ABSTRACT

The language of the DNA is translated into polypeptides through the gateway of genetic code table which is known as the most complex molecular process in the cell. With few exceptions in the extranuclear DNA in eukaryotes, all life forms follow the standard genetic code table. The genetic code is considered to be robust in the sense that it exhibits redundancy and tolerant to most harmful mutations. This robustness is essential to ensure the fidelity of genetic information and preserves the biological significance. The variations in the population are common in nature. The impact of such variations can be lethal in the absence of the process of natural selection. However, the variations are due to genetic drift in smaller populations where chance events play a pivotal role in shaping the fate of the variants. The significant progress in molecular evolution has led to the augmentation of various biological databases with extensive genome sequence data obtained through high-throughput techniques. Many inter-species studies in the past decades have provided the intriguing evolutionary relationship between species. However, the availability of ample number of genome sequences of the same species in various databases has enabled modern day researchers to study molecular evolution research in a different dimension as "intra-species study". In our endeavour, this present PhD thesis describes some intriguing efforts to understand the normalization of base substitution bias across genomes, to understand the amino acids exchangeability, to understand the impact of DNA replication and transcription on different intergenic regions as well as the impact of various fundamental processes on co-transcribed genes considering multiple strains of bacteria *Escherichia coli* and *Salmonella enterica*. The results of our analysis are extensively outlined in Chapters 1 through 5. The conclusion and prospects for future research from this study are separately addressed in Chapter 6.

The thesis is organized in the following order:

I. Chapter 1 includes the introduction to the thesis, objectives of the PhD work, and review of literature in the context of the objectives of the thesis. The introduction section briefly covers the brief history of the journey of molecular evolution as well as explains the basic terminologies used in the field. Following this, the objectives of the PhD research are

outlined. Subsequently, the literature review section succinctly incorporates recent advancements and knowledge relevant to the PhD objectives.

II. Chapter 2 discusses the importance of consideration of codon degeneracy and pretermination codons (PTC) in estimation of $\frac{ti}{tv}$ ratio in coding sequences. It also discusses the methodology of implementation of the improved $\frac{ti}{tv}$ ratio and describes the application of the equation in visualization of frequencies of variations among different substitutions. It also describes about the accessibility of the software developed in our laboratory to estimate different $\frac{ti}{tv}$ ratios across the genomes accessible at https://github.com/CBBILAB/CBBI.git.

III. Chapter 3 discusses the non-synonymous variations in the coding sequences. It highlights the detrimental impact of non-synonymous transitions and transversions in coding sequences. It follows the methodology developed in Chapter 2. Our findings in terms of 64*64 matrix and 20*20 matrix unveil many novel findings regarding individual amino acids level exchangeability in *Escherichia coli*. It also unearths the role of codon degeneracy in non-synonymous variations. The findings provide many key points to ponder about the organization of the genetic code table's evolutionary prospects.

IV. Chapter 4 discusses the role of replication-associated mutations and transcription-associated mutations towards bringing variation in genomes. It describes a comparison between intra operon intergenic regions and inter operon intergenic regions in terms of their polymorphism spectra. Surprisingly, the intra-operon intergenic regions were discovered with a minimum of 2.5-fold lesser polymorphism frequency than the inter-operon intergenic regions. The former was believed to have higher mutation frequency due to the repetitive cycles of transcription. It hinted towards selection in the intra-operon intergenic regions. This work has been published as "**Beura P. K., Sen P., Aziz R., Satapathy S. S., Ray S. K The transcribed intergenic regions exhibit lower frequency of nucleotide polymorphism than the untranscribed intergenic regions in the genomes of Escherichia coli and Salmonella enterica. Journal of Genetics, 2023."**

V. Chapter 5 discusses the polymorphism spectra comparison between two co-transcribed genes. It described the scenario giving rise to mutation in the prokaryotic genomes. Adverse to our anticipation, the polymorphism spectra, transition to transversion ratio, individual

strains level study as well as the phylogeny revealed the polymorphism spectra between two co-transcribed genes to be different. It also provides the role of higher degenerate codons in the establishment of different polymorphism spectra between co-transcribed genes. This work has been published as "**Beura P.K., Sen P., Aziz R., Chetia C., Dash M., Satapathy S. S., Ray S. K Difference in synonymous polymorphism related to codon degeneracy between co-transcribed genes in the genome of Escherichia coli. Curr. Sci. 2023**."

VI. Chapter 6 summarizes the overall work done in the PhD work including the key findings. It lists the prospects of possible future work based on the work done in this thesis.