Abstract

The majority of the genetic information of an organism is stored in the form of Deoxyribonucleic acid (DNA) sequences bundled into one or more chromosomes. DNA consists of two strands twisted around each other in a right handed fashion. Each strand is a sequence of four nucleotide bases, adenine, cytosine, guanine, and thymine, denoted by characters A, C, G, and T. The two strands are the reverse complement of each other. They are held together with the help of double hydrogen bonds (between complementary bases A and T) and triple hydrogen bonds (between complementary bases G and C). The pairing of purine (R:A/G) and pyrimidine (Y:T/C) in the double-stranded DNA is important for its structural stability and function. DNA sequence length varies from organism to organism, and in the case of bacteria is of order 10^6 bases. The sub-sequences of the DNA strands called genes are associated with different functions. The nucleotide sequence of a protein coding gene is translated in the form of triplets called codons to produce the amino acid sequence of a protein. The base sequence of a tRNA gene facilitates amino acid transportation to the site of protein synthesis owing to its typical secondary structure. A typical bacterium genome consists of a few thousand genes that essentially determine all the features.

Though the double-stranded DNA is the permanent store of all the information of an organism, it is continuously evolving under the influence of several intrinsic and extrinsic forces. As the different segments of a DNA sequence are associated with different functions, the impact of these evolutionary forces differs along the DNA sequence. Recent developments in genome sequencing techniques, gene annotation algorithms and public computational infrastructure

provide a huge volume of genetic information in several databases such as NCBI, EMBL, DDBJ, RNAStructure etc. These advancements facilitate computational studies to understand the evolutionary mechanism at the molecular level. In this thesis, we have computationally analyzed a large volume of bacterial genome sequences to address some evolutionary factors such as codon usage bias, single nucleotide polymorphism, and secondary structure motifs.

Synonymous codons coding for the same amino acid are not used in equal frequencies in a gene. This unequal usage of synonymous codons, known as codon usage bias (CUB), is observed across genomes of all kingdoms of life, from prokaryotes to eukaryotes and viruses. Genome (G+C)% among bacteria varies from less than 20% to more than 75%. Accordingly, G/C rich codons are used more frequently in G+C rich genomes, and the reverse is also true for G+C poor genomes. The pattern of CUB also differs among genes within an organism. The expression levels of growth-associated genes are much higher than other genes in fast-growing bacteria. According to the selection-mutation-drift (SMD) theory, weakly expressed genes codon usage bias is determined mainly by genome composition. In contrast, selection pressure plays a significant role in deciding codon usage bias in highly expressed genes. In this study, we have analyzed in detail the codon usage bias across several bacterial genomes considering machine learning-based approaches. This study demonstrates a commonality among bacteria regarding the behaviour of specific codons with regard to gene expression. Codon Adaptation Index (CAI) is a widely used method for measuring CUB in a gene. We have provided a web portal available at http://14.139.219.242:8003/cai to calculate CAI online. The reference set of high expression genes for more than six hundred bacteria species, yeast, and human are currently available in the web portal for calculating CAI.

Single nucleotide base substitution is the most commonly observed mutations among the organisms. As each of the four DNA bases can mutate to any of the three other bases, there are twelve possible directional substitution mutation types that include four transitions in which a purine (or a pyrimidine) is replaced

by another purine (or a pyrimidine) (R→R; Y→Y) and eight transversions in which a purine (or a pyrimidine) is replaced by a pyrimidine (or a purine) (R→Y; Y→R). These directional substitution mutations do not occur at equal frequencies in bacterial genomes for mechanistic reasons such as unequal stability among different base pairs, the differential propensity of bases to damages such as deamination, oxidation, and radiation as well as selective reasons such as the differential impact on the structure and function of DNA, RNA, and proteins. We have investigated single nucleotide polymorphism (SNPs) in intergenic regions (IRs) and four-fold degenerate sites (FFS) of the genes in genomes of three γ-Proteobacteria and two Firmicutes to understand the mechanism of substitution mutations. In general, similarity among these bacteria in terms of the SNP pattern at IRs was observed. The polymorphism patterns at FFS were unlike IRs and were also inconsistent among the five amino acids within a genome. The SNP pattern at FFS of an amino acid was comparable to the codon usage pattern.

Unlike DNA, ribonucleic acid (RNA) is primarily single-stranded in nature, but the bases of RNA have a tendency to form pairs with the help of hydrogen bonds that leads to the folding of RNA to attain a definite shape. This folded structure is called the secondary structure of RNA. Because of the secondary structure, RNA performs several important function. For instance, transfer RNA folds to form a typical clover leaf shape to participate in the process of translation. Computational algorithms are extensively used in the prediction of RNA secondary structure. In this thesis, we have developed a graph-based approach to predict the RNA secondary structure. We also reviewed RNA secondary structure estimation algorithms based on machine learning principles emphasizing recent advancements in deep learning methods.

Transversion and transition mutations have variable effects on the stability of RNA secondary structure, considering that the former destabilizes the double helix geometry to a greater extent by introducing purine: purine (R: R) or pyrimidine: pyrimidine (Y: Y) base pairs. Therefore, transversion frequency is likely lower than transition in the secondary structure regions of RNA genes. Transfer

RNA genes have a well-defined secondary structure stem, and loop motifs, and the details are available in the genomic tRNA database: GtRNAdb for computational studies. In this report, we have analyzed transition and transversion frequencies in tRNA genes secondary structure and compared with the intergenic regions in five bacterial species: *Escherichia coli, Klebsiella pneumoniae, Salmonella enterica, Staphylococcus aureus* and *Streptococcus pneumoniae* using a large genome sequence data set. Our analysis provides additional supporting evidence towards role of the secondary structure on different base substitutions.

Keywords: codon degeneracy, codon-usage-bias, optimal-codon, codon adaptation index, translational selection, gene expression, molecular evolution, single nucleotide polymorphism, base substitution, transition, transversion, intergenic region, four-fold degenerate sites, RNA secondary structure, machine learning, deep learning, pseudoknots, tRNA secondary structure.