Chapter 1

Introduction

The genome of a bacterium consists of the bacterium's complete genetic information. Essentially, it determines the phenotypes and genotypes of the bacterium. This genetic information is organized into a collection of genes, which are subsequences inscribed in the two strands of a Deoxyribonucleic acid (DNA). The two strands run in opposite directions with respect to each other (antiparallel). They are held together by hydrogen bonds between complementary bases of W (A and T with two hydrogen bonds) and S (G and C with three hydrogen bonds) nucleotides. In both the DNA strands, information is stored as a code made up of these four nucleotides.

Since whole genome sequencing of the first bacterium *Haemophilus influenza* in 1995, there has been a considerable development in genomic and other molecular research technologies, particularly in the field of genome sequencing projects. Further rapid information technology developments helped produce a huge volume of genomic information in online public databases such as DDBJ, EMBL, and NCBI. As per a recent study in 2020, more than 2,00,000 genomes from different phyla have been sequenced along with a variety of micro-organisms [250]. These genomic data sequenced across diverse micro-organisms have invited experts from computer science, mathematics and biological domains to work together for new insights into genome content, adaptation, and evolution. Cross-platform research has resulted in additional information on RNA and protein structures available in public databases such as GtRNAdb, RNA STRAND, PDB, etc. In this genomic era, our understandings of evolutionary aspects of microbial genomes have increased significantly owing to recent developments in data science and machine learning algorithms that facilitate large-scale analysis of genome sequences to understand evolution at the molecular level.

Different segments of a DNA sequence are associated with different biological functions by virtue of the primary base sequences of the segment or the secondary structure produced out of the base sequence. Nucleotide sequence of a protein-coding genes is read in the form of base triplets called codons, mapped to the genetic code table to produce the amino acid sequence. The base sequence of a tRNA gene facilitates amino acid transportation to the site of protein synthesis because of its typical clover leaf-like secondary structure. Therefore, different segments of the DNA sequence are under influence of different evolutionary forces [136][27][68][69][73][170].

Non uniform usage of synonymous codons in protein coding genes of organisms is known since Clarke (1970) for about five decades. Progress in genome sequencing technologies in 1980s facilitated accumulation of genome sequences in public databases that helped scientists to put forward theories that guides us to understand codon usage [58][59][60][82][83][84][191][190][192][63][57][13][136]. Grantham and co-researchers proposed the genome hypothesis in 1980 that states all genes in a genome, or more loosely genome type, tend to have the same coding strategy. The genome hypothesis mainly indicates the presence of some selection mechanism in a genome operating for the differential usage of synonymous codons and the selection mechanism is variable across organisms. In support of selection on codon usage in 1982 [57] reported codon usage difference between high and low expression genes in *Escherichia coli*. Ikemura [82][83][84] demonstrated the positive correlation between codon abundance and the cytosolic abundance of cognate tRNA in three organisms *Escherichia coli*, *Salmonella typhimurium*, and *Saccharomyces cerevisiae*, suggesting selection for efficient translation. Muto and Osawa (1987) [136] demonstrated that genome G+C% in bacteria strongly influences codon usage bias such that in G+C% high genomes, synonymous codons ending with G/C are preferred to synonymous codons ending with A/T whereas the reverse is true in case of G+C% low genomes. Using population genetics based methodology, Sharp et al. in 2005 reported that the selection strength is variable among bacteria [193] and finding from Rocha in 2004 [168] pointed out that the growth rate in organisms imposes selection on codon usage in organisms. Further lifestyle of the organisms is also reported to be a selection factor on codon usage in bacteria [18]. The machine learning based study [210] reported that that the translational selection on codon is ubiquitous in bacteria. Our understanding on mechanism of selection on codon usage was further increased by the work of Ran and Higgs [160]. These scientists showed that the selection rules for two-fold degenerate codons and four-fold degenerate codons are different. Wald et al. in 2012 also observed the same difference between two-fold degenerate codons and four-fold degenerate codons with respect to selection [227]. Bulmer in 1990 proposed the Selection-Mutation-Drift (SMD) theory according to which selection predominates in determining codon usage bias in high expression genes whereas nucleotide substitution predominates in determining codon usage bias in low expression genes [20].More recently, codon usage bias has been demonstrated to influence co-translational protein folding [10]. Ribosome profile has distinctly demonstrated the difference among synonymous codons regarding the translation speed [162]. Genome composition (G+C%) that varies approximately from 13.0 to 75.0 in bacteria is the other major factor influencing codon usage [143][170]. Single nucleotide base substitution is one of the leading factor influencing genome composition in bacteria and therefore, the codon usage in organisms is influenced by combined effect of nucleotide substitution and selection forces.

The stem-loop hypothesis or the Nussinov-Forsdyke hypothesis proposed genome wide selection for formation of DNA secondary structure (DNA stem-loop regions) which is advantageous to the cell [113]. A sequence containing an inverted repeat (e.g. NNNATGCNNNGCATNNN) has palindrome-like characteristics with the potential to fold back on itself forming a hairpin like stem-loop structures having significant functions, for example tRNA clover leaf structure and rho-independent transcription termination site in prokaryotes. Organisms those have accepted point mutations which increased the probability of stem-loop formation reported to have evolutionary advantage [45].

In this thesis, considering large-scale bacterial genome sequences available from public databases, we have addressed some of the important aspects of molecular evolution, such as codon usage bias, nucleotide polymorphism, strand asymmetry, RNA secondary structure in bacterial genomes.

1.1 Codon usage bias in bacterial genomes

In the universal genetic code table (Table 1.1), 64 base triplets called codons are assigned to 20 amino acids and translation stop signals. The number of codons are more than the number of amino acids and therefore, codon to amino acid mapping is many to one. But the number of codons assigned to the amino acids are not uniform: two amino acids Met and Trp are encoded by one codon each; nine amino acids Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Tyr are encoded by two codons each, Ile is encoded by three codons; five amino acids Val, Pro, Thr, Ala and Gly are encoded by four codons each; and the three amino acids Leu, Ser and Arg are encoded by six codons each. The three remaining codons represents the translational stop signals. Codons that encode the same amino acid are called as synonymous codons, which are known to be used non-randomly in organisms, a phenomenon known as codon usage bias. The pattern and magnitude of codon usage bias is different from organisms to organisms and even among the genes within a genome. Base composition and replication strand asymmetry are considered some of the mutational factors [47][168] influencing codon usage bias in genomes

AA	Codon	AA	Codon	AA	Codon	AA	Codon
Phe	UUU	Ser	UCU	Tyr	UAU	Cys	UGU
	UUC		UCC		UAC		UGC
Leu	UUA		UCA	TER	UAA	TER	UGA
	UUG		UCG		UAG	Trp	UGG
	CUU	Pro	CCU	His	CAU	Arg	CGU
	CUC		CCC		CAC		CGC
	CUA		CCA	Gln	CAA		CGA
	CUG		CCG		CAG		CGG
Ile	AUU	Thr	ACU	Asn	AAU	Ser	AGU
	AUC		ACC		AAC		AGC
	AUA		ACA	Lys	AAA	Arg	AGA
Met	AUG		ACG		AAG		AGG
Val	GUU	Ala	GCU	Asp	GAU	Gly	GGU
	GUC		GCC		GAC		GGC
	GUA		GCA	Glu	GAA		GGA
	GUG		GCG		GAG		GGG

Table 1.1: Genetic code table

of bacteria. Optimum translational selection for either accurate or fast codon recognition is a selection factor influencing CUB among the genes in a genome [160][194]. Codons used more frequently in high expression genes than the low expression genes are considered the optimal codons (OCs) [68][192][195].

In last four decades, different computational methods have been proposed to measure codon usage bias in a gene sequence. Relative synonymous codon usage (RSCU) estimates biasness in synonymous codon usage for individual amino acids in a gene sequence. Effective number of codons [237] estimate overall biasness in codon usage in a gene. These metrics used in RSCU and Nc assumes equal base frequencies. The null expectation for the usage of each codon depends upon the GC content, meaning that in a GC-rich genome, codons with G and C will tend to be used more than those with A and T without being preferred. Effective number of codons prime (Nc′) and its variants [137][181] estimate biasness in codon usage in a gene taking GC content into account. Codon Adaptation Index (CAI) [192] estimates the extent of biasness in codon usage in a gene towards codons that are known to be used favorably in highly expressed genes. Software implementations of these measures are available for public used in the form of various packages, stand-alone programs and also in the form of websites [145][209][131][181]. Implementations also exist in libraries of common programming languages, such as, BioPerl [201], seqinR [25], and EMBOSS [167]. Here we briefly present mathematical equations of the methods RSCU and CAI considered in this thesis.

1. **Relative Synonymous Codon Usage (RSCU)**

   Relative Synonymous Codon Usage (RSCUi) value for i^{th} Codon is calculated as the observed number of occurrences divided by the number expected if all synonymous codons for an amino acid were used equally frequently. For synonymous Codon i of an n-fold degenerate amino acid, RSCU is as shown in equation 1.1

   $$RSCU_i = X_i/(1/n\sum_{i=1}^{n} X_i) \tag{1.1}$$

   X_i is the number of occurrences of Codon i and n is degeneracy (1, 2, 3, 4 or 6).

   Minimum and maximum possible RSCU values varies between 0.0 and degeneracy of the amino acid codons respectively. For equal usage of synonymous codons, RSCU values are 1.0 each for the synonymous codons. Sum of RSCU values of all synonymous codons for an amino is always equal to the degeneracy of the amino acid codons.

2. **Codon Adaptation Index (CAI)**

   The index uses a set of highly expressed genes which is used as a reference to categorize any unknown gene of same species to be highly or lowly expressed. It is also used for assessing the adaptation of viral genes to their hosts, and for making comparisons of codon usage in different organisms [192][145].

   $$CAI = (\pi_{i=1}^{L} w_i)^{1/L} \tag{1.2}$$

   Here relative adaptedness values (w_i) are calculated for each of the 61 amino acid encoding codons considering a reference set of organism-specific high expression genes. L is codon count in the gene sequence excluding Met and Trp codons. For codon i, w_i is defined in equation 1.3.

$$where \quad w_i = RSCU_i/(max(RSCU)) \tag{1.3}$$

Here $RSCU_i$ and $max(RSCU)$ are the relative synonymous codon usage values for a codon i and the maximum RSCU value among all the synonymous codons of the codon i, respectively.

CAI values of the genes vary between 0.0 to 1.0 denoting high and low expression genes, respectively.

1.2 Single nucleotide polymorphism in bacterial genomes

Biological information is stored in DNA by the linear order of the bases in it. During DNA replication, repair or recombination, in a very low frequency, one base can be replaced by any other three bases, which causes change in the information in the DNA. These changes in DNA sequence due to base replacement are known as substitution. Base replacement between A and G or between C and T is called as a transition (*ti*). Base replacement between R (A, G) and Y (C, T) bases is called transversion (*tv*) Figure 1-1.

Each of the four bases can change to three other bases resulting in total of twelve directional base substitutions. Out of these twelve substitutions, four are transitions, and eight are transversions. Though the theoretically possible number of transversions is two times the number of transitions, the observed transition frequency is two times the transversion frequency [187][39][105][204][123][184]. This four times higher transition frequency compared to transversion is attributed to mechanistic reasons such as unequal stability among different base pairs, dif-

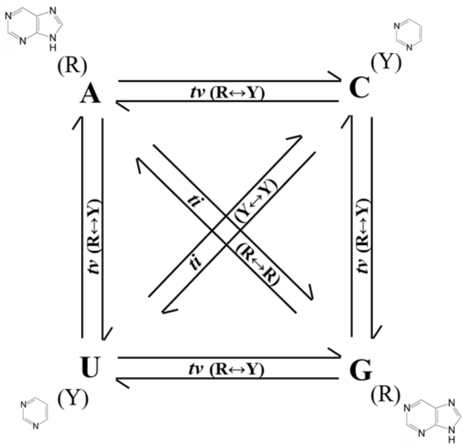

Figure 1-1: Different substitution mutations in the genome
Figure presents twelve possible directional base substitutions in a sequence. In theory, out of the four bases A, C, G, and U any one base can be replaced by the other three bases resulting into twelve base substitutions. Out of these twelve substitutions, the four are called transitions (ti) in which a purine (R: A/G) (or a pyrimidine (Y:C/U)) is replaced by another purine (R) (or a pyrimidine (Y)); eight different substitutions called transversion (tv) in which a purine (R) (or a pyrimidine (Y)) is replaced by a pyrimidine (Y) (or a purine (R))

ferential propensity of bases to damages such as deamination, oxidation, and radiation, as well as selective reasons such as differential impact on the structure and function of DNA, RNA, and proteins [211][114][47][26]. The polymorphism analysis in different genomic regions such as the Intergenic Regions (IRs), Four-fold Degenerate sites (FFS), stem-loop regions in RNA secondary structures etc. helps to understand different aspects of molecular evolution.

1.3 RNA secondary structure estimation

Primary base sequences of single stranded RNAs tend to form pairs with their alternate bases (A:U, G:C and G:U) to form secondary structures consisting of typical stem and loop motifs. Bonds which are formed by base pairing are called internal bonds or edges. Some bases will form pairs while others may not and they form loops. Any bounded region by external or internal edges is called a face.

A hypothetical example containing faces of a secondary structure is shown in Figure 1-2. The figure in left depicts the typical motifs or faces such as hair-pin loop (H: region bounded by the bases in positions 21, 22, 23, 24, and 25), stacking region (S: region bounded by the bases in positions 2, 3, 31, and 32. A stem is constituted by two or more stacking regions as in positions 5, 6, 7, 12, 13, 14), bulge loop (BL: region bounded by the bases in positions 4, 5, 14, 15, and 16), interior loop (I: region bounded by the bases in positions 18, 19, 20, 26, 27, 28, and 29), bifurcation loop (BF: region bounded by the bases in positions 3, 4, 16, 17, 18, 29, 30, and 31). Pseudoknots (PK) are RNA tertiary structures which are formed between unpaired bases of RNA secondary structure. Here in this figure, pseudo-knotted structure is formed when the 10^{th} and 1^{th} bases of the H loop pair with bases at 23^{rd} and 22^{nd} positions, respectively. They are non-nested or overlapping in nature as can be illustrated in the form of an arc diagram given as depicted in the right hand side figure. Pseudo-knot is represented as a dotted stacking region. The corresponding arc diagram in the right panel shows that the motifs are nested/non-nested in nature with respect to the absence / presence of pseudo-knots.

Owing to this secondary structure, RNA performs several important cellular functions for example, the clover leaf structure of the tRNA. In order to supplement resource-intensive scientific experiments, interdisciplinary approaches based on computational algorithms have been developed for estimating RNA secondary structure. Various estimation algorithms based on machine learning principles [103][37][77][23] emphasizing on recent advancements in deep learning methods[225][248][126][23] have been incorporated to estimate secondary structure of RNA.

A huge volume of available DNA sequences in public databases provides scope for the opened direction of structure prediction and genomic analysis. Determining secondary structure of RNAs helps to determine gene expression, regulation, stability, and function [51]. It also helps to understand the genetic disease, create new drugs, and allows biologists understand molecules in cell

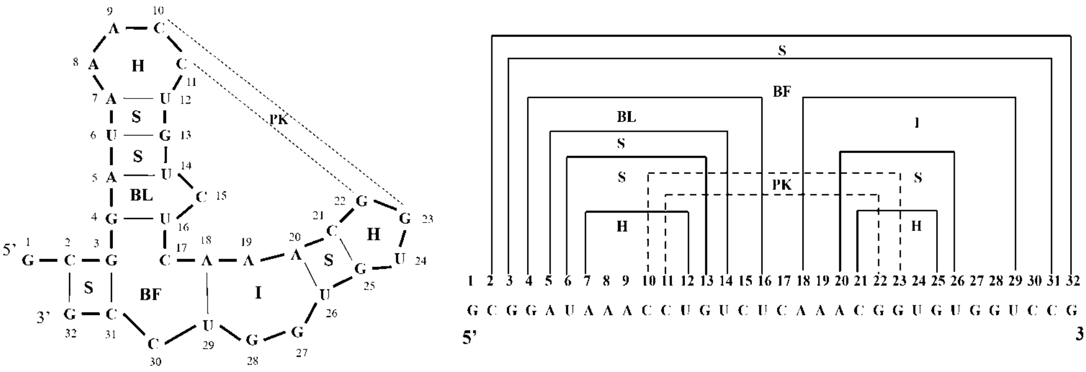

Figure 1-2: Different faces in a RNA secondary structure
A two-panel figure presenting the motifs of secondary structures with a pseudoknot (left panel) and corresponding arc diagram (right panel) of a hypothetical RNA sequence. The consecutive bases in the primary sequence of RNA are linked by external edges. Bonds which are formed by hydrogen bonds between two distant bases (A:U, G:C and G:U) are called internal edges. Figures in left depict the typical motifs or faces such as hair-pin loop (H), stacking region (S).A stem is constituted by two or more stacking regions, bulge loop (BL), interior loop (I),bifurcation loop (BF). Pseudoknots (PK) are RNA tertiary structures which are formed between unpaired bases of RNA secondary structure. They are non-nested or overlapping in nature as can be illustrated in the form of an arc diagram in right. Pseudo-knot is represented as dotted stacking region. The corresponding arc diagrams in the right panel shows that the motifs are nested/non-nested in nature with respect to absence / presence of pseudo-knots.

[216]. There are different types of RNA(s), e.g., transfer RNA (tRNA), messenger RNA(mRNA), ribosomal RNA (rRNA), signal recognition particle RNA (SRP RNA) etc.

1.3.1 Thermodynamic features considered in RNA secondary structure prediction algorithms

To determine the optimal RNA secondary structure among several sub-optimal structures, thermodynamic energies are taken into consideration. Minimum free energy (MFE) is an energy model used to determine the thermodynamic stability of an RNA structure. Nearby bases in RNA sequence pair to reduce thermodynamic free energy and make the secondary structure more stable. Therefore, stability of an RNA secondary structure can be inversely proportional to its associated free energy. The total free energy of a secondary structure is calculated by combining energies associated with individual motifs in the structure. Stems are comparatively stable and therefore free energy associated with stems is less in comparison to loops. Bigger loops make the secondary structure unsteady. Gibbs free energy model is used to calculate free energy of a structure. In this model, overall free energy of a structure is estimated based on the local arrangement of the bases in the motif, so it is termed as nearest neighbour model [139][158]. Gibbs free energy can be described as follows:

$$\Delta G = \Delta H - T\Delta S \tag{1.4}$$

Where ΔG represents Gibbs Energy, ΔH represents enthalpy which is the total energy to create a system, or it is also sum of total internal energies and T is the temperature. ΔS is entropy which measures molecular disorder, or randomness of a system [139][258]. For a Watson-Crick stacking, Gibbs Energy is calculated as:

ΔG°37 Watson-Crick = ΔG°37 intermolecular initiation + ΔG°37 AU end penalty (per AU end) +ΔG°37 symmetry (self-complementary duplexes) + P[ΔG°37 stacking]

These values and the computations are based on Turners lab [220]. For example, for a Watson-Crick stacking:

5‘ CGCAGCU 3’
3‘ GCGUCGA 5’

ΔG°37 Watson-Crick = ΔG°37 intermolecular initiation + ΔG°37 AU end penalty + ΔG°37 (CG followed by GC) + ΔG°37 (GC followed by CG) + ΔG°37 (CG followed by AU) + ΔG°377 (AU followed by GC) + ΔG°37 (GC followed by CG) + ΔG°37 (CG followed by UA)

= 4.9 kcal/mol + 0.45 kcal/mol 2.36 kcal/mol 3.42 kcal/mol 2.11 kcal/mol 2.08 kcal/mol 3.42 kcal/mol 2.08 kcal/mol.

$= 10.12$ kcal/mol

1.4 Challenges

Evolutionary information is accumulated over millions of years in the DNA of organisms in the form of a sequence of four nucleotides denoted by characters A, T, G and C. Size of the DNA sequence varies from few Kilo bases in virus to several Giga bases in human. Different parts of the DNA are associated with different functions of the organism. In the process of translation, the nucleotide sequence of the protein coding gene is utilized to produce a protein which is a sequence of 20 amino acids. The standard version of the genetic code table defines non-random translational mapping of 64 nucleotide triplets (or codons) to 20 amino acids and three stop signals. Though more than 10^{84} alternative genetic code tables are possible theoretically [97], only one standard genetic code table is used universally across bacteria with some variations in alternative start codons and the use of the stop codons UAG and UGA for the rare amino acid selenocysteine and pyrrolysine, respectively. It is to be noted here that there are thirty two other genetic code tables with slight variations reported in National Center for Biotechnology Information (NCBI) web portal (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), for example, the universal codons AUA (isoleucine) and UGA (stop) coded for methionine and tryptophan, respectively, in human mitochondria. Evolution of the genetic code is elegantly presented in a research article by Osawa and co-researchers in 1992 [142]

The large volume of genome sequences available in online public databases such as DDBJ and NCBI have invited experts from computer science, statistics, and other fields to understand biological processes and increase our understanding of the biological system. Computationally intensive techniques are used to recognize patterns and visualization of biological systems. Different parts of the DNA being associated with different biological functions, understanding the significance of the sequence, developing suitable computational methods, generate result and their interpretation are challenging tasks for bioinformaticians.

Experiments such as Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography are often used to determine accurate RNA secondary structure. Because of the secondary structure, RNA performs several important functions, for instance, tRNA folds to form a typical cloverleaf shape and participate in the translation. But, laboratory experiments to determine RNA structures are expensive and time-consuming. However, developing computational approaches to estimate these structures with reasonable accuracy are also challenge.

1.5 Objective of the thesis

The aim of this research is to understand factors such as base substitution, strand asymmetry, codon usage bias, RNA secondary structure influencing evolution in bacteria by computational analysis of genome sequences. The objectives are as follows:-

(A) To analyze codon usage bias (CUB) in bacterial genomes

  (i) To find out the difference in CUB between the high expression genes and all other genes using machine learning algorithms

  (ii) To find out codons preferred/avoided in the high-expression genes

  (iii) To understand the role of codon degeneracy on selected codon usage bias

(B) To analyze single nucleotide polymorphism spectrum in bacterial genomes

  (i) To develop a computational methodology for estimating base substitution patterns in bacterial genomes

  (ii) To segregate bacterial genome regions such as CDS, IRs into leading and lagging strand

   (iii) To compare single nucleotide polymorphism spectra at Four-Fold Degenerate sites (FFS) and Intergenic Regions (IRs)

(C) To estimate RNA secondary structure

   (i) To understand the principle behind RNA secondary structure motifs

   (ii) To develop a computational method to estimate RNA secondary structure motifs

(D) To analyze the role of secondary structure motifs on base substitution

   (i) To estimate secondary structure motifs in tRNA genes sequences and segregate stem and loops regions

   (ii) To estimate base substitution pattern in stem and loop regions in tRNA genes and do a comparative analysis

1.6 Organization of the thesis

The thesis is organized as follows:

- Chapter 1 introduces the thesis and the challenges. It describes the three different sub-objectives that we addressed. It also describes the organization of the thesis.
- Chapter 2 presents the machine learning-based study on codon usage difference between the high and the low expression genes in genomes of bacteria. The chapter also describes the web-server to calculate Codon Adaptation Index with organism-specific reference set of genes.
- Chapter 3 presents single nucleotide polymorphism (SNPs) in intergenic regions (IRs) and four-fold degenerate sites (FFS) in genomes of three γ-Proteobacteria and two Firmicutes to understand the mechanism of substitution mutations.

- Chapter 4 presents a graph theory-based method to estimate RNA secondary structure and a survey on a wide range of algorithms based on machine learning and specifically deep learning methods available in public domain.

- Chapter 5 presents base substitutions in tRNA gene sequences in genomes of five bacteria and discuss possible role of stem and loop motifs of the secondary structure on the base substitutions.

- Chapter 6 concludes the thesis by summarizing the works done and also lists the possible future research directions in this area.