the mapping of the points that represent the row coordinates and column coordinates corresponding to the first two principal axes, when the total inertia is decomposed along principal axes. This symmetric map of a simple correspondence analysis represents rows and column by points. Row points closer together have more similar column profiles and column points that are closer together have more similar row profiles. Symmetric maps of the same data matrix from the correspondence analysis that visualises only the row points representing the species, Figure 5.13(c), and the column points representing the codons, Figure 5.13(d) with the point sizes being proportional to the mass or the relative frequency of the points were also created. In the asymmetric map (Figure 5.14 (a, b)), rows are represented in the principal coordinates and columns in standard coordinates multiplied by the square root of the corresponding masses, giving reconstructions of the standardized residuals. Rows are represented by points and columns by arrows. Trends in the codon usage pattern of the overlapping part of the genomes cannot be inferred from these plots due to the lack of any clear clustering patterns.

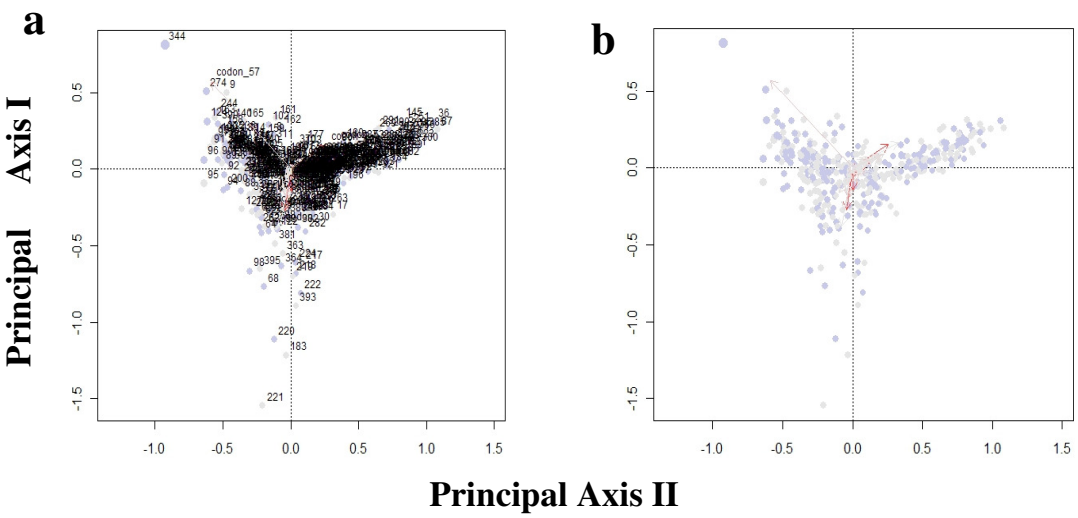


Figure 5.14. Asymmetric map from the correspondence analysis of the 2 way contingency table. (a) is with labels for the codons (codon_1 to codon_64) and the species (numbers from 1 to 413) and (b) is without the labels. In both the plots (a, b), columns, the 64 codons, are represented by arrows and rows by points.

5.8 Self-organizing map

As the principal component analysis yielded little information about the codon usage bias patterns, self-organizing maps were created to check for the patterns in codon

usage bias by the codons in the overlapping part of the genomes of the microbes belonging to different phyla. Self-organizing map for the classification was created using the function `wccsom()` from the R package `wccsom`. The input was the 413 × 64 data matrix with observed relative percentages of the 64 different codons for 413 species. The grids used for classification, x from 2 to 8 and y from 2 to 8, were checked for the optimality using the χ^2 test of independence of the grouping generated by the grid and the different phyla used in the analyses. P-values of all the grids used for doing the χ^2 tests were significant indicating a strong association between the classification and the different Phyla.

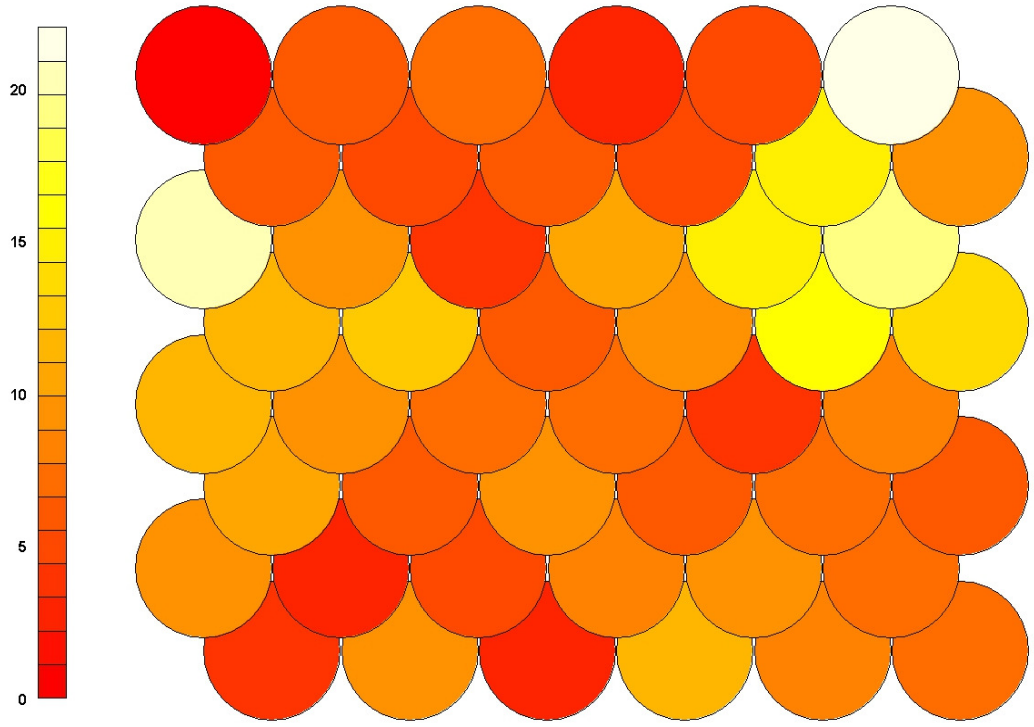


Figure 5.11. Mapping counts for the self-organizing map using the optimal 6×8 grid. Each shade of the color indicates the number of phyla belonging to one of the 48 groups generated from the SOMs. The number of members belonging to each of the groups according to the color is given on the color legend in the left hand side of the figure. The color indicates mapping of the counts in the different groups generated by the selected grid (6×8) of the self-organizing map.

Table 4.1 shows that the optimal grid for the classification is 6×8 as this combination gave the least p-value for the χ^2 test. The self-organizing map with the grid 6×8 was therefore taken for further analyses and for visualisation of the mapping. The mapping count for each of the 48 different groups created as a part of the clustering using self-

organizing maps (Figure 5.11) and the mapping given by the self-organizing map using the plotting character `bullet` (Figure 5.12) was created. The table 5.2 gives the number of organisms in each phylum belonging to the 48 different groups generated by the optimal grid for the self-organizing map.

5.8.1 Residual analysis of the χ^2 test

The Table 5.3 presents the residuals of the χ^2 test for testing independence of the 48 groups generated by the 6×8 grid used for creating the self-organized map and the 16 phyla to which the microbes belonged to. The cells in the Table 5.3 with the absolute value of the residuals greater than or equal to 2 are highlighted. The cells with a positive value for the residuals are overrepresented, i.e., the observed number of organisms in a phylum belonging to a particular group is greater than the expected number of organisms (members) that should belong to that group. The cells with a negative value for the residuals are underrepresented i.e., the observed number of microbes in a phylum belonging to a particular group is lower than expected number of microbes in that group. Here, analyses of the residuals indicate that the residuals with absolute value above 2 are overrepresented, indicating that the members of that phylum are present in greater numbers than the expected number in the particular group. In this analysis the highest value of the residual, 9.37, is for the Phylum Spirochaetes in the group 8.

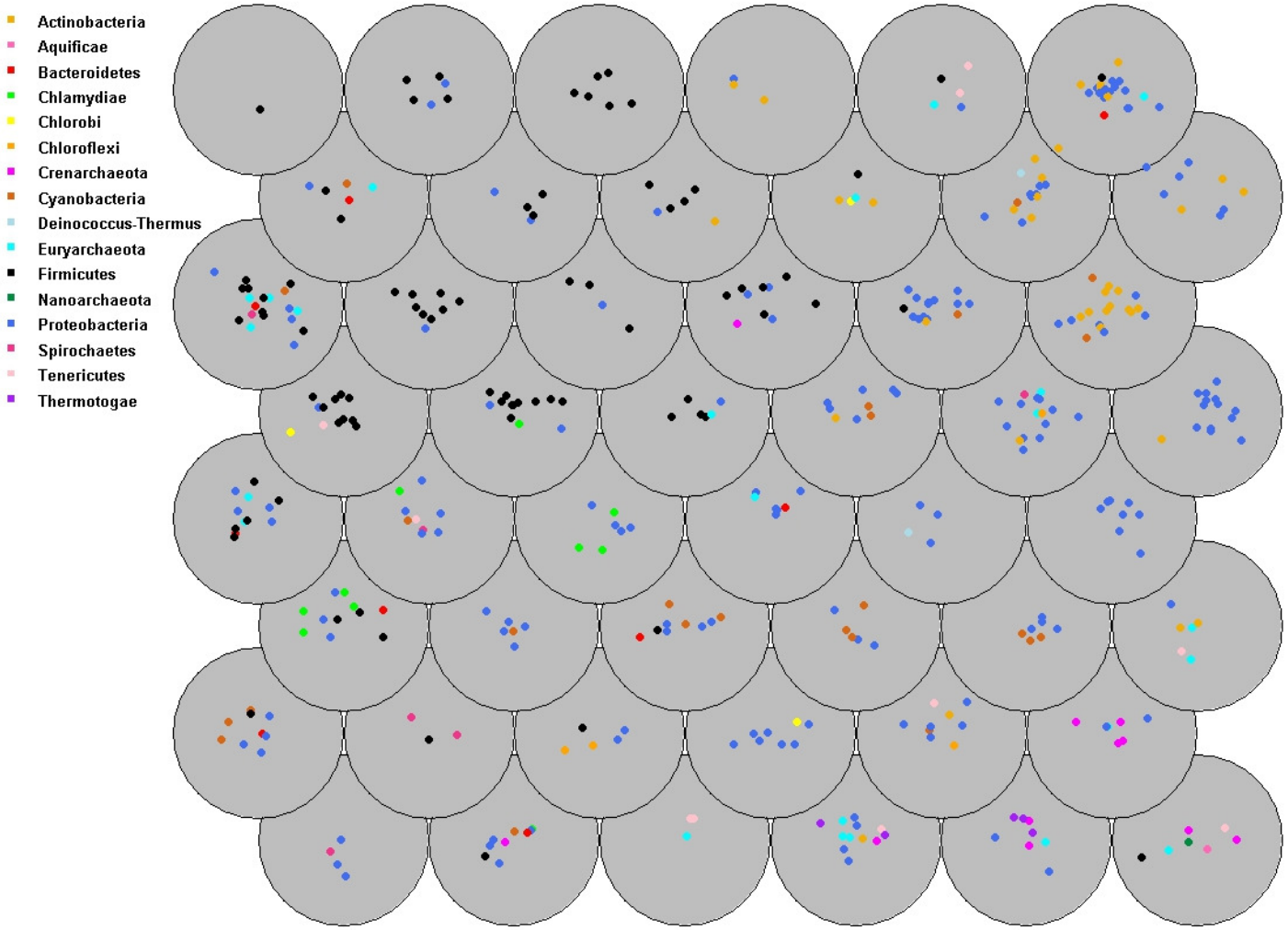


Figure 5.12. Results of the clustering using SOM. The grid used is the 6×8. The plotting character `bullet` is colored according to the phylum.

Table 5.2. Number of organisms in the 16 different phyla under study that belong to each of the 48 different groups derived from the 6 × 8 grid SOM classification. The total number of organisms from the different phyla that belonged to each group is also given in the table. The table shows that group 48 has the maximum number of members, 22, and the group 42 has the least number of members, 1.

Groups	*Actinobacteria*	*Aquificae*	*Bacteroidetes*	*Chlamydiae*	*Chlorobi*	*Chloroflexi*	*Crenarchaeota*	*Cyanobacteria*	*Deinococcus-Thermus*	*Euryarchaeota*	*Firmicutes*	*Nanoarchaeota*	*Proteobacteria*	*Spirochaetes*	*Tenericutes*	*Thermotogae*	*Group Total*
Group1	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	4
Group2	0	0	1	1	0	0	1	1	0	0	1	0	4	0	0	0	9
Group3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	3
Group4	1	0	0	0	0	0	1	0	0	3	0	0	4	0	1	2	12
Group5	0	0	0	0	0	0	2	0	0	1	0	0	2	0	0	3	8
Group6	0	1	0	0	0	0	2	0	0	1	1	1	0	0	1	0	7
Group7	0	0	1	0	0	0	0	3	0	0	1	0	4	0	0	0	9
Group8	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	3
Group9	0	0	0	0	0	2	0	0	0	0	1	0	2	0	0	0	5
Group10	0	0	0	0	1	0	0	0	0	0	0	0	7	0	0	0	8
Group11	1	0	0	0	0	1	0	1	0	0	0	0	5	0	1	0	9
Group12	0	0	0	0	0	0	4	0	0	1	0	0	2	0	0	0	7
Group13	0	0	1	4	0	0	0	0	0	0	3	0	3	0	0	0	11
Group14	0	0	0	0	0	0	0	1	0	0	0	0	5	0	0	0	6
Group15	0	0	1	0	0	0	0	3	0	0	1	0	4	0	0	0	9
Group16	0	0	0	0	0	0	0	3	0	0	0	0	3	0	0	0	6
Group17	0	0	0	0	0	0	0	3	0	0	0	0	4	0	0	0	7
Group18	2	0	0	0	0	0	0	0	0	2	0	0	1	0	1	0	6
Group19	0	0	1	0	0	0	0	0	0	2	5	0	4	0	0	0	12
Group20	0	0	0	1	0	0	0	1	0	0	0	0	5	1	1	0	9
Group21	0	0	0	3	0	0	0	0	0	0	0	0	4	0	0	0	7
Group22	0	0	1	0	0	0	0	0	0	1	0	0	5	0	0	0	7
Group23	0	0	0	0	0	0	0	0	1	0	0	0	3	0	0	0	4
Group24	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	8
Group25	0	0	0	0	1	0	0	0	0	0	9	0	1	0	1	0	12
Group26	0	0	0	1	0	0	0	0	0	0	10	0	2	0	0	0	13
Group27	0	0	0	0	0	0	0	0	0	1	4	0	1	0	0	0	6
Group28	1	0	0	0	0	0	0	2	0	0	0	0	6	0	0	0	9
Group29	1	0	0	0	0	1	0	0	0	2	0	0	11	1	0	0	16
Group30	1	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	14
Group31	0	0	1	0	0	0	0	1	0	4	9	0	4	1	0	0	20
Group32	0	0	0	0	0	0	0	0	0	0	8	0	1	0	0	0	9
Group33	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	4
Group34	0	0	0	0	0	0	1	0	0	0	6	0	3	0	0	0	10
Group35	1	0	0	0	0	0	1	0	0	0	1	0	12	0	0	0	15
Group36	11	0	0	0	0	0	0	2	0	0	0	0	6	0	0	0	19
Group37	0	0	1	0	0	0	0	1	0	1	2	0	1	0	0	0	6
Group38	0	0	0	0	0	0	0	0	0	0	3	0	2	0	0	0	5
Group39	1	0	0	0	0	0	0	0	0	0	4	0	1	0	0	0	6
Group40	2	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	5
Group41	6	0	0	0	0	0	0	1	1	0	0	0	7	0	0	0	15
Group42	3	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	9
Group43	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Group44	0	0	0	0	0	0	0	0	0	0	4	0	2	0	0	0	6
Group45	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	7
Group46	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3
Group47	0	0	0	0	0	0	0	0	0	1	1	0	1	0	2	0	5
Group48	5	0	1	0	0	0	0	0	0	1	1	0	14	0	0	0	22

Table 5.3 Residuals of the χ^2 test of independence of the 48 groups from the SOM classification and the 16 different phyla. The residuals with absolute value equal to or above 2 are highlighted in different colors viz., green for 6 and above, yellow for 3-5, and cyan for 2-2.99999999.

Groups	*Actinobacteria*	*Aquificae*	*Bacteroidetes*	*Chlamydiae*	*Chlorobi*	*Chloroflexi*	*Crenarchaeota*	*Cyanobacteria*	*Deinococcus-Thermus*	*Euryarchaeota*	*Firmicutes*	*Nanoarchaeota*	*Proteobacteria*	*Spirochaetes*	*Tenericutes*	*Thermotogae*
Group1	-0.6066620	-0.09841357	-0.2952407	-0.3112110	-0.17045730	-0.19682713	-0.3264009	-0.4821260	-0.1391778	-0.4719749	-0.92320108	-0.09841357	0.97184184	3.9072298	-0.3112110	-0.2200594
Group2	-0.9099929	-0.14762035	1.8151836	1.6753527	-0.25568595	-0.29524070	1.5528769	0.6595752	-0.2087667	-0.7079623	-0.66267654	-0.14762035	0.06147007	-0.3615945	-0.4668165	-0.3300891
Group3	-0.5253847	-0.08522865	-0.2556859	-0.2695167	-0.14762035	-0.17045730	-0.2826714	-0.4175334	-0.1205315	2.0377874	-0.79951559	-0.08522865	-1.13709200	-0.2087667	7.1511751	-0.1905771
Group4	-0.0990857	-0.17045730	-0.5113719	-0.5390333	-0.29524070	-0.34091459	1.2034951	-0.8350668	-0.2410630	2.8523100	-1.59903118	-0.17045730	-0.51531135	-0.4175334	1.3161396	4.8660674
Group5	-0.8579496	-0.13917780	-0.4175334	-0.4401188	-0.24106302	-0.27835560	3.8711500	-0.6818292	-0.1968271	0.8307140	-1.30560349	-0.13917780	-0.77977833	-0.3409146	-0.4401188	9.3285504
Group6	-0.8025383	7.55095684	-0.3905667	-0.4116935	-0.22549381	-0.26037782	4.2001174	-0.6377928	-0.1841149	0.9772655	-0.40246735	7.55095684	-1.73693672	-0.3188964	2.0172981	-0.2911113
Group7	-0.9099929	-0.14762035	1.8151836	-0.4668165	-0.25568595	-0.29524070	-0.4896013	3.4251038	-0.2087667	-0.7079623	-0.66267654	-0.14762035	0.06147007	-0.3615945	-0.4668165	-0.3300891
Group8	-0.5253847	-0.08522865	-0.2556859	-0.2695167	-0.14762035	-0.17045730	-0.2826714	-0.4175334	-0.1205315	-0.4087422	0.45124175	-0.08522865	-1.13709200	9.3713052	-0.2695167	-0.1905771
Group9	-0.6782687	-0.11002971	-0.3300891	-0.3479445	-0.19057705	8.86839479	-0.3649273	-0.5390333	-0.1556055	-0.5276840	-0.06333772	-0.11002971	-0.10556257	-0.2695167	-0.3479445	-0.2460339
Group10	-0.8579496	-0.13917780	-0.4175334	-0.4401188	3.90722980	-0.27835560	-0.4616005	-0.6818292	-0.1968271	-0.6674733	-1.30560349	-0.13917780	1.91293447	-0.3409146	-0.4401188	-0.3112110
Group11	0.1889167	-0.14762035	-0.4428610	-0.4668165	-0.25568595	3.09182621	-0.4896013	0.6595752	-0.2087667	-0.7079623	-1.38480163	-0.14762035	0.56921287	-0.3615945	1.6753527	-0.3300891
Group12	-0.8025383	-0.13018891	-0.3905667	-0.4116935	-0.22549381	-0.26037782	8.8320226	-0.6377928	-0.1841149	0.9772655	-1.22128024	-0.13018891	-0.58548429	-0.3188964	-0.4116935	-0.2911113
Group13	-1.0060351	-0.16320044	1.5528769	7.2345747	-0.28267145	-0.32640087	-0.5412746	-0.7995156	-0.2308003	-0.7826818	0.42860436	-0.16320044	-0.79955349	-0.3997578	-0.5160851	-0.3649273
Group14	-0.7430061	-0.12053151	-0.3615945	-0.3811541	-0.20876670	-0.24106302	-0.3997578	1.1030521	-0.1704573	-0.5780488	-1.13068579	-0.12053151	1.50118600	-0.2952407	-0.3811541	-0.2695167
Group15	-0.9099929	-0.14762035	1.8151836	-0.4668165	-0.25568595	-0.29524070	-0.4896013	3.4251038	-0.2087667	-0.7079623	-0.66267654	-0.14762035	0.06147007	-0.3615945	-0.4668165	-0.3300891
Group16	-0.7430061	-0.12053151	-0.3615945	-0.3811541	-0.20876670	-0.24106302	-0.3997578	4.4901190	-0.1704573	-0.5780488	-1.13068579	-0.12053151	0.25747523	-0.2952407	-0.3811541	-0.2695167
Group17	-0.8025383	-0.13018891	-0.3905667	-0.4116935	-0.22549381	-0.26037782	-0.4317878	4.0659291	-0.1841149	-0.6243641	-1.22128024	-0.13018891	0.56596814	-0.3188964	-0.4116935	-0.2911113
Group18	1.9487617	-0.12053151	-0.3615945	-0.3811541	-0.20876670	-0.24106302	-0.3997578	-0.5904814	-0.1704573	2.8818666	-1.13068579	-0.12053151	-0.98623554	-0.2952407	2.2424566	-0.2695167
Group19	-1.0507693	-0.17045730	1.4441521	-0.5390333	-0.29524070	-0.34091459	-0.5653429	-0.8350668	-0.2410630	1.6290452	1.52786218	-0.17045730	-0.51531135	-0.4175334	-0.5390333	-0.3811541
Group20	-0.9099929	-0.14762035	-0.4428610	1.6753527	-0.25568595	-0.29524070	-0.4896013	0.6595752	-0.2087667	-0.7079623	-1.38480163	-0.14762035	0.56921287	2.4039340	1.6753527	-0.3300891
Group21	-0.8025383	-0.13018891	-0.3905667	6.8752812	-0.22549381	-0.26037782	-0.4317878	-0.6377928	-0.1841149	-0.6243641	-1.22128024	-0.13018891	0.56596814	-0.3188964	-0.4116935	-0.2911113
Group22	-0.8025383	-0.13018891	2.1698152	-0.4116935	-0.22549381	-0.26037782	-0.4317878	-0.6377928	-0.1841149	0.9772655	-1.22128024	-0.13018891	1.14169436	-0.3188964	-0.4116935	-0.2911113
Group23	-0.6066620	-0.09841357	-0.2952407	-0.3112110	-0.17045730	-0.19682713	-0.3264009	-0.4821260	7.0458761	-0.4719749	-0.92320108	-0.09841357	0.97184184	-0.2410630	-0.3112110	-0.2200594
Group24	-0.8579496	-0.13917780	-0.4175334	-0.4401188	-0.24106302	-0.27835560	-0.4616005	-0.6818292	-0.1968271	-0.6674733	-1.30560349	-0.13917780	2.45147703	-0.3409146	-0.4401188	-0.3112110
Group25	-1.0507693	-0.17045730	-0.5113719	-0.5390333	3.09182621	-0.34091459	-0.5653429	-0.8350668	-0.2410630	-0.8174845	4.02937687	-0.17045730	-1.83446583	-0.4175334	1.3161396	-0.3811541
Group26	-1.0936754	-0.17741758	-0.5322527	1.2213489	-0.30729626	-0.35483516	-0.5884275	-0.8691651	-0.2509063	-0.8508648	4.34411952	-0.17741758	-1.52211066	-0.4345825	-0.5610436	-0.3967178
Group27	-0.7430061	-0.12053151	-0.3615945	-0.3811541	-0.20876670	-0.24106302	-0.3997578	-0.5904814	-0.1704573	1.1519089	2.40699021	-0.12053151	-0.98623554	-0.2952407	-0.3811541	-0.2695167
Group28	0.1889167	-0.14762035	-0.4428610	-0.4668165	-0.25568595	-0.29524070	-0.4896013	2.0423395	-0.2087667	-0.7079623	-1.38480163	-0.14762035	1.07695566	-0.3615945	-0.4668165	-0.3300891
Group29	-0.3891417	-0.19682713	-0.5904814	-0.6224220	-0.34091459	2.14664591	-0.6528017	-0.9642521	-0.2783556	1.1748070	-1.84640217	-0.19682713	1.56287658	1.5920204	-0.6224220	-0.4401188
Group30	-0.2538728	-0.18411492	-0.5523448	-0.5822225	-0.31889640	-0.36822985	-0.6106401	-0.9019752	-0.2603778	-0.8829842	-1.72715108	-0.18411492	2.83589938	-0.4509876	-0.5822225	-0.4116935
Group31	-1.3565374	-0.22005942	0.8545641	-0.6958890	-0.38115410	-0.44011885	-0.7298545	-0.1504801	-0.3112110	2.7347795	2.29540575	-0.22005942	-1.57354202	1.3161396	-0.6958890	-0.4920678
Group32	-0.9099929	-0.14762035	-0.4428610	-0.4668165	-0.25568595	-0.29524070	-0.4896013	-0.7231891	-0.2087667	-0.7079623	4.39219910	-0.14762035	-1.46175831	-0.3615945	-0.4668165	-0.3300891
Group33	-0.6066620	-0.09841357	-0.2952407	-0.3112110	-0.17045730	-0.19682713	-0.3264009	-0.4821260	-0.1391778	-0.4719749	2.32636182	-0.09841357	-0.55138655	-0.2410630	-0.3112110	-0.2200594
Group34	-0.9592168	-0.15560551	-0.4668165	-0.4920678	-0.26951665	-0.31121102	1.4215799	-0.7623082	-0.2200594	-0.7462578	2.65069899	-0.15560551	-0.63097512	-0.3811541	-0.4920678	-0.3479445
Group35	-0.3235841	-0.19057705	-0.5717312	-0.6026576	-0.33008914	-0.38115410	-0.6320726	0.1374515	-0.2695167	-0.9139754	-1.22841552	-0.19057705	2.17693554	-0.4668165	-0.6026576	-0.4261432
Group36	6.9973470	-0.21448740	-0.6434622	-0.6782687	-0.37150307	-0.42897479	-0.7113742	0.8525979	-0.3033310	-1.0286454	-2.01207012	-0.21448740	-0.76490342	-0.5253847	-0.6782687	-0.4796084
Group37	-0.7430061	-0.12053151	2.4039340	-0.3811541	-0.20876670	-0.24106302	-0.3997578	1.1030521	-0.1704573	1.1519089	0.63815221	-0.12053151	-0.98623554	-0.2952407	-0.3811541	-0.2695167
Group38	-0.6782687	-0.11002971	-0.3300891	-0.3479445	-0.19057705	-0.22005942	-0.3649273	-0.5390333	-0.1556055	-0.5276840	1.87432723	-0.11002971	-0.10556257	-0.2695167	-0.3479445	-0.2460339
Group39	0.6028778	-0.12053151	-0.3615945	-0.3811541	-0.20876670	-0.24106302	-0.3997578	-0.5904814	-0.1704573	-0.5780488	2.40699021	-0.12053151	-0.98623554	-0.2952407	-0.3811541	-0.2695167
Group40	2.2704152	-0.11002971	-0.3300891	-0.3479445	5.05664444	-0.22005942	-0.3649273	-0.5390333	-0.1556055	1.3673897	-0.06333772	-0.11002971	-1.46797946	-0.2695167	-0.3479445	-0.2460339
Group41	3.9324745	-0.19057705	-0.5717312	-0.6026576	-0.33008914	-0.38115410	-0.6320726	0.1374515	3.4408292	-0.9139754	-1.78777121	-0.19057705	0.21045615	-0.4668165	-0.6026576	-0.4261432
Group42	2.3867359	-0.14762035	-0.4428610	-0.4668165	-0.25568595	-0.29524070	-0.4896013	-0.7231891	-0.2087667	-0.7079623	-1.38480163	-0.14762035	1.07695566	-0.3615945	-0.4668165	-0.3300891
Group43	-0.3033310	-0.04920678	-0.1476203	-0.1556055	-0.08522865	-0.09841357	-0.1632004	-0.2410630	-0.0695889	-0.2359874	1.70477473	-0.04920678	-0.65650037	-0.1205315	-0.1556055	-0.1100297
Group44	-0.7430061	-0.12053151	-0.3615945	-0.3811541	-0.20876670	-0.24106302	-0.3997578	-0.5904814	-0.1704573	-0.5780488	2.40699021	-0.12053151	-0.36438015	-0.2952407	-0.3811541	-0.2695167
Group45	-0.8025383	-0.13018891	-0.3905667	-0.4116935	-0.22549381	-0.26037782	-0.4317878	-0.6377928	-0.1841149	-0.6243641	4.51040998	-0.13018891	-1.73693672	-0.3188964	-0.4116935	-0.2911113
Group46	3.2813499	-0.08522865	-0.2556859	-0.2695167	-0.14762035	-0.17045730	-0.2826714	-0.4175334	-0.1205315	-0.4087422	-0.79951559	-0.08522865	-0.25765568	-0.2087667	-0.2695167	-0.1905771
Group47	-0.6782687	-0.11002971	-0.3300891	-0.3479445	-0.19057705	-0.22005942	-0.3649273	-0.5390333	-0.1556055	1.3673897	-0.06333772	-0.11002971	-0.78677101	-0.2695167	5.4000986	-0.2460339
Group48	2.0915763	-0.23080027	0.7518494	-0.7298545	-0.39975780	-0.46160054	-0.7654779	-1.1306858	-0.3264009	-0.2034383	-1.70322570	-0.23080027	1.46728768	-0.5653429	-0.7298545	-0.5160851

5.9 Heat map visualisation

5.9.1 Heat map of groups versus phyla

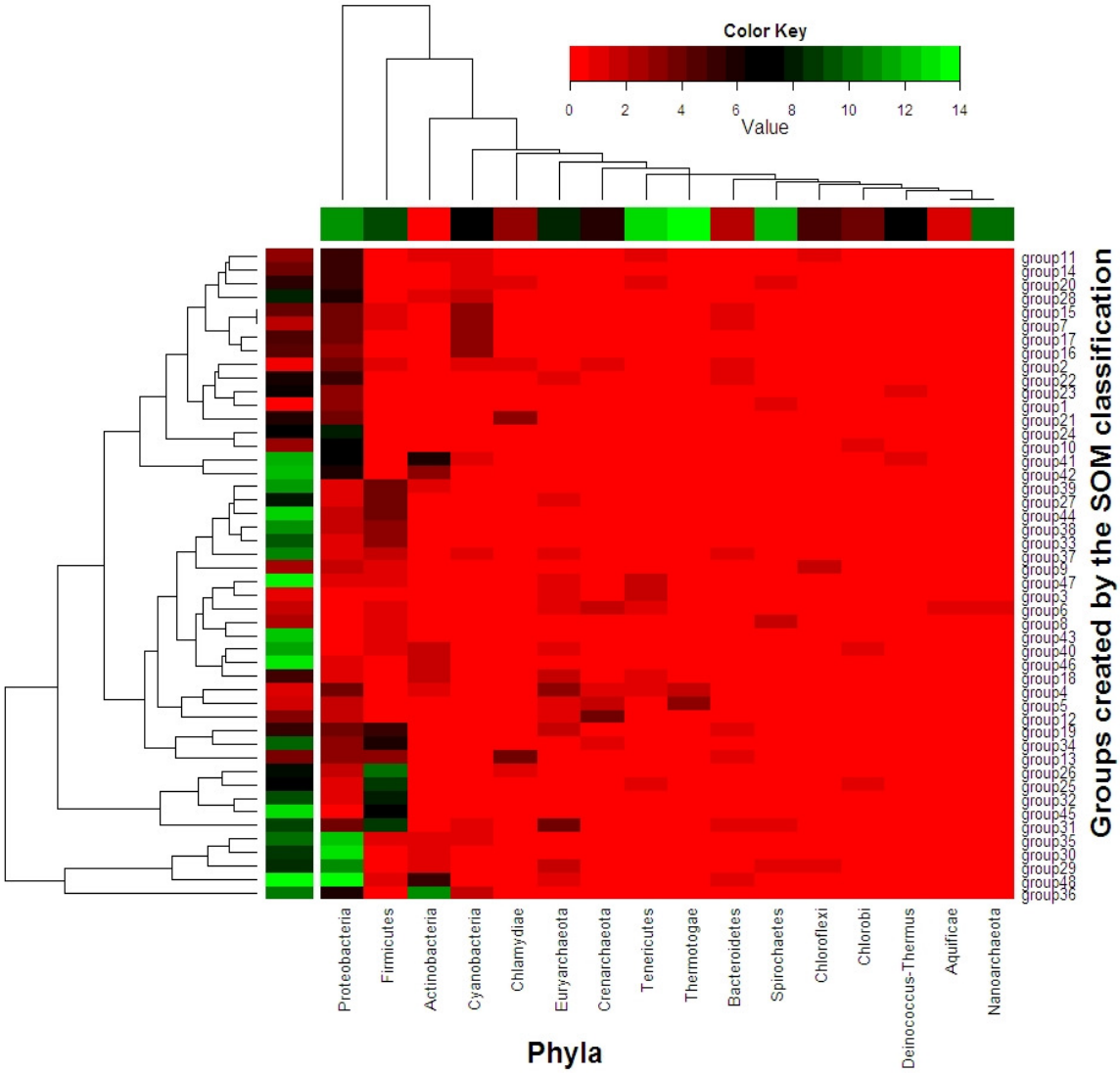


Figure 5.15. Heat map of 48 groups generated by the optimal SOM grid versus 16 different phyla used in the study. The colors scheme used is green-red and the color shade for each tile is proportional to the number of species of each Phylum in a particular group shown by the color key on the top right. Green colored cells have the highest proportion of members and are strongly associated with grouping generated by SOM and red, the least proportion of members and least associated.

The heat map generated to visualise the strength of the association between the clustering generated by the self-organizing map using the optimal SOM grid, 6×8, and the taxonomy information at the phylum level is in Figure 5.15. The heat map was created using a

48×16 matrix from the 48 different groups generated by the optimal SOM grid making the rows and members of the 16 phyla making the columns. The rows and columns were ordered by their respective mean values and dendrograms were added to the top and to the left side. The color of the rectangular tile corresponds to the respective value in the matrix elaborated by the color key. The heat map clearly shows that phyla Proteobacteria and Firmicutes are the most associated with the grouping generated by the SOM as shown by the green colored rectangular tiles which stands for higher values.

5.9.2 Heat map of residuals

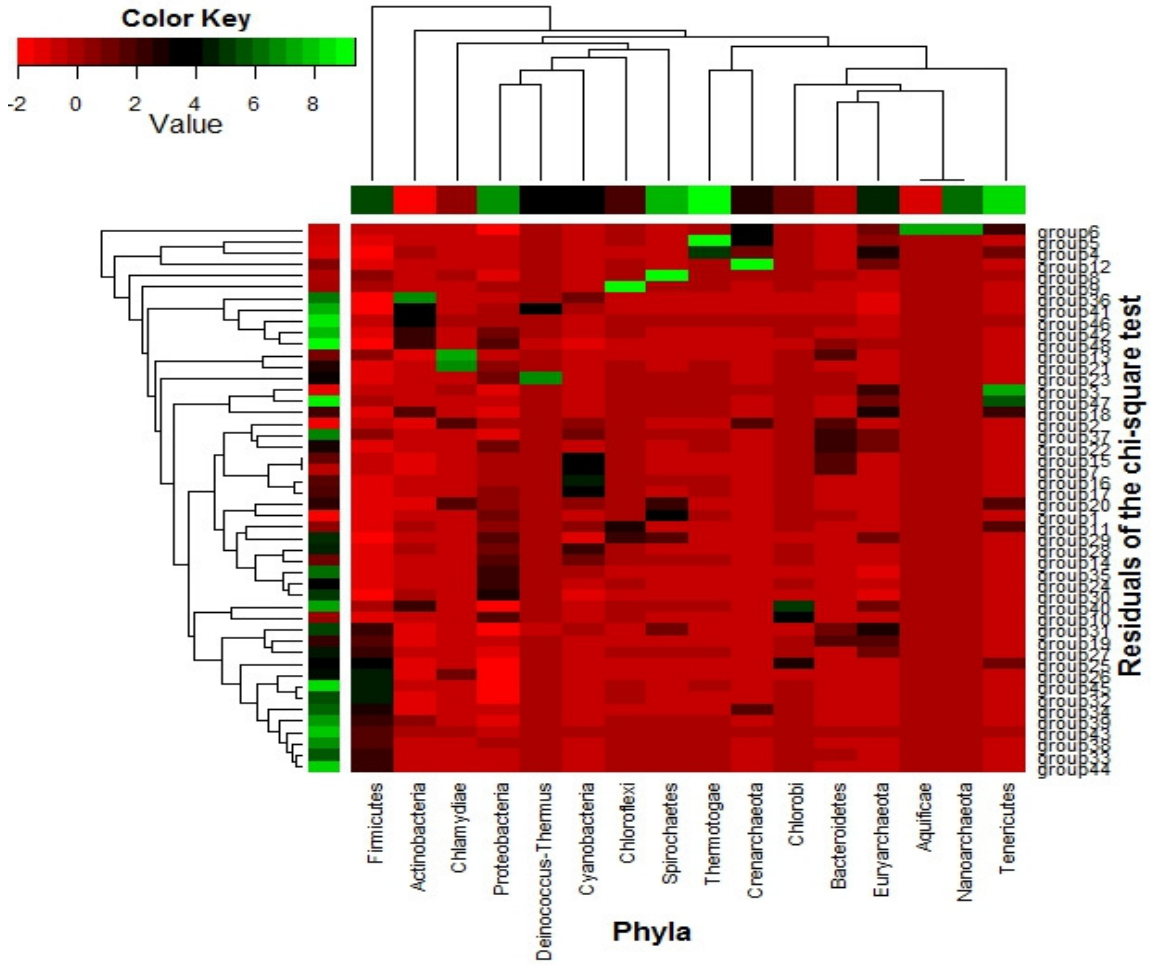


Figure 5.16. Heat maps of the residuals of the χ^2 test of independence of the 48 groups generated by the self-organizing map and the 16 phyla. The values for the color indicated by the tiles are given in the color key on the top-left. Green colored cells shows higher values for residuals and red ones, the lowest.

The residuals of the χ^2 test used for testing independence of the grouping created by the grid, 6×8, and the 16 different phyla were visualised using R function `heatmap()` in Figure 5.16. The heat map highlights the deviations in the standardised residuals. The color key indicating the values of the residuals in each rectangular tile is given on the top-left of the plot. Green colored cells have the highest value for the residuals and the red ones the least. Phylum Chloroflexi in group 9, Spirochaetes in group 8, Thermotogae in group 5 and Crenarchaeota in group 12 are the ones that are highest associated with the grouping.

5.9.3 Heat map of groups versus codons

A heat map was generated using the codebook vectors from the clustering generated by the self-organizing map with the chosen grid (6×8) (Figure 5.17).

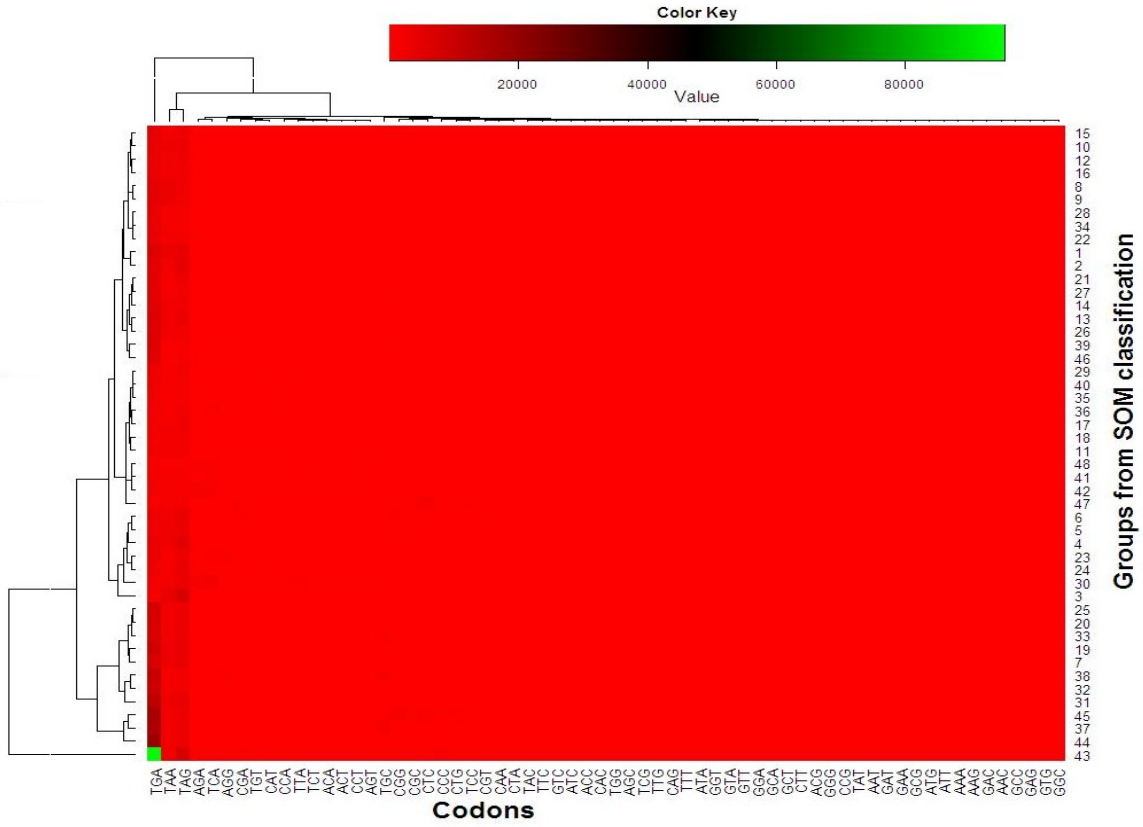


Figure 5.17. Heat map of the 48 groups from the optimal SOM clustering versus the 64 codons (with row and column dendrograms). Y-axis is for the codons and x-axis for the groups. The values of the codebook vectors, as given by the color of the tile, are given in the color key. The codon TGA shows the highest values for the code book vector.

The rows represent the 48 groups generated by the optimal grid 6×8 and the columns shows the 64 codons. This heat map was created with the ordering of both rows and columns and with row and column dendrograms. It shows that the codon TGA is the most overrepresented followed by TAA and TAG.

5.10 Comparison of the overlap numbers with other databases

The Table 5.4 shows the results for the comparison of the overlap numbers with those in two other databases, BPhyOG (Luo *et al*., 2007) and PairWise Neighbours ((Pallejà *et al*., 2009). There were 107 microbial species in common between the three studies. The total overlap numbers from this study were exactly the same as in the BPhyOG database for 24 species and rest of them, except a few, had almost similar overlap numbers. But the total overlap numbers in the PairWise Neighbours database are considerably greater than that from the other two studies. The total overlap numbers of these 107 species were arranged ascendingly with respect to the PairWise Neighbours database and were plotted along with the overlap numbers for those species from this study and BphyOG in Figure 5.18.

Table 5.4. Comparison of the number of overlaps from this study to those available from the BPhyOG (Luo *et al*., 2007) and PairWise Neighbours (Pallejà *et al*., 2009). The cases having same overlap numbers between this study and that given by the BPhyOG are highlighted in yellow.

SpeciesName	*GenBank identifiers*	*GenBank accession numbers*	*BPhyOG*	*Current study*	*Pairwise Neighbours*
Acinetobacter sp. ADP1	50083297	NC_005966	317	316	370
Aeropyrum pernix K1	118430835	NC_000854	514	297	315
Aquifex aeolicus VF5	15282445	NC_000918	587	587	637
Archaeoglobus fulgidus DSM 4304	11497621	NC_000917	825	820	877
Bacillus anthracis str. Ames	30260195	NC_003997	493	499	554
Bacillus anthracis str. Ames Ancestor	50196905	NC_007530	495	495	555
Bacillus cereus ATCC 10987	42779081	NC_003909	614	611	685
Bacillus halodurans C-125	57596592	NC_002570	468	472	536
Bacillus subtilis subsp.subtilis str.168	255767013	NC_000964	473	482	532
Bacillus thuringiensis serovar konkukian str.97-27	49476684	NC_005957	482	487	549
Bacteroides thetaiotaomicron VPI-5482	29345410	NC_004663	364	367	413
Bartonella henselae str. Houston-1	49474831	NC_005956	212	199	233
Bartonella quintana str. Toulouse	49473688	NC_005955	128	111	140
Bdellovibrio bacteriovorus HD100	42521650	NC_005363	627	627	765
Bifidobacterium longum NCC2705	58036264	NC_004307	152	153	186
Bordetella bronchiseptica RB50	33598993	NC_002927	885	885	973

Bordetella pertussis Tohama I	33591275	NC_002929	568	503	560
Borrelia burgdorferi B31	15594346	NC_001318	189	188	210
Bradyrhizobium japonicum USDA 110	27375111	NC_004463	1331	1335	1438
Campylobacter jejuni subsp.*jejuni* NCTC 11168	15791399	NC_002163	534	519	621
Caulobacter crescentus CB15	16124256	NC_002696	703	704	810
Chlamydia muridarum Nigg	29337300	NC_002620	175	175	184
Chlamydophila caviae GPIC	29839769	NC_003361	187	187	201
Chlamydophila pneumoniae CWL029	15617929	NC_000922	187	186	197
Chlamydophila pneumoniae J138	15835535	NC_002491	198	198	209
Chlamydophila pneumoniae TW-183	33241335	NC_005043	214	214	226
Chlorobium tepidum TLS	21672841	NC_002932	335	334	357
Chromobacterium violaceum ATCC 12472	34495455	NC_005085	558	561	670
Clostridium acetobutylicum ATCC824	15893298	NC_003030	304	305	356
Clostridium perfringens str.13	18308982	NC_003366	143	143	174
Corynebacterium glutamicum ATCC 13032	58036263	NC_003450	362	362	434
Desulfotalea psychrophila LSv54	51243852	NC_006138	432	413	471
Enterococcus faecalis V583	29374661	NC_004668	377	379	460
Escherichia coli CFT073	26245917	NC_004431	1388	1280	1528
Escherichia coli O157-H7 EDL933	16445223	NC_002655	810	776	977
Escherichia coli O157:H7 str. Sakai	15829254	NC_002695	873	846	1069
Geobacter sulfurreducens PCA	39995111	NC_002939	524	524	583
Gloeobacter violaceus PCC 7421	37519569	NC_005125	632	632	700
Haemophilus ducreyi 35000HP	33151282	NC_002940	207	207	273
Haemophilus influenzae Rd KW20	16271976	NC_000907	150	150	228
Halobacterium sp. NRC-1	15789340	NC_002607	329	331	388
Helicobacter hepaticus ATCC 51449	32265499	NC_004917	328	330	406
Helicobacter pylori 26695	15644634	NC_000915	278	279	337
Helicobacter pylori J99	15611071	NC_000921	245	245	312
Lactobacillus plantarum WCFS1	28376974	NC_004567	317	304	369
Lactococcus lactis subsp.lactis Il1403	15671982	NC_002662	271	271	326
Leifsonia xyli subsp.xyli str. CTCB07	50953925	NC_006087	253	253	307
Listeria monocytogenes str.4b F2365	85700163	NC_002973	322	321	371
Mesoplasma florum L1	50364815	NC_006055	65	65	99
Mesorhizobium loti MAFF303099	57165207	NC_002678	1102	1100	1239
Methanocaldococcus jannaschii DSM 2661	15668172	NC_000909	253	242	265
Methanococcus maripaludis S2	45357563	NC_005791	92	91	109
Methanopyrus kandleri AV19	20093440	NC_003551	459	455	473
Methanosarcina acetivorans str. C2A	20088899	NC_003552	466	474	512
Methanosarcina mazei Go1	21226102	NC_003901	279	276	304
Mycobacterium avium subsp.*paratuberculosis* K-10	41406098	NC_002944	878	878	971
Mycobacterium bovis AF2122/97	31791177	NC_002945	823	795	904
Mycobacterium tuberculosis CDC1551	50953765	NC_002755	981	974	1074
Mycobacterium tuberculosis H37Rv	57116681	NC_000962	784	818	929
Mycoplasma mobile 163K	47458835	NC_006908	105	104	147
Mycoplasma mycoides subsp.mycoides SC str. PG1	127763381	NC_005364	102	103	141
Nanoarchaeum equitans Kin4-M	38349555	NC_005213	227	227	243
Neisseria meningitidis MC58	77358697	NC_003112	160	157	185
Neisseria meningitidis serogroup A strain Z2491	15793034	NC_003116	206	166	227
Nostoc sp. PCC 7120	17227497	NC_003272	298	299	336
Oceanobacillus iheyensis HTE831	23097455	NC_004193	347	347	409

Onion yellows phytoplasma OY-M	255961248	NC_005303	46	52	57
Pasteurella multocida subsp.multocida str. Pm70	15601865	NC_002663	167	165	232
Picrophilus torridus DSM 9790	48477072	NC_005877	430	430	462
Porphyromonas gingivalis W83	34539880	NC_002950	278	277	299
Prochlorococcus marinus str. MIT 9313	33862273	NC_005071	358	357	420
Propionibacterium acnes KPA171202	50841496	NC_006085	501	497	570
Pseudomonas aeruginosa PA01	110645304	NC_002516	754	753	888
Pyrobaculum aerophilum str. IM2	18311643	NC_003364	748	748	807
Pyrococcus furiosus DSM 3638	18976372	NC_003413	675	665	719
Rickettsia conorii str. Malish 7	15891923	NC_003103	213	212	225
Salmonella enterica subsp.enterica serovar Typhi str. CT18	16758993	NC_003198	651	612	781
Shewanella oneidensis MR-1	24371600	NC_004347	390	389	515
Shigella flexneri 2a str.2457T	30061571	NC_004741	605	599	739
Sinorhizobium meliloti 1021	15963753	NC_003047	390	415	471
Staphylococcus aureus subsp.aureus MRSA252	49482253	NC_002952	308	299	374
Staphylococcus aureus subsp.aureus MSSA476	49484912	NC_002953	299	294	371
Staphylococcus aureus subsp.aureus Mu50	57634611	NC_002758	318	313	384
Staphylococcus aureus subsp.aureus MW2	21281729	NC_003923	302	302	376
Staphylococcus aureus subsp.aureus N315	29165615	NC_002745	288	293	361
Staphylococcus epidermidis ATCC 12228	27466918	NC_004461	299	299	359
Streptococcus agalactiae 2603V/R	22536185	NC_004116	287	288	361
Streptococcus agalactiae NEM316	25010075	NC_004368	283	267	345
Streptococcus mutans UA159	24378532	NC_004350	267	266	326
Streptococcus pneumoniae TIGR4	194172857	NC_003028	313	313	370
Streptococcus pyogenes M1 GAS	15674250	NC_002737	235	237	294
Streptococcus pyogenes MGAS315	21909536	NC_004070	290	292	347
Streptococcus pyogenes SSI-1	28894912	NC_004606	275	277	325
Sulfolobus tokodaii str.7	24473558	NC_003106	862	861	898
Thermoplasma volcanium GSS1	13540831	NC_002689	275	273	304
Thermosynechococcus elongatus BP-1	22297544	NC_004113	465	468	508
Thermotoga maritima MSB8	15642775	NC_000853	724	724	759
Thermus thermophilus HB27	46198308	NC_005835	666	668	733
Treponema denticola ATCC 35405	42516522	NC_002967	484	484	554
Treponema pallidum subsp.pallidum str. Nichols	15638995	NC_000919	276	275	289
Tropheryma whipplei str. Twist	28572175	NC_004551	196	112	131
Tropheryma whipplei TW08/27	32447382	NC_004572	112	192	213
Ureaplasma parvum serovar 3 str. ATCC 700970	13357558	NC_002162	75	75	96
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	32490749	NC_004344	67	65	73
Wolbachia endosymbiont of Drosophila melanogaster	42519920	NC_002978	159	159	173
Xanthomonas campestris pv.campestris str. ATCC 33	21229478	NC_003902	705	693	765
Yersinia pestis strain CO92	16120353	NC_003143	425	414	508

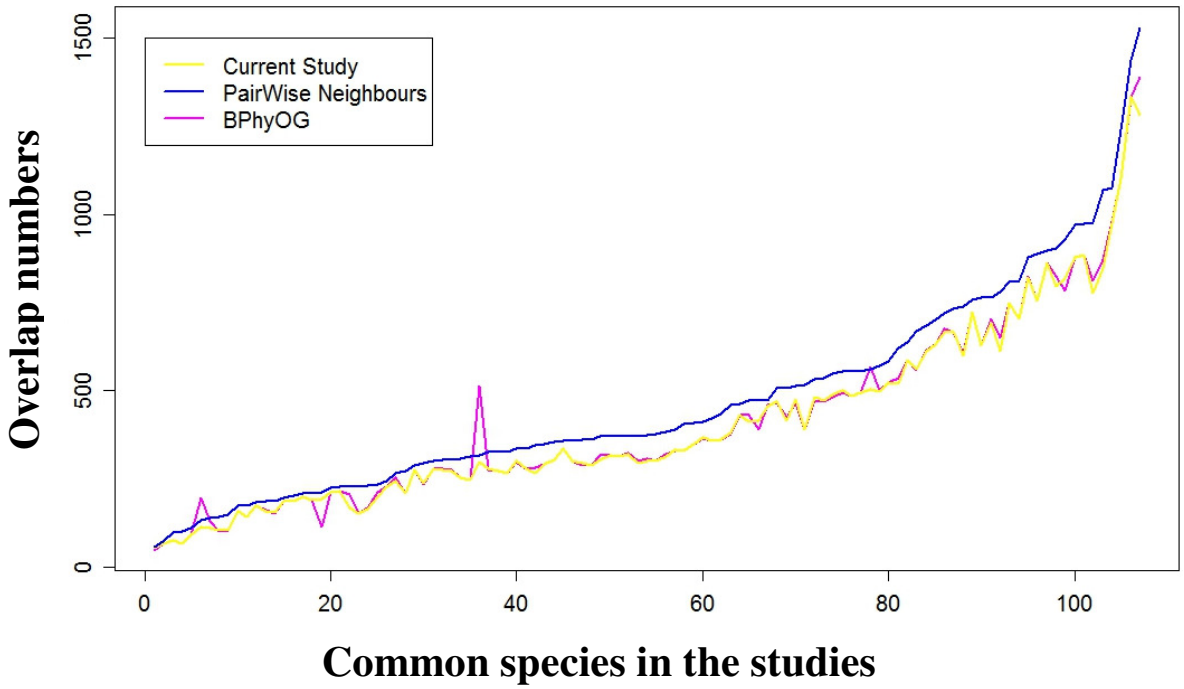


Figure 5.18. Comparison of the overlap numbers obtained from the current study with those from the BPhyOG and PairWise Neighbours database. The total overlap numbers from the PairWise Neighbours database were ordered ascendingly and plotted along with the overlap numbers for the respective species from this study and BPhyOG. Yellow line is for this study, blue for PairWise Neighbours and magenta for BPhyOG.

6 Discussion

This study was aimed to find the overlapping genes in the completely sequenced microbial genomes computationally, calculate the codon usage by the codons in the overlapping parts and to check whether there is any codon usage bias in the overlapping part of the genomes compared to the overall codon usage of the genome. The codon usage for the genomes were obtained from the codon tables of the codon usage database (Nakamura *et al*., 2000), the codon usage tabulated from the GenBank (Nakamura *et al*., 1997; Benson *et al*., 2010). A total of 435 microbial species were included in this study for which the complete genomes were available from the NCBI's (Sayers *et al*., 2010) GenBank and for which the codon usage (Aota *et al*., 1988; Nakamura *et al*., 1997) were available in the CUTG database.

Studies on the overlapping genes have been conducted ever since their discovery in the bacteriophage ΦX174 (Barell *et al*., 1976; Sanger *et al*., 1977), which came out as a surprise while sequencing its genome. The occurrence of genomic overlaps added to the crisis of the linear model of the gene (Portin, 1993; Portin, 2009). Several studies have been conducted to monitor the existence of genomic overlaps in viruses, prokaryotes and eukaryotes. Fukuda *et al*., (Fukuda *et al*., 2003) analysed the whole genomic sequences of 50 bacterial species and in 2004, Johnson and Chisholm (Johnson and Chisholm 2004) analysed 198 microbial genomes, both from NCBI's GenBank, to find overlapping genes. Luo *et al*. created a server named BPhyOG (Luo *et al*., 2007) containing information about the overlapping genes from a total of 177 prokaryotic genomes, which were used to to reconstruct phylogenies.

In the present work, overlapping genes were defined as pairs of adjacent genes whose coding sequences overlap partially or completely and they were found based on the annotations from the genomic files of the selected microbes obtained from the GenBank. The 3' and 5' untranslated regions are not taken into account in the current study for finding the overlapping gene. Studies conducted by Luo *et al*., Fukuda *et al*., and Johnson and Chisholm also defined overlapping genes pairs as adjacent genes, located on either

DNA strand sharing one or more nucleotides in their coding sequence (Fukuda et al., 2003; Luo et al., 2007; Johnson and Chisholm 2004). On the other hand, studies conducted by Pallejà et al., which lead to the creation of the PairWise Neighbours database used computational prediction of the Shine-Dalgarno (SD) sequence by calculating the base pairing free energy between translation initiation regions and the 16S rRNA 3' tail and the annotations of the coding sequences from the genomic file for finding the genomic overlaps (Pallejà et al., 2009). The Shine-Dalgarno (SD) sequence is responsible for an efficient translation and can be used to determine the coding sequences (Ma et al., 2002; Ponnala, 2010). In the study conducted by Luo et al., the accession number or species name of the genomic sequence of prokaryotes were used as the input for calculating the overlapping genes and for generating the phylogenetic tree. We used the GenBank identifiers for calculating the number of overlapping genes.

Current study analysed 435 species and overlapping genes were present in all of them although the 3' and 5' untranslated regions were not taken into account. This is in agreement with the findings from the other studies, showing that overlapping genes are ubiquitous in the microbial genomes, as approximately one third of the genes in all the microbial genomes are overlapping (Fukuda *et al*., 2003; Johnson and Chisholm 2004; Luo *et al*., 2007). The complete genomes of the prokaryotes used in this study were collected from the NCBI's GenBank Release 176.0 whereas Luo *et al*. collected from the NCBI ftp server in August 2004 (Luo *et al*., 2007).

Although the genomes of prokaryotes are compact without introns, a couple of exceptions were found in this study indicated by the `joined split` location of the coding sequence in the GenBank annotations. Such kind of locations was found in 13 of the genomes that were excluded from the analysis. Out of them, 9 `joined split` location were due the circular nature of their genome with coding potential of the sequences spanning from near the end position of the coding sequence to the position after the start. Four of them contained introns viz., *Clostridium difficile* 630, *Cyanothece* sp. PCC 8801, *Pyrococcus horikoshii* OT3 and *Methylococcus capsulatus* str. Bath. Also, there were 23 genomes with multiple chromosomes that were excluded from the analysis. The names of

these species are listed in the Figure 6.1. Three of them having 3 chromosomes are given first in the list and the other 20 species have 2 chromosomes.

Burkholderia cenocepacia AU 1054
Burkholderia cenocepacia HI2424
Burkholderia sp.383
Brucella suis 1330
Brucella melitensis biovar Abortus 2308
Deinococcus radiodurans R1
Haloarcula marismortui ATCC 43049
Leptospira borgpetersenii serovar Hardjo-bovis JB197
Leptospira borgpetersenii serovar Hardjo-bovis L550
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
Leptospira interrogans serovar Lai str.56601
Ochrobactrum anthropi ATCC 49188
Paracoccus denitrificans PD1222
Photobacterium profundum SS9
Pseudoalteromonas haloplanktis TAC125
Ralstonia eutropha JMP134
Rhodobacter sphaeroides 2.4.1
Rhodobacter sphaeroides ATCC 17029
Vibrio cholerae O395
Vibrio fischeri ES114
Vibrio parahaemolyticus RIMD 2210633
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016

Figure 6.1. The list of microbial species that were excluded from the analysis due to the presence of multiple chromosomes in their genomes. The first three organisms have three chromosomes and the others have two chromosomes.

An attempt was made to compare the overlap numbers from this study to those in the web servers, BPhyOG (Luo *et al*., 2007) and PairWise Neighbours (Pallejà *et al*., 2009). The results are presented in the Table 5.3. The comparison was done for all the species that were common between the two databases, viz., BPhyOG and PairWise Neighbours, and this study. Altogether, there were 107 species common between the BPhyOG comprising a total of 177 genomes, PairWise Neighbours database consisting 678 genomes and the 435 different genomes included in this study. The numbers of overlaps were the same in both this study and the BPhyOG for 24 out of the 107 species compared. The difference in the number of the overlapping genes may be due to the newer versions and updates of the GenBank release that may cause alterations in the size of the genome and the coding

sequences as revealed by the current annotations in GenBank and those for the same species in the study conducted by Luo *et al* (Luo *et al*., 2007). Hence, the current study gives the updated picture about the overlap numbers in the studied species. The overlap numbers in PairWise Neighbours database are considerably greater for all the compared species (Figure 5.18). The overlap numbers from this study and BPhyOG are almost similar except for a few big differences which can be assumed to be due to the updates of the annotations and the newer versions of the genomic files. The differences in PairWise Neighbours data arise largely arise from including the 5' Shine-Dalgarno sequences.

It has been proved that the proportions or the numbers of overlapping genes are decided by the niche and life style, i.e., adaptive pressure of the organism. In general, obligatory intracellular parasites show reduction of the genome size by forming more genomic overlaps than the free-living organisms (Sakharkar *et al*., 2005). Reducing the genome size will result in the formation of increased amount of pseudo genes as the organism gets adapted to the host resulting in convergent evolution (Sakharkar and Chow, 2005).

Studies have also been done to analyse the significance of the presence of overlapping genes, e.g., the natural anti-sense transcript study in prokaryotes (Lacatena and Cesareni, 1981; Itoh and Tomizawa, 1980) helped to find the importance of genomic overlaps in the cell regulation. Overlapping genes are conserved in closely related species and studies of Luo *et al*. on γ-Proteobacteria genomes concluded that overlappping genes can be used as genomic markers for estimating the phylogentic distance of completely sequenced microbial genomes (Luo *et al*., 2006). One of the databases that have been constructed with this aim is BPhyOG, (Luo *et al*., 2007) with overlapping genes from 177 bacterial species, and it estimates the phylogenetic trees based on the overlapping gene pairs. Another database is OGtree (Jiang *et al*., 2008) that reconstructs the genome tree of some prokaryotes based on the distance between the overlapping genes by using the information from both the overlapping gene content and order. The updated version of the database OGtree2 (Cheng *et al*., 2010) takes into account genomic rearrangements and uses regulatory regions to define the overlapping gene for reconstructing genome trees more precisely.

Strand-wise overlap distribution was made to examine the proportion of the same strand overlaps in the microbial species chosen for the analyses. From the study on strand distribution, it became evident that, unidirectional overlaps were the only type of overlaps present in this study, when only the overlaps of length above nine bases were taken into account. When all the overlaps are taken into account without considering any threshold for length, overlaps are present on both strands. This agrees with the previous studies which showed that unidirectional overlaps are far more common than the overlaps in the opposite strands in prokaryotes (Fukuda *et al*., 2003; Sabath *et al.,* 2008; Cock and Whitworth, 2007, 2010) as indicated in the Figure 5.7. Usually all the genes in prokaryotic operons are transcribed in the same direction. The presence of only unidirectional overlaps among overlaps with length above nine bases in the genomes of analysed microbes indicate that these overlapping genes may have a tendency to be the grouped into operons which may then transcribed and translated together leading to the regulation of gene expression by transcriptional and translational coupling and thereby playing a role in the evolution of the genes as described by the other studies (Normark *et al*., 1983; Chen *et al*., 1990; Inokuchi *et al*., 2000; Krakauer and Plotkin, 2002; Sakharkar *et al*., 2005). As unidirectional overlaps are widely conserved in prokaryotes indicated by other studies, conserved operons for the genomic unidirectional overlaps in the current should also show functional associations and hence can be used to predict the links between the overlapping gene pairs (Dandekar *et al*., 1998; Overbeek *et al*., 1999; Johnson and Chisholm, 2004).

There are different opinions about the formation of unidirectional overlaps. Some of the studies show that unidirectional overlaps are formed due to the loss of stop codons or frame shift or mutation at the 3' end (Sakharkar *et al*., 2005). On the contrary, another study based on the reading frame bias exhibited by the unidirectional overlaps states that these overlaps are developed by the extension of the N terminal end or 5' end (Cock and Whitworth, 2010). It has been found that amongst the unidirectional overlapping genes, the most prominent overlap length is 4 bases that must have created as a result of the fusion of the termination codon of one gene with the initiation codon of the other

indicating the role it plays in transcriptional coupling. Thus the neighbouring gene play a role in bringing the neighbouring genes in contact with the translational machinery and hence regulate the gene as in trp operon where TrpC and TrpF overlap by several base pairs (Zheng *et al*., 2002).

The one of the aims of the current study was to find the codon usage by the overlapping part in the analysed genomes. Codon usage of an organism is very important as it serves as a crucial indicator of molecular evolution (Duret, 2002). It has been proved that there are similar codon preferences between related microbial species (Ikemura and Ozek, 1983) and difference in codon preferences between taxonomically unrelated groups and even between the different genes within species (Sharp *et al*., 1986; Shields and Sharp, 1987; Nakamura *et al*., 2000). These patterns of codon usage among the different species and genes can be distinguished by the codon usage and hence helps to characterize different species, determine open reading frame, find optimal protein expression etc.

The current study was focused on finding whether the codons in the overlapping part of the genomes are used with the same frequency as compared to the overall usage of the codons in the genomes. The difference in the frequency of usage of the alternative synonymous codons constitute what is called the codon usage bias and the main interest in the study was to find patterns of codon usage bias by the overlapping part of the genomes in the phylum level of taxonomy. The codon usages of the analysed genomes were obtained through the codon frequency tables available from the CUTG database, which contains the codon usages tabulated from the GenBank (Aota *et al*., 1988). The codon usage from the overlapping parts were calculated by counting the codons from the overlapping parts after taking into account of the phase of the coding sequences involved in the overlap.

The relative normalized codon usage percentages were calculated from the codons in the overlapping part of the genomes to measure the codon usage bias. The number of each of the 64 different codons used by the overlapping part of the genes in a genome were compared with the number of codons used by the genome obtained from the codon usage

table of that species and the relative normalized percentages were calculated keeping the expected percentages from the codon usage table as 100 %. While calculating the relative percentages of codons in the overlapping part, it was observed that some species had unexpected codons, the codons that were not used by the genome as indicated by the codon usage table from the codon usage database. Ambiguous codons were present in *Escherichia coli* CFT073, *Escherichia coli* O157:H7 EDL933, *Bacillus cereus* ATCC 10987, *Chlamydophila pneumoniae* AR39, *Mycobacterium tuberculosis* CDC1551 and *Treponema pallidum* subsp. pallidum str. Nichols.

Different measures that have been used in the previous studies for the calculation of the codon usage bias include e.g.

- N_c, the effective number of codons in a gene measures the bias in the equal usage of codons (Wright, 1990). The value is between 20 and 61, the number of codons in the universal genetic code coding for amino acids, according to the usage of the genetic code. The value of 20 means that the codon is extremely biased and a value of 61 indicate no bias (all codons are used equally).
- Relative synonymous codon usage (RSCU) is a measure of synonymous codon usage variation among the genes. It is the ratio of the observed frequency of codons to the expected if all the synonymous codons for those amino acids are used equally (Sharp and Li, 1987; Hassan *et al*., 2009).
- Codon adaptation index (CAI) (Sharp and Li 1987) is used for summarizing codon usage by comparing all the genes in the genome with an optimal codon usage inferred from a reference set of presumed high-expression genes.
- Multivariate analyses like principal component analysis and correspondence analyses that reduce the dimensions of the codon usage data keeping the variation have been used for analysing and visualising the codon usage trends in the genomes (Grocock and Sharp, 2002; Angellotti *et al*., 2007; Hassan *et al*., 2009).

The χ^2 test was used in the current study for testing the goodness of fit of the observed frequency of the codons from the overlapping part with the expected frequency of codons as given from the codon usage table, analogous to other studies trying to compare the codon usage differences (McInerney, 1998). The results of the χ^2 test gave a significant difference (p-value <0.001) in the codon usage by the overlapping part for all the analysed species. The five most biased codons and the five least biased codons from all the species were obtained computationally by comparison of the number of codons from the overlap part to that available from the codon usage tables at CUTG. Out of the 413 organisms studied, the most biased codon was TGA. 73% of the analysed species had TGA as their most biased codon, 18.6% of the species had TAG as the most deviating codon and 4.85% of the species had TAA as the most deviating codon. The second most biased codon present in the majority of the investigated microbes was TAG. 56.5% of the total analysed species had TAG, 22.3% had TAA and 14.5 % had TGA as the second most deviating codon. The codons TGA, TAA and TAG are the terminator codons for the translation table number 11, the bacterial, archaeal, and plant plastid code. Hence, the terminator codons, TGA, TAA and TAG, are the most biased codons exhibited by the overlapping part in the analysed genomes. This finding coincides with the fact that the overlapping part are formed by the conjunction of the termination codon of one gene with the initiation codon of the other indicating the role these overlaps play in the regulation of gene expression by transcriptional coupling in line with the results of other studies in this field (Krakauer and Plotkin, 2002; Sakharkar *et al*., 2005; Cock and Whitworth, 2010).

As there was bias in the codon usage by the overlapping part of the genomes, in order to find whether there are any patterns in the codon usage bias by the overlapping part in the phylum level of taxonomy, multivariate analyses such as principal component analysis, self-organizing maps and correspondence analysis along with the visualisations using the heat maps were used. Multivariate analysis methods such as correspondence analysis and principal component analysis have also been used to study heterogeneous codon usage in a variety of species (Grantham *et al*., 1980; Greenacre, 1984; Kanaya *et al*., 1996a; Kunst *et al*., 1997; Angellotti *et al*., 2007; Hassan *et al*., 2009). The unsupervised artificial

neural network algorithm, self-organizing maps has been used for analysing codon usage in genes from bacteria, cluster and visualise genes of individual species at a higher resolution than that from the principal component analysis (Kanaya *et al*., 2001; Wang *et al*., 2001; Supek and Vlahovicek, 2004).

Although the principal component analysis and the correspondence analysis did not give any significant patterns of codon usage in the phylum level of taxonomy in this study, the clustering using the SOM with the optimal grid 6×8, the grid with the lowest p-value from the χ^2 test of independence, yielded a significant result for classification of the codon usage bias patterns. The grid used for the creating the self-organizing map has got significant role in the classification results as indicated by the previous studies (Wang *et al*., 2001; Supek and Vlahovicek, 2004). As the classifications with the different grids showed a significant result, it should be inferred from this study that there is a definite trend in the codon usage bias pattern in the phylum level of taxonomy.

On the other hand, other studies on the codon usage patterns concluded that the similarity between species in average codon usage is a short range phenomenon, generally rapidly diminishing beyond the genus level (Mitreval *et al*., 2006). Heat map visualisations were generated for visualising the strength of the association between the clustering given by the self-organizing map using the optimal SOM grid, 6×8, and the taxonomy information at the phylum level of taxonomy and is given in Figure 5.15. The phyla Proteobacteria, Firmicutes and Actinobacteria are the most associated with the grouping generated by the SOM. The residuals of the χ^2 test of independence of the grouping were also visualised using the heat maps (Figure 5.16) to explore the phyla that affected the classification the most. The SOM classification and the heat maps clearly indicate the stop codon usage bias by the overlapping part of the genomes compared to the normal parts of the genome indicating a role in transcriptional coupling and regulation of gene expression.

Codon usage bias can arise due to a number of evolutionary processes. The GC (guanine and cytosine) content of the genome which is influenced by the mutational pressure is the main factor influencing codon usage and amino acid composition (Sueoka, 1988; Wilquet and Van de Casteele, 1999; Mitreval et al., 2006; Hu *et al*., 2007; Hassan *et al*., 2009).

Other important factors that affect the codon usage patterns in various organisms are abundance of tRNAs (Ikemura, 1985) for highly expressed mRNAs that need translational efficiency (Gouy and Gautier,1982; Sharp and Li, 1987; Goetz and Fuglsang, 2005), selection on the mRNA secondary structure stability (Chamary and Hurst, 2005), facilitation of correct co-translational protein folding in vertebrates (Oresic *et al*., 2003), directional and strand specific mutational bias (Gupta and Ghosh, 2001; Gupta *et al*., 2004; Guo and Yuan, 2009), environmental factors (Lobry and Necsulea, 2006; Basak *et al*., 2007), secondary structure of proteins(Wright, 1990; Gu *et al*., 2004), replicational and transcriptional selection (Gupta *et al*., 2004), translational selection (Hassan *et al*., 2009; Sharp *et al*., 2010), horizontal gene transfer (Grocock and Sharp, 2002; Guo and Yuan, 2009) etc.

Codon bias is positively correlated to selection against missense and nonsense mutations and to gene length (Qin *et al*., 2004; Stoletzki and Eyre-Walker, 2007; Tuller *et al*., 2010). The two different types of translational errors that can occur are missense mutations where a wrong amino acid is incorporated into the peptide chain which can have various deleterious effects or can predispose to other mutations (Ninio 1991).and nonsense mutations forming stop codons in the coding regions leading to premature termination of peptide synthesis. Synonymous codons usage can affect the rate of missense and nonsense errors. Codon bias is expected to be highest in codons that encode the most important amino acid sites as selection is acting to minimize translational errors. Highly conserved genes and sites have higher codon bias indicating selection against missense errors. Also codon bias increases along the length of genes, indicating selection against nonsense errors.

The details of the genomes that were used in this analysis including their scientific names and the GenBank identifiers are given in the Appendix. The study can be extended by creating database to include the information about the overlap numbers, their strand and phase information, the patterns of bias shown by the codons in the overlapping part in the studied 16 Phyla. Analysing the amino acid usage patterns by the codons in the overlap part can also provide further insights into selection against the missense and nonsense

mutations. A possible improvement of the study can be done by taking into account the untranslated regions while finding the overlapping genes. As stated by other studies on the overlapping genes, the overlap numbers and extent in each of the analysed microbes and hence the codon usage are also are affected by the presence of the annotation errors like the misprediction of start and stop codons, problems in the identification of the open reading frame etc (Devos and Valencia, 2001; Pallejà *et al*., 2008, 2009).

7 Conclusions

The ultimate goal of the study was to find the codon usage bias patterns, if any, exhibited by the codons in the overlapping part of the analysed microbial genomes. In order to accomplish the goal, the presence and prevalence of overlapping genes in the selected microbial species was investigated using the annotations from the GenBank. The analyses using the Perl scripts indicated that there are considerable numbers of overlaps in all of the studied organisms. Threshold for the overlap length was kept at 9 base pairs and only those overlaps passing this threshold were chosen for calculating the codon usage, codon usage bias analyses and statistical significance tests. Unidirectional or same strand overlaps were only overlap type present among the overlaps of length greater than nine base pairs. These points to the fact that these overlapping genes in the analysed microbes are predisposed to be organised into operons, as all the genes in an operon are transcribed and translated in the same direction. By this mechanism of transcriptional and translational coupling, it is evident that the analysed overlapping genes would be able to exert a role in the regulation of gene expression and also in the evolution of the gene.

There are differences in the numbers of overlapping gene in the microbial species from this study, BPhyOG and Pairwise Neighbours. Some of the overlap numbers given by the BPhyOG were exactly the same as that calculated in this study and the others were in a similar range. But the overlap numbers given by the PairWise Neighbours for almost all of the organisms are greater due to the different definition used for the coding regions by including the Shine-Dalgarno sequence.

Codon usage of the overlapping part of the genomes was obtained by counting all the codons from the overlapping part of the genomes after calculating their relative phase. Comparison of the codon usage by the codons in the overlapping part of the genomes to the overall codon usage of the genomes obtained from the codon usage tables at the CUTG, codon usage database, showed a significant difference in the codon usage patterns (p-value <2.2e-16 from the χ^2 test) for all of the studied organisms.

Taxonomy information for the analysed microbes at different levels were obtained and the phylum level of taxonomy was used to study the codon usage bias patterns within the phyla. Analyses of the codon usage by the overlapping part of the genomes showed a very clear bias in the codon usage in the overlapping part compared to the overall codon usage of the genome from the codon frequencies of the codon usage table at CUTG database. The most biased codons in the overlapping part were the stop codons. The stop codon, TGA, was the most biased codon present in most of (73%) the analysed species and TAG was the second most biased codon present in the overlap part of the analysed microbes (56.5 % of the total analysed species).

SOM visualization generated by the optimal grid makes it evident that there are significant codon usage bias patterns in the phylum level of taxonomy and the heat map visualization shows that the phyla Proteobacteria, Firmicutes and Actinobacteria are the most associated with the codon usage bias patterns generated by the SOM. As principal component analysis and correspondence analysis did not yield any significant pattern for the codon usage bias in the phylum level of taxonomy and as self-organizing map showed a significant bias patterns, it can be inferred from this study that the self-organizing maps are more effective than the principal component analysis and correspondence analysis when analysing the codon usage patterns of a large number of genes.

Codon usage bias can be attributed to a number of factors, most important of which are the selection against translational errors (selection for translational efficiency), GC content of the genome, strand specific mutational bias, environmental factors, horizontal gene transfer etc. Also there is positive correlation between codon bias to selection against mis-sense and non-sense mutations and to gene length. Hence this stop codon bias exhibited by the overlaps in the analysed microbes points to the fact that they play a substantial role in the selection against translational errors viz., nonsense and mis-sense mutations.

Overlapping genes in prokaryotes are highly conserved as any change in the overlapping part affect both the genes involved in the overlap. Biased codon usage is highest in the most conserved genes and sites and it also depend on the level of gene expression. As the genomic overlaps are highly conserved, a greater bias in the codon usage in this part compared to the other parts is expected. Codon usage bias in the microbial genomes has significance also in the lateral gene transfer events. The phylum Actinobacteria has the highest median overlap number from amongst the 16 different phyla analysed. As the overlap numbers are influenced by the life style and niche of the organism, this can indicate that this phylum may contain more intracellular organisms which acquire the overlapping patterns in the genome by reducing the genome size.

The abundance of the stop codons or the stop codon bias in the same strand overlaps also indicate the role they play in the transcriptional and translational coupling showing that they have a role in the gene expression regulation. Thus, analysis and quantification of the patterns of codon usage has a very significant role in understanding the relevant mechanisms underlying the biased usage of synonymous codons and to gain idea on the molecular basis of evolution and environmental adaptation of the living organisms.

8 References

Agresti A. (2002). Categorical Data Analysis, 2nd edition. Hoboken, NJ.

Alff-Steinberger C. (1969). The genetic code and error transmission. *Proc Natl Acad Sci U S A*, 64: 584-91.

Angellotti MC, Bhuiyan SB, Chen G, Wan XF. (2007). CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res,* 35: W132–136.

Aota S, Gojobori T, Ishibashi F, Maruyama T, Ikemura T. (1988). Codon usage tabulated from the GenBank Genetic Sequence Data. *Nucleic Acids Res,* 16: r315–r402.

Aota S, Ikemura T. (1986). Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res,* 14: 6345-6355.

Atkins JF, Steitz JA, Anderson CW, Model P. (1979). Binding of mammalian ribosomes to MS2 phage RNA reveals an overlapping gene encoding a lysis function. *Cell*, 18: 247–256.

Bains W. (1987). Codon distribution in vertebrate genes may be used to predict gene length. J *Mol Biol,* 197: 379–388.

Barrell BG, Air GM, Hutchison CA 3rd. (1976). Overlapping genes in bacteriophage *ΦX174*. *Nature,* 264: 34–41.

Basak S, Roy S, Ghosh TC. (2007). On the origin of synonymous codon usage divergence between thermophilic and mesophilic prokaryotes. *FEBS Lett*, 581: 5825-30.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. (2010). GenBank. *Nucleic Acids Res,* 38: D46–D51.

Bennetzen JL, Hall BD. (1982). Codon selection in Yeast. *J Biol Chem,* 257: 3026–3031.

Bernardi G, Bernardi G. (1986). Compositional constraints and genome evolution. *J Mol Evol*, 24: 1–11.

Binns N, Masters M. (2002). Expression of the *Escherichia coli* pcnB gene is translationally limited using an inefficient start codon: a second chromosomal example of translation initiated at AUU. *Mol Microbiol,* 44: 1287– 1298.

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* 277: 1453–1462.

Bofkin L, Goldman N. (2007). Variation in evolutionary processes at different codon positions. *Mol Biol Evol,* 24: 513–521.

Boi S, Solda G, Tenchini ML. (2004). Shedding light on the dark side of the genome: overlapping genes in higher eukaryotes. *Current Genomics*, 5: 509-524.

Bosher JM, Labouesse M. (2000). RNA interference: genetic wand and genetic watchdog. *Nat Cell Biol,* 2**:** E31–E36.

Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, Rabinowitz JD. (2006). Conservation of the metabolomic response to starvation across two divergent microbes. *Proc Natl Acad Sci U S A,* 103: 19302–19307.

Burge CB, Karlin S. (1998). Finding the genes in genomic DNA. *Curr Opin Struct Biol*, 8: 346–354.

Chang LJ, Pryciak P, Ganem D, Varmus HE. (1989). Biosynthesis of the reverse transcriptase of hepatitis B viruses involved *de novo* translational initiation not ribosomal frameshifting. *Nature,* 337: 364–368.

Chamary JV, Hurst LD. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol*, 6: R75.

Chen SM, Takiff HE, Barber AM, Dubois GC, Bardwell JC, Court DL. (1990). Expression and characterization of RNase III and Era proteins. Products of the rnc operon of *Escherichia coli*. *J Biol Chem,* 265: 2888–2895.

Cheng CH, Yang CH, Chiu HT, Lu CL. (2010). Reconstructing genome trees of prokaryotes using overlapping genes. *BMC Bioinform*, 11: 102.

Clark MA, Baumann L, Thao ML, Moran NA, Baumann P. (2001). Degenerative minimalism in the genome of a Psyllid endosymbiont. *J Bacteriol,* 183: 1853–1861.

Cock PJ, Whitworth DE. (2007). Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. *J Mol Evol*, 64: 457-462.

Cock PJ, Whitworth DE. (2010). Evolution of relative reading frame bias in unidirectional prokaryotic gene overlaps. *Mol Biol Evol*, 27: 753-756.

Cooper PR, Smilinich NJ, Day CD, Nowak NJ, Reid LH, Pearsall RS, Reece M, Prawitt D, Landers J, Housman DE, Winterpacht A, Zabel BU, Pelletier J, Weissman BE, Shows TB, Higgins MJ. (1998). Divergently transcribed overlapping genes expressed in liver and kidney and located in the 11p15.5 imprinted domain. *Genomics,* 49: 38–51.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. (1961). General nature of the genetic code for proteins. *Nature*, 192: 1227–1232.

Dandekar T, Snel B, Huynen M, Bork P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci,* 23: 324-328.

Delihas N, Rokita SE, Zheng P. (1997). Natural antisense RNA/target RNA interactions: possible models for antisense oligonucleotide drug design. *Nat Biotechnol,* 15: 751–753.

Devos D, Valencia A. (2001). Intrinsic errors in genome annotation. *Trends Genet*, 17: 429-31.

Dillon LS. (1987). The Gene: Its Structure, Function and Evolution. *Plenum Press New York*. p. 12.

D'Onofrio G, Ghosh TC, Bernardi G. (2002). The base composition of the genes is correlated with the secondary structures of the encoded proteins. *Gene,* 300: 179–187.

Dunnill P. (1966). Triplet nucleotide-amino-acid pairing; a stereochemical basis for the division between protein and non-protein amino-acids. *Nature*, 210: 1265-7.

Duret L. (2002). Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*, 12: 640-9.

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A,* 95: 14863–14868.

Fahey ME, Moore TF, Higgins DG. (2002). Overlapping antisense transcription in the human genome. *Comp Funct Genomics,* 3: 244-53.

Fukuda Y, Nakayama Y, Tomita M. (2003). On dynamics of overlapping genes in bacterial genomes. *Gene*, 323: 181-187.

Fukuda Y, Washio T, Tomita M. (1999). Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res,* 27: 1847– 1853.

Goetz RM, Fuglsang A. (2005). Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. *Biochem Biophys Res Commun*, 327: 4-7.

Golderer G, Dlaska M, Grobner P, Piendl W. (1995). TTG serves as an initiation codon for the ribosomal protein MvaS7 from the archaeon *Methanococcus vannielii*. *J Bacteriol,* 177: 5994–5996.

Gouy M, Gautier C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res,* 10: 7055-7074.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res,* 9: r43–r74.

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res,* 8: r49–r62.

Greenacre M. J. (1984). Theory and Applications of Correspondence Analysis. *Academic Press*, London, UK.

Grocock RJ, Sharp PM. (2002). Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene,* 289: 131–139.

Grundström T, Jaurin B. (1982). Overlap between *ampC* and *frd* operons on the *Escherichia coli* chromosome. *Proc Natl Acad Sci USA*, 79: 111-1115.

Gu W, Zhou T, Ma J, Sun X, Lu Z. (2004). The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Biosystems,* 73: 89–97.

Guo FB, Yuan JB. (2009). Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in Lawsonia intracellularis. *DNA Res*, 2: 91-104.

Gupta SK, Ghosh TC. (2001). Gene expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa. *Gene*, 273: 63-70.

Gupta SK, Bhattacharyya TK, Ghosh TC. (2004). Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *J Biomol Struct Dyn*, 21: 527-36.

Han da X, Wang HY, Ji ZL, Hu AF, Zhao YF. (2010). Amino acid homochirality may be linked to the origin of phosphate-based life. *J Mol Evol*, 70: 572-82.

Hassan S, Mahalingam V, Kumar V. (2009). Synonymous codon usage analysis of thirty two mycobacteriophage genomes. *Adv Bioinformatics*, 2009: 316936.

Hatfield D, Diamond A. (1993). UGA: a split personality in the universal genetic code. *Trends Genet,* 9: 69– 70.

Henikoff S, Keene MA, Fechtel K, Fristrom JW. (1986*).* Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite DNA strands*. Cell,* 44: 33–42.

Higgs PG. (2009). A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct,* 4:16.

Hu J, Zhao X, Zhang Z, Yu J. (2007). Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res Microbiol*, 158: 363-70.

Ikemura T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E.coli* translational system. *J Mol Biol*, 151: 389–409.

Ikemura T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol,* 2: 13–34.

Ikemura T, Ozeki H. (1983). Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb Symp Quant Biol*, 47: 1087-97.

Inokuchi Y, Hirashima A, Sekine Y, Janosi L, Kaji A. (2000). Role of ribosome recycling factor (RRF) in translational coupling. *EMBO J,* 19: 3788– 3798.

Itoh T, Tomizawa J. (1980). Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Pro Natl Acad Sc USA,* 77: 2450-2454.

Jiang LW, Lin KL, Lu CL. (2008). OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes. *Nucleic Acids Res,* 36: W475–480.

Johnson ZI, Chisholm SW. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome Res,* 14: 2268–2272.

Jolliffe I. (2002). Principal component analysis. *Springer,* 2nd edition.

Kanaya S, Kudo Y, Nakamura Y, Ikemura T. (1996). Detection of genes in *Escherichia coli* sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage. *Comput Appl Biosci,* 12: 213–225.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238: 143-55.

Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E coli* O157 genome. *Gene,* 276: 89–99.

Karlin S, Mrazek J. (1996). What drives codon choices in human genes? *J Mol Biol*, 262: 459–472.

Keese PK, Gibbs A. (1992). Origins of genes: ‘‘big bang’’ or continuous creation? *Proc Natl Acad Sci U S A,* 89: 9489–9493.

Kim DS, Cho CY, Huh JW, Kim HS, Cho HG. (2009). EVOG: a database for evolutionary analysis of overlapping genes. *Nucleic Acids Res,* 37: D698-702.

Kingsford C, Delcher AL, Salzberg SL. (2007). A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. *Mol Biol Evol,* 24: 2091-8.

Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y; RIKEN GER Group; GSL Members. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res,* 13: 1324-1334.

Knight RD, Freeland SJ, Landweber LF. (2001). Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet,* 2: 49–58.

Knight RD, Freeland SJ, Landweber LF. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol,* 2: RESEARCH0010.

Kohonen T. (1982). Self-organized formation of topologically correct feature maps. *Biol Cybern*, 43: 59–69.

Kohonen T. (1990). The self-organizing map. *Proc IEEE,* 78: 1464–1480.

Kohonen T, Oja E, Simula O, Visa A, Kangas J. (1996). Engineering applications of the self-organizing map. *Proc IEEE,* 84: 1358–1384.

Koonin EV, Novozhilov AS. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, 61: 99-111.

Korbel JO, Jensen LJ, von Mering C, Bork P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol,* 22: 911-917.

Kozak M. (1983). Comparison of initiation of protein synthesis in prokaryotes, eukaryotes, and organelles. *Microbiol Rev,* 47: 1–45.

Krakauer DC. (2000). Stability and evolution of overlapping genes. *Evolution*, 54: 731–739.

Krakauer DC, Plotkin JB. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci USA,* 99: 1405–1409.

Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, Miller P, Gerstein MB, Snyder M. (2002). An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol,* 20: 58–63.

Kumar M, Carmichael GG. (1998). Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol Mol Biol Rev,* 62: 1415-1434.

Lacatena RM, Cesareni G. (1981). Base pairing of RNA I with its complementary sequence in the primer precursor inhibits ColE1 replication. *Nature,* 294: 623-626.

Lapidot M, Pilpel Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep,* 7: 1216-22.

Levin DB, Whittome B. (2000). Codon usage in nucleopolyhedroviruses. *J Gen Virol,* 81: 2313-2325.

Lillo F, Krakauer DC. (2007). A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. *Biol Direct*, 2: 22.

Lloyd AT, Sharp PM. (1992). Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 20: 5289–5295.

Lobry JR, Gautier C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res,* 22: 3174–3180.

Lobry JR, Necsulea A. (2006). Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene*, 385: 128-36.

Luo Y, Fu C, Zhang DY, Lin K. (2006). Overlapping genes as rare genomic markers: the phylogeny of gamma-Proteobacteria as a case study. *Trends Genet,* 22: 593-596.

Luo Y, Fu C, Zhang DY, Lin K. (2007). BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. *BMC Bioinformatics*, 8: 266.

Lü H, Zhao WM, Zheng Y, Wang H, Qi M, Yu XP. (2005). Analysis of synonymous codon usage bias in Chlamydia. *Acta Biochim Biophys Sin*, 1: 1-10.

Ma J, Campbell A, Karlin S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol,* 184: 5733–5745.

McCutcheon JP, McDonald BR, Moran NA. (2009). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet,* 5: e1000565.

McGirr KM, Buehuring GC. (2006). Tax & rex: overlapping genes of the Deltaretrovirus group. *Virus Genes,* 32: 229–239.

McInerney JO. (1998). **Replicational and transcriptional selection on codon usage in** *Borrelia burgdorferi***.** *Proc Natl Acad Sci USA,* **95:** 10698-1703.

McLachlan AD, Staden R, Boswell DR. (1984). A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res,* 12: 9567–9575.

Mitreval M, Wendl MC, Martin J, Wylie T, Yin Y, Larson A, Parkinson J, Waterston RH, McCarter JP. (2006). Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biology,* 7: R75.

Miyata T, Yasunaga T. (1978). Evolution of overlapping genes. *Nature*, 272: 532–535.

Morel Y, Bristow J, Gitelman SE, Miller WL. (1989). Transcript encoded on the opposite strand of the human steroid 21- hydroxylase/complement component C4 gene locus. *Proc Natl Acad Sci USA,* 86: 6582–6586.

Moreno-Hagelsieb G, Collado-Vides J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics, 18: S329-36.

Morrison, D. F. (1976). Multivariate Statistical Methods. New York: McGraw-Hill.

Nakamura Y, Gojobori T, Ikemura T. (2000). Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucl Acids Res,* 28: 292.

Nakamura Y, Gojobori T, Ikemura T. (1997). Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res, 25*: *244-245*.

Normark S, Bergström S, Edlund T, Grundström T, Jaurin B, Lindberg FP, Olsson O. (1983). Overlapping genes. *Annu Rev Genet,* 17: 499– 525.

Oresic M, Dehn M, Korenblum D, Shalloway D. (2003). Tracing specific synonymous codon-secondary structure correlations through evolution. *J Mol Evol*, 56: 473-84.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A,* 96: 2896-2901.

Pallejà A, Harrington ED, Bork P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions?. BMC Genomics, 9: 335.

Pallejà A, Reverter T, Garcia-Vallvé S, Romeu A. (2009). PairWise Neighbours database: overlaps and spacers among prokaryote genomes. *BMC Genomics,* 10: 281.

Pavesi A. (2000). Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. *J Mol Evol,* 50: 284–295.

Pavesi A. (2006). Origin and evolution of overlapping genes in the family Microviridae. *J Gen Virol,* 87: 1013–1017.

Pavesi A, De Iaco B, Granero MI, Porati A. (1997). On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *J Mol Evol,* 44: 625–631.

Polard P, Prère MF, Chandler M, Fayet O. (1991). Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *J Mol Biol,* 222: 465– 477.

Ponnala L. (2010). A plausible role for the presence of internal shine-dalgarno sites. *Bioinform Biol Insights*, 4: 55-60.

Portin P. (1993). The concept of the gene: short history and present status. Q *Rev Biol,* 68: 173-223.

Portin P. (2009). The elusive concept of the gene. *Hereditas,* 146: 112–117.

Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168: 2245-60.

Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation. *J Virol,* 83: 10719–10736.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www. R-project.org.

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV. (2002). Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet*, 18: 228–232.

Roymondal U, Das S, Sahoo S. (2009). Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. *DNA Res*, 16: 13-30.

Sakharkar KR, Chow VT. (2005). Strategies for genome reduction in microbial genomes. *Genome Inform*, 16: 69-75.

Sakharkar KR, Sakharkar MK, Verma C, Chow VT. (2005). Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int J Syst Evol Microbiol,* 55: 1205–1209.

Sammet SG, Bastolla U, Porto M. (2010). Comparison of translation loads for standard and alternative genetic codes. *BMC Evol Biol*, 10: 178.

Samuel CE. (1989). Polycistronic animal virus mRNAs. *Prog Nucleic Acid Res Mol Biol,* 37: 127–153.

Sander C, Schulz GE. (1979). Degeneracy of the information contained in amino acid sequences: evidence from overlaid genes. *J Mol Evol,* 13: 245–252.

Sanna CR, Li WH, Zhang L. (2008). Overlapping genes in the human and mouse genomes. *BMC Genomics,* 9: 169.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 38: D5–D16.

Sazuka T, Ohara O. (1996). Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp.strain PCC6803 by amino-terminal protein sequencing. *DNA Res,* 3: 225– 232.

Sueoka N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A*, 85: 2653-7.

Shah P, Gilchrist MA. (2010). Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. *PLoS Genet*, 6: e1001128.

Sharp PM, Cowe E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast,* 7: 657–678.

Sharp PM, Li WH. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. *Nucleic Acids Res,* 14: 7737–7749.

Sharp PM, Li WH. (1987). The Codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res,* 15: 1281–1295.

Sharp PM, Tuohy TMF, Mosurski KR. (1986). Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*, 14: 5125-5143.

Sharp PM, Emery LR, Zeng K. (2010). Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*, 365: 1203-12.

Shaw DC, Walker JE, Northrop FD, Barrell BG, Godson GN, Fiddes JC. (1978). Gene K, a new overlapping gene in *bacteriophage G4*. *Nature,* 272: 510-515.

Shendure J, Church GM. (2002). Computational discovery of sense antisense transcription in the human and mouse genomes. *Genome Biol,* 3: 0044.1-0044.14.

Shields DC, Sharp PM. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res*, 15: 8023–8040.

Solda G, Suyama M, Pelucchi P, Boi S, Guffanti A, Rizzi E, Bork P, Tenchini ML, Ciccarelli FD. (2008). Non-random retention of protein-coding overlapping genes in Metazoa, *BMC Genomics*, 9: 174.

Spencer CA, Gietz RD, Hodgetts RB. (1986). Overlapping transcription units in the dopa decarboxylase region of *Drosophila*. *Nature,* 322: 279–281.

Spiers AJ, Bergquist PL. (1992). Expression and regulation of the RepA protein of the RepFIB replicon from plasmid P307. *J Bacteriol,* 174: 7533– 7541.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12: 1611–1618.

Stoletzki N, Eyre-Walker A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol*, 24: 374-81.

Su WY, Xiong H, Fang JY. (2010). Natural antisense transcripts regulate gene expression in an epigenetic manner. *Biochem Biophys Res Commun,* 396: 177-81.

Supek F, Vlahovicek K. (2004). INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics*, 20: 2329-30.

Terryn N, Rouze P. (2000). The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci,* 5: 394–396.

Thompson LA. (2009). R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis (2002) 2nd edition.

Tuller T, Waldman YY, Kupiec M, Ruppin E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A*, 107: 3645-50.

Vanhée-Brossollet C, Vaquero C. (1998). Do natural antisense transcripts make sense in eukaryotes? *Gene*, 211: 1-9.

Veeramachaneni V, Makałowski W, Galdzicki M, Sood R, Makałowska I. (2004). Mammalian overlapping genes: the comparative perspective. *Genome Res,* 14: 280-286.

von der Malsburg C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik,* 14: 85–100.

Wagner EG, Altuvia S, Romby P. (2002). Antisense RNAs in bacteria and their genetic elements. *Adv Genet,* 46: 361-398.

Wagner EG, Simons RW. (1994). Antisense RNA control in bacteria, phages, and plasmids. *Annu Rev Microbiol,* 48: 713-742.

Wan XF, Xu D, Kleinhofs A, Zhou J. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol Biol,* 4: 19.

Wang G, Nie L, Tan H. (2003). Cloning and characterization of sanO, a gene involved in nikkomycin biosynthesis in *Streptomyces ansochromogenes*. *Lett Appl Microbiol,* 37: 452–457.

Wang HC, Badger J, Kearney P, Li M. (2001). Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol Biol Evol*, 18: 792-800.

Wang X, Zhou D, Qin L, Dai E, Zhang J, Han Y, Guo Z, Song Y, Du Z, Wang J, Wang J, Yang R. (2006). Genomic comparison of *Yersinia pestis* and *Yersinia pseudotuberculosis* by combination of suppression subtractive hybridization and DNA microarray. *Arch Microbiol,* 186: 151-159.

Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. (1997). An information-intensive approach to the molecular pharmacology of cancer. *Science,* 275: 343–349.

Weinstein JN. (2008). Biochemistry. A postgenomic visual icon. *Science,* 319: 1772–1773.

Williams T, Fried M. (1986). A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. *Nature,* 322: 275–279.

Wilquet V, Van de Casteele M. (1999). The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Res Microbiol*, 150: 21-32.

Wong JT. (1975). A co-evolution theory of the genetic code. *Proc Natl Acad Sci U S A*, 72: 1909-12.

Wong JT. (2005). Coevolution theory of the genetic code at age thirty. *Bioessays*, 27: 416-25.

Wright F. (1990). The 'effective number of codons' used in a gene. *Gene,* 87: 23–29.

Yelin R, Dahary D, Sorek R. and 13 other authors. (2003). Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol,* 21: 379–386.

Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. (2002). Computational identification of operons in microbial genomes. *Genome Res*, 12: 1221-30.

9 Appendix

List of the species along with their GenBank identifiers used for the analysis

The species name and the GenBank identifiers from the completed genomes of the microbes were collected and checked for their availability in the codon usage database (Nakamura *et al*., 2000). The table prepared according to the thesis plan which includes all the microbial species that are present both in the NCBI (Sayers *et al*., 2010) ftp complete genomes (prokaryotes) and in codon usage database is given below. The table contains the species name and their respective GenBank identifiers that were used in this analysis.

Number	Species from NCBI ftp	GenBank Identifiers
1	*Escherichia coli* 536	110640213
2	*Escherichia coli* APEC O1	117622295
3	*Escherichia coli* CFT073	26245917
4	*Escherichia coli* O157: H7 EDL933	16445223
5	*Escherichia coli* UTI89	91209055
6	*Acinetobacter baumannii* ATCC 17978	126640115
7	*Acinetobacter* sp. ADP1	50083297
8	*Actinobacillus succinogenes* 130Z	152977688
9	*Actinobacillus pleuropneumoniae* L20	126207488
10	*Aeromonas salmonicida* subsp. *salmonicida* A449	145297124
11	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	117617447
12	*Aeropyrum pernix* K1	118430835
13	*Anaplasma phagocytophilum* HZ	88606690
14	*Anaplasma marginale str. Florida*	222474741
15	*Anaplasma marginale str. St. Maries*	56416370
16	*Aquifex aeolicus* VF5	15282445
17	*Archaeoglobus fulgidus* DSM 4304	11497621
18	*Arthrobacter* sp. FB24	116668568
19	*Arthrobacter aurescens* TC1	119960487
20	*Aster yellows witches'-broom phytoplasma* AYWB	85057280
21	*Azoarcus* sp. BH72	119896292
22	*Bacillus amyloliquefaciens* FZB42	154684518
23	*Bacillus anthracis* str. Ames	30260195
24	*Bacillus anthracis* str. 'Ames Ancestor'	50196905
25	*Bacillus cereus* ATCC 10987	42779081
26	*Bacillus halodurans* C-125	57596592
27	*Bacillus licheniformis* ATCC 14580	163119169
28	*Bacillus subtilis* subsp. subtilis str. 168	255767013

29	*Bacillus thuringiensis* str. Al Hakam	118475778
30	*Alcanivorax borkumensis SK2*	110832861
31	*Alkaliphilus metalliredigens* QYMF	150387853
32	*Anabaena variabilis* ATCC 29413	75906225
33	*Acaryochloris marina* MBIC11017	158333233
34	*Acidiphilium cryptum* JF-5	148259021
35	*Acidothermus cellulolyticus* 11B	117927211
36	*Acidovorax* sp. JS42	121592436
37	*Anaeromyxobacter* sp. Fw109-5	153002879
38	*Anaeromyxobacter dehalogenans* 2CP-C	86156430
39	*Bacillus cereus* E33L	52140164
40	*Bacillus clausii* KSM-K16	56961782
41	*Bacillus thuringiensis* serovar konkukian str. 97-27	49476684
42	*Bacteroides fragilis* NCTC 9343	60679597
43	*Bacteroides fragilis* YCH46	53711291
44	*Bacteroides thetaiotaomicron* VPI-5482	29345410
45	*Bartonella bacilliformis* KC583	121601635
46	*Bartonella henselae* str. Houston-1	49474831
47	*Bartonella quintana* str. Toulouse	49473688
48	*Bdellovibrio bacteriovorus* HD100	42521650
49	*Bifidobacterium adolescentis* ATCC 15703	119025018
50	*Bifidobacterium longum* NCC2705	58036264
51	*Bifidobacterium longum*	189438863
52	*Bordetella avium* 197N	187476514
53	*Bordetella bronchiseptica* RB50	33598993
54	*Bordetella pertussis Tohama I*	33591275
55	*Borrelia afzelii* PKo	111114823
56	*Borrelia burgdorferi* B31	15594346
57	*Borrelia garinii* PBi	51598263
58	*Borrelia hermsii* DAH	187917883
59	*Borrelia recurrentis* A1	203287471
60	*Borrelia turicatae* 91E135	119952806
61	*Bradyrhizobium* sp. BTAi1	148251626
62	*Bradyrhizobium* sp. ORS278	146337175
63	*Bradyrhizobium japonicum* USDA 110	27375111
64	*Caldicellulosiruptor saccharolyticus* DSM 8903	146295085
65	*Campylobacter curvus* 525.92	154173617
66	*Campylobacter fetus* subsp.fetus 82-40	118474057
67	*Campylobacter hominis* ATCC BAA-381	154147866
68	*Campylobacter jejuni* subsp. jejuni NCTC 11168	15791399
69	*Campylobacter jejuni* subsp. jejuni 81-176	121612099
70	*Campylobacter jejuni* subsp. doylei 269.97	153950938
71	*Candidatus Blochmannia pennsylvanicus* str. BPEN	71891793
72	*Candidatus Carsonella ruddii* PV	116334902
73	*Candidatus Methanoregula boonei* 6A8	154149549
74	*Candidatus Phytoplasma mali*	194246403

75	*Candidatus Vesicomyosocius okutanii* HA	148244169
76	*Carboxydothermus hydrogenoformans* Z-2901	78042616
77	*Caulobacter crescentus* CB15	16124256
78	*Chlamydia muridarum* Nigg	29337300
79	*Chlamydia trachomatis* A/HAR-13	76788711
80	*Chlamydophila abortus* S26/3	62184647
81	*Chlamydophila caviae* GPIC	29839769
82	*Chlamydophila felis* Fe/C-56	89897807
83	*Chlamydophila pneumoniae* AR39	58021288
84	*Chlamydophila pneumoniae* CWL029	15617929
85	*Chlamydophila pneumoniae* J138	15835535
86	*Chlamydophila pneumoniae* TW-183	33241335
87	*Chlorobium chlorochromatii* CaD3	78187984
88	*Chlorobium phaeobacteroides* DSM 266	119355857
89	*Chlorobium tepidum* TLS	21672841
90	*Chloroflexus aurantiacus* J-10-fl	163845603
91	*Chromobacterium violaceum* ATCC 12472	34495455
92	*Chromohalobacter salexigens* DSM 3043	92112136
93	*Clavibacter michiganensis* subsp. michiganensis NCPPB 382	148271178
94	*Clavibacter michiganensis* subsp. sepedonicus	170780462
95	*Clostridium acetobutylicum* ATCC 824	15893298
96	*Clostridium beijerinckii* NCIMB 8052	150014892
97	*Clostridium botulinum* A str. ATCC 3502	148378011
98	*Clostridium botulinum* A str. Hall	153934468
99	*Clostridium botulinum* F str. Langeland	153937894
100	*Clostridium cellulolyticum* H10	220927459
101	*Clostridium kluyveri* DSM 555	153952670
102	*Clostridium novyi* NT	118442852
103	*Clostridium perfringens* str. 13	18308982
104	*Clostridium perfringens* ATCC 13124	110798562
105	*Clostridium perfringens* SM101	110801439
106	*Clostridium tetani* E88	28209834
107	*Clostridium thermocellum* ATCC 27405	125972525
108	*Colwellia psychrerythraea* 34H	71277742
109	*Corynebacterium diphtheriae* NCTC 13129	38232642
110	*Corynebacterium efficiens* YS-314	25026556
111	*Corynebacterium glutamicum* ATCC 13032	62388892
112	*Corynebacterium glutamicum* ATCC 13032	58036263
113	*Corynebacterium glutamicum* R	145294042
114	*Corynebacterium jeikeium* K411	68535062
115	*Cupriavidus taiwanensis*	188590795
116	*Synechococcus* sp. JA-3-3Ab	86604733
117	*Dehalococcoides* sp. BAV1	147668652
118	*Dehalococcoides* sp. CBDB1	73747956
119	*Dehalococcoides ethenogenes* 195	57233530

120	*Deinococcus geothermalis* DSM 11300	94984109
121	*Desulfitobacterium hafniense* DCB-2	219666071
122	*Desulfitobacterium hafniense* Y51	89892746
123	*Desulfotalea psychrophila* LSv54	51243852
124	*Desulfotomaculum reducens* MI-1	134297881
125	*Dichelobacter nodosus* VCS1703A	146328629
126	*Ehrlichia canis* str. Jake	73666633
127	*Ehrlichia ruminantium* str. Gardel	58616727
128	*Ehrlichia ruminantium* str. Welgevonden	57238731
129	*Enterobacter* sp. 638	146309667
130	*Enterococcus faecalis* V583	29374661
131	*Escherichia coli* O157: H7 str. Sakai	15829254
132	*Fervidobacterium nodosum* Rt17-B1	154248705
133	*Finegoldia magna* ATCC 29328	169823697
134	*Flavobacterium johnsoniae* UW101	146297766
135	*Flavobacterium psychrophilum* JIP02/86	150024114
136	*Francisella tularensis* subsp.*tularensis* FSC198	110669657
137	*Francisella tularensis* subsp.*tularensis* WY96-3418	134301169
138	*Francisella tularensis* subsp.*holarctica*	89255449
139	*Francisella tularensis* subsp.*holarctica* OSU18	115313981
140	*Francisella tularensis* subsp.*tularensis* SCHU S4	255961454
141	*Frankia* sp. CcI3	86738724
142	*Frankia alni* ACN14a	111219505
143	*Geobacillus kaustophilus* HTA426	56418535
144	*Geobacillus thermodenitrificans* NG80-2	138893679
145	*Geobacter metallireducens* GS-15	78221228
146	*Geobacter sulfurreducens* PCA	39995111
147	*Gloeobacter violaceus* PCC 7421	37519569
148	*Gramella forsetii* KT0803	120434372
149	*Granulibacter bethesdensis* CGDNIH1	114326664
150	*Haemophilus ducreyi* 35000HP	33151282
151	*Haemophilus influenzae* Rd KW20	16271976
152	*Haemophilus influenzae* 86-028NP	162960935
153	*Haemophilus influenzae* PittEE	148825133
154	*Hahella chejuensis* KCTC 23*96*	83642913
155	*Halobacterium* sp. NRC-1	15789340
156	*Haloquadratum walsbyi* DSM 16790	110666976
157	*Halorhodospira halophila* SL1	121996810
158	*Helicobacter acinonychis* str. Sheeba	109946640
159	*Helicobacter hepaticus* ATCC 51*449*	32265499
160	*Helicobacter pylori* 26695	15644634
161	*Helicobacter pylori* HPAG1	108562424
162	*Helicobacter pylori* J99	15611071
163	*Herminiimonas arsenicoxydans*	134093294
164	*Hyperthermus butylicus* DSM 5456	124026906
165	*Hyphomonas neptunium* ATCC 15444	114797051

166	*Idiomarina loihiensis L2TR*	56459112
167	*Jannaschia* sp. CCS1	89052491
168	*Kineococcus radiotolerans* SRS30216	255961475
169	*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578	152968582
170	*Lactobacillus acidophilus* NCFM	159162017
171	*Lactobacillus brevis* ATCC 367	116332681
172	*Lactobacillus casei* ATCC 334	116493574
173	*Lactobacillus delbrueckii* subsp. bulgaricus ATCC 11842	104773257
174	*Lactobacillus delbrueckii* subsp. bulgaricus ATCC BAA-365	116513228
175	*Lactobacillus gasseri* ATCC 33323	116628683
176	*Lactobacillus helveticus* DPC 4571	161506634
177	*Lactobacillus plantarum* WCFS1	28376974
178	*Lactobacillus sakei* subsp. sakei 23K	81427616
179	*Lactococcus lactis* subsp. lactis Il1403	15671982
180	*Lactococcus lactis* subsp. cremoris MG1363	125622882
181	*Lactococcus lactis* subsp. cremoris SK11	116510843
182	*Lawsonia intracellularis* PHE/MN1-00	94986445
183	*Legionella pneumophila* str. Lens	54292964
184	*Legionella pneumophila* str. Paris	54295983
185	*Legionella pneumophila* subsp. pneumophila str. Philadelphia 1	52840256
186	*Leifsonia xyli* subsp. xyli str. CTCB07	50953925
187	*Leuconostoc mesenteroides* subsp. mesenteroides ATCC 8293	116617174
188	*Listeria monocytogenes* str. 4b F2365	85700163
189	*Listeria welshimeri serovar* 6b str. SLCC5334	116871422
190	*Lysinibacillus sphaericus* C3-41	169825618
191	*Magnetococcus* sp. MC-1	117923318
192	*Magnetospirillum magneticum* AMB-1	83309099
193	*Mannheimia succiniciproducens* MBEL55E	52424055
194	*Maricaulis maris* MCS10	114568554
195	*Marinobacter aquaeolei* VT8	120552944
196	*Marinomonas* sp. MWYL1	152994043
197	*Mesoplasma florum* L1	50364815
198	*Mesorhizobium loti* MAFF303099	57165207
199	*Metallosphaera sedula* DSM 5348	146302785
200	*Methanothermobacter thermautotrophicus* str. Delta H	15678031
201	*Methanobrevibacter smithii* ATCC 35061	148642060
202	*Methanococcoides burtonii* DSM 6242	91772082
203	*Methanococcus aeolicus Nankai*-3	150400439
204	*Methanocaldococcus jannaschii* DSM 2661	15668172
205	*Methanococcus maripaludis* C5	134045046
206	*Methanococcus maripaludis* C7	150401930
207	*Methanococcus maripaludis* S2	45357563
208	*Methanococcus vannielii* SB	150398760
209	*Methanocorpusculum labreanum* Z	124484829
210	*Methanoculleus marisnigri* JR1	126177952
211	*Methanopyrus kandleri* AV19	20093440
212	*Methanosaeta thermophila* PT	116753325

213	*Methanosarcina acetivorans* C2A	20088899
214	*Methanosarcina mazei* Go1	21226102
215	*Methanosphaera stadtmanae* DSM 3091	84488831
216	*Methylibium petroleiphilum* PM1	124265193
217	*Methylobacillus flagellatus* KT	91774356
218	*Mycobacterium* sp. JLS	126432613
219	*Mycobacterium* sp. KMS	119866057
220	*Mycobacterium* sp. MCS	108796981
221	*Mycobacterium avium* 104	118462219
222	*Mycobacterium avium* subsp.paratuberculosis K-10	41406098
223	*Mycobacterium bovis* AF2122/97	31791177
224	*Mycobacterium bovis* BCG str. Pasteur 1173P2	121635883
225	*Mycobacterium smegmatis* str. MC2 155	118467340
226	*Mycobacterium tuberculosis* CDC1551	50953765
227	*Mycobacterium tuberculosis* F11	148821191
228	*Mycobacterium tuberculosis* H37Ra	148659757
229	*Mycobacterium tuberculosis* H37R*v*	57116681
230	*Mycobacterium ulcerans* Agy99	118615919
231	*Mycobacterium vanbaalenii* PYR-1	120401028
232	*Mycoplasma hyopneumoniae* 232	54019969
233	*Mycoplasma hyopneumoniae* 7448	72080342
234	*Mycoplasma hyopneumoniae* J	71893359
235	*Mycoplasma mobile* 163K	47458835
236	*Mycoplasma mycoides* subsp.mycoides SC str. PG1	127763381
237	*Mycoplasma synoviae* 53	71894025
238	*Myxococcus xanthus* DK 1622	108756767
239	*Nanoarchaeum equitans* Kin4-M	38349555
240	*Natronomonas pharaonis* DSM 2160	76800655
241	*Neisseria gonorrhoeae* FA 1090	59800473
242	*Neisseria meningitidis* MC58	77358697
243	*Neisseria meningitidis* Z2491	15793034
244	*Neorickettsia sennetsu* str. Miyayama	88607955
245	*Nitratiruptor* sp. SB155-2	152989753
246	*Nitrobacter hamburgensis* X14	92115633
247	*Nitrobacter winogradskyi* Nb-255	75674199
248	*Nitrosococcus oceani* ATCC 19707	77163561
249	*Nocardia farcinica* IFM 10152	54021964
250	*Nocardioides* sp. JS614	119714272
251	*Nostoc punctiforme* PCC 73102	186680550
252	*Nostoc* sp. PCC 7120	17227497
253	*Novosphingobium aromaticivorans* DSM 12444	87198026
254	*Oceanobacillus iheyensis* HTE831	23097455
255	*Oenococcus oeni* PSU-1	116490126
256	*Onion yellows phytoplasma* OY-M	255961248
257	*Candidatus Protochlamydia amoebophila* UWE25	46445634
258	*Pasteurella multocida* subsp.multocida str. Pm70	15601865

259	*Pediococcus pentosaceus* ATCC 25745	116491818
260	*Pelobacter carbinolicus* DSM 2380	90960985
261	*Pelobacter propionicus* DSM 2379	118578449
262	*Pelotomaculum thermopropionicum* SI	147676335
263	*Picrophilus torridus* DSM 9790	48477072
264	*Polaromonas* sp. JS666	91785913
265	*Polaromonas naphthalenivorans* CJ2	121602919
266	*Porphyromonas gingivalis* ATCC 33277	188993864
267	*Porphyromonas gingivalis* W83	34539880
268	*Prochlorococcus marinus* str. AS9601	123967536
269	*Prochlorococcus marinus* subsp.marinus str. CCMP1375	33239452
270	*Prochlorococcus marinus* subsp.pastoris str. CCMP1986	33860560
271	*Prochlorococcus marinus* str. MIT 9313	33862273
272	*Prochlorococcus marinus* str. MIT 9211	159902540
273	*Prochlorococcus marinus* str. MIT 9301	126695337
274	*Prochlorococcus marinus* str. MIT 9303	124021714
275	*Prochlorococcus marinus* str. MIT 9312	78778385
276	*Prochlorococcus marinus* str. MIT 9515	123965234
277	*Prochlorococcus marinus* str. NATL1A	124024712
278	*Prochlorococcus marinus* str. NATL2A	162958048
279	*Propionibacterium acnes* KPA171202	50841496
280	*Pseudomonas aeruginosa* PAO1	110645304
281	*Pseudomonas aeruginosa* PA7	152983466
282	*Pseudomonas aeruginosa* UCBPP-PA14	116048575
283	*Pseudomonas fluorescens* Pf-5	70728250
284	*Pseudomonas fluorescens* SBW25	229587578
285	*Pseudomonas mendocina* ymp	146305042
286	*Pseudomonas putida* F1	148545259
287	*Pseudomonas putida* GB-1	167031021
288	*Pseudomonas stutzeri* A150*1*	146280397
289	*Pseudomonas syringae* pv.phaseolicola 1448A	71733195
290	*Psychrobacter* sp. PRwf-1	148651817
291	*Psychrobacter arcticus* 273-4	71064581
292	*Psychrobacter cryohalolentis* K5	93004831
293	*Psychromonas ingrahamii* 37	119943794
294	*Pyrobaculum aerophilum str.* IM2	18311643
295	*Pyrobaculum arsenaticum* DSM 13514	145590267
296	*Pyrobaculum calidifontis* JCM 11548	126458628
297	*Pyrobaculum islandicum* DSM 4184	119871520
298	*Pyrococcus furiosus* DSM 3638	18976372
299	*Rhizobium* sp. NGR234	227820587
300	*Rhizobium etli* CFN 42	86355669
301	*Rhodobacter sphaeroides* ATCC 17025	146276058
302	*Rhodoferax ferrireducens* T118	89898822
303	*Rhodopseudomonas palustris* BisA53	115522030
304	*Rhodopseudomonas palustris* BisB18	90421528

305	*Rhodopseudomonas palustris* BisB5	91974482
306	*Rhodopseudomonas palustris* HaA2	86747127
307	*Rhodopseudomonas palustris* TIE-1	192288433
308	*Rhodospirillum rubrum* ATCC 11170	83591340
309	*Rickettsia bellii* RML369-C	91204815
310	*Rickettsia conorii* str. Malish 7	15891923
311	*Rickettsia felis* URRWXCal2	67458392
312	*Rickettsia typhi* str. Wilmington	51473215
313	*Roseiflexus* sp. RS-1	148654188
314	*Roseobacter denitrificans* OCh 114	110677421
315	*Rubrobacter xylanophilus* DSM 9941	108802856
316	*Saccharophagus degradans* 2-40	90019649
317	*Saccharopolyspora erythraea* NRRL 2338	134096620
318	*Salinibacter ruber* DSM 13855	83814055
319	*Salinispora tropica* CNB-440	145592566
320	*Salmonella enterica* subsp.arizonae serovar 62:z4, z23:--	161501984
321	*Salmonella enterica* subsp.enterica serovar Newport str. SL254	194442203
322	*Salmonella enterica* subsp.enterica serovar Typhi str. CT18	16758993
323	*Shewanella* sp. MR-4	113968346
324	*Shewanella* sp. MR-7	114045513
325	*Shewanella* sp. W3-18-1	120596833
326	*Shewanella baltica* OS185	152998555
327	*Shewanella denitrificans* OS217	91791369
328	*Shewanella frigidimarina* NCIMB 400	114561188
329	*Shewanella loihica* PV-4	127510935
330	*Shewanella oneidensis* MR-1	24371600
331	*Shewanella putrefaciens* CN-32	146291111
332	*Shigella boydii* Sb227	82542618
333	*Shigella dysenteriae* Sd197	82775382
334	*Shigella flexneri* 2a str. 301	24111450
335	*Shigella flexneri* 2a str. 2457T	30061571
336	*Shigella flexneri* 5 str. 8401	110804074
337	*Shigella sonnei* Ss046	74310614
338	*Sinorhizobium medicae* WSM419	150395228
339	*Sinorhizobium meliloti* 1021	15963753
340	*Sodalis glossinidius* str. 'morsitans'	85057978
341	*Sphingomonas wittichii* RW1	148552929
342	*Sphingopyxis alaskensis* RB2256	103485498
343	*Staphylococcus aureus* subsp. aureus COL	57650036
344	*Staphylococcus aureus* subsp. aureus JH1	150392480
345	*Staphylococcus aureus* subsp. aureus JH9	148266447
346	*Staphylococcus aureus* subsp. aureus MW2	21281729
347	*Staphylococcus aureus* subsp. aureus Mu50	57634611
348	*Staphylococcus aureus* subsp. aureus N315	29165615
349	*Staphylococcus aureus* subsp. aureus NCTC 8325	88193823
350	*Staphylococcus aureus* subsp. aureus str. Newman	151220212
351	*Staphylococcus aureus* RF122	82749777

352	*Staphylococcus aureus* subsp. aureus MRSA252	49482253
353	*Staphylococcus aureus* subsp. aureus MSSA476	49484912
354	*Staphylococcus epidermidis* ATCC 12228	27466918
355	*Staphylococcus epidermidis* RP62A	57865352
356	*Staphylococcus haemolyticus* JCSC1435	70725001
357	*Staphylococcus saprophyticus* subsp.saprophyticus ATCC 15305	73661309
358	*Staphylothermus marinus* F1	126464913
359	*Streptococcus agalactiae* 2603V/R	22536185
360	*Streptococcus agalactiae* A909	76786714
361	*Streptococcus agalactiae* NEM316	25010075
362	*Streptococcus equi* subsp.zooepidemicus	225867617
363	*Streptococcus mitis* B6	289166909
364	*Streptococcus mutans* UA159	24378532
365	*Streptococcus pneumoniae* D39	116515308
366	*Streptococcus pneumoniae* TIGR4	194172857
367	*Streptococcus pyogenes* M1 GAS	15674250
368	*Streptococcus pyogenes* MGAS10270	94989509
369	*Streptococcus pyogenes* MGAS10394	50913346
370	*Streptococcus pyogenes* MGAS10750	94993396
371	*Streptococcus pyogenes* MGAS2096	94991497
372	*Streptococcus pyogenes* MGAS315	21909536
373	*Streptococcus pyogenes* MGAS9429	94987631
374	*Streptococcus pyogenes* str. Manfredo	139472888
375	*Streptococcus pyogenes* SSI-1	28894912
376	*Streptococcus sanguinis* SK36	125716887
377	*Streptococcus suis* 05ZYH33	146317663
378	*Streptococcus suis* 98HAH33	146319850
379	*Streptococcus suis* P1/7	253752822
380	*Streptococcus thermophilus* CNRZ1066	55821993
381	*Streptococcus thermophilus* LMD-*9*	116626972
382	*Streptococcus thermophilus* LMG 18311	55820103
383	*Sulfolobus acidocaldarius* DSM 639	70605853
384	*Sulfolobus tokodaii* str.7	24473558
385	*Sulfurovum* sp. NBC37-1	152991597
386	*Symbiobacterium thermophilum* IAM 14863	51891138
387	*Synechococcus* sp. CC9311	113952711
388	*Synechococcus* sp. CC9605	78211558
389	*Synechococcus* sp. CC9902	78183584
390	*Synechococcus* sp. PCC 7002	170076636
391	*Synechococcus* sp. RCC307	148241099
392	*Synechococcus* sp. WH 7803	148238336
393	*Synechococcus elongatus* PCC 7942	81298811
394	*Synechococcus* sp. WH 8102	33864539
395	*Syntrophobacter fumaroxidans* MPOB	116747452
396	*Syntrophomonas wolfei* subsp.*wolfei str. Goettingen*	114565576
397	*Syntrophus aciditrophicus* SB	85857845

398	*Thermobifida fusca* YX	72160406
399	*Thermofilum pendens* Hrk 5	119718918
400	*Thermoplasma volcanium* GSS1	13540831
401	*Thermosipho melanesiensis* BI429	150019913
402	*Thermosynechococcus elongatus* BP-1	22297544
403	*Thermotoga* sp. RQ2	170287807
404	*Thermotoga maritima* MSB8	15642775
405	*Thermotoga neapolitana* DSM 4359	222098974
406	*Thermotoga petrophila* RKU-1	148269145
407	*Thermus thermophilus* HB27	46198308
408	*Thiobacillus denitrificans* ATCC 25259	74316018
409	*Thiomicrospira crunogena* XCL-2	118139508
410	*Treponema denticola* ATCC 35405	42516522
411	*Treponema pallidum* subsp.pallidum str. Nichols	15638995
412	*Trichodesmium erythraeum* IMS101	113473942
413	*Tropheryma whipplei* TW08/27	28572175
414	*Tropheryma whipplei* str. Twist	32447382
415	*Ureaplasma parvum serovar 3* str. ATCC 700970	13357558
416	*Verminephrobacter eiseniae* EF01-2	121607004
417	*Wigglesworthia glossinidia* endosymbiont of Glossina brevipalpis	32490749
418	Wolbachia endosymbiont strain TRS of Brugia malayi	58584261
419	Wolbachia endosymbiont of Drosophila melanogaster	42519920
420	*Wolbachia* sp. wRi	225629872
421	*Xanthobacter autotrophicus* Py2	154243958
422	*Xanthomonas albilineans*	285016821
423	*Xanthomonas campestris* pv. campestris str. 8004	66766352
424	*Xanthomonas campestris* pv. campestris str. ATCC 33913	21229478
s425	*Xanthomonas axonopodis* pv. citri str. 306	21240774
426	*Xanthomonas oryzae* pv .oryzae KACC10331	58579623
427	*Xanthomonas oryzae* pv. oryzae MAFF 311018	84621657
428	*Yersinia enterocolitica* subsp.enterocolitica 8081	123440403
429	*Yersinia pestis Antiqua*	108805998
430	*Yersinia pestis* CO92	16120353
431	*Yersinia pestis Nepal*516	108810166
432	*Yersinia pestis* biovar Microtus str.91001	45439865
433	*Yersinia pseudotuberculosis* IP 32953	51594359
434	*Yersinia pseudotuberculosis* IP 31758	153946813
435	*Zymomonas mobilis* subsp.mobilis ZM4	283856168