A Study on Single Nucleotide Variations in Different Regions of *Escherichia coli* Genome Sequences

A thesis submitted in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy

Mr. Pratyush Kumar Beura
Registration No: TZ23000709 of 2023

Department of Molecular Biology and Biotechnology
Tezpur University, Tezpur-784028
Assam, India

November 2024

Chapter 6

Conclusion And Future Prospects

CHAPTER 6

CONCLUSION AND FUTURE PROSPECTS

Behind the curtain of a redundant and robust genetic code table there lies a plethora of possibilities of research in molecular evolution. The genetic code table identifies heterogeneity in terms of the puzzling phenomenon of codon usage bias. However, this codon usage bias has helped us to study SNVs in coding sequences of many organisms. The findings of the research work as well as the future prospects are described as follows:

- We observed a distinctive role of codon degeneracy and PTC in the comparative study between the conventional $\frac{ti}{tv}$ and improved $\frac{ti}{tv}$. Our improved estimator correctly justifies the normalization of observed *ti* (*tv*) and estimated *ti* (*tv*) in terms of the composition of different degeneracy classes. The differential strength of selection of synonymous and non-synonymous variations encourages to estimate the $\frac{ti}{tv}$ separately for both synonymous and non-synonymous variations**.** The comparatively lower Pearson correlation value between *Sti* and *Nti* between both the estimators gives a sanity check to the accuracy of the improved estimator in visualisation of the process of selection. In the future, applying this estimator to other organisms could help in accurately estimating improved $\frac{ti}{tv}$ values and also verify the impact of codon degeneracy across different species.

- The interesting part regarding the non-synonymous variations was revealed in form of the role of codon degeneracy in pursuance of *Nti,* the higher *Nti* values in FFD codons than the TFD codons. However, the reason behind this *ti* bias between the FFD codons and the TFD codons is yet to be understood in terms of the confounding effect of mutation and selection. Among several findings, the six times higher frequency value between Cys→Ser (SB) over Cys→Ser (FB) as well as AUG→AUA have shown a new direction in understanding the structure and evolution of the genetic code table. Among different intrinsic factors, hydrophobicity, stereochemistry, and economy play vital roles in amino acid exchangeabilities. Therefore, the amino acid exchangeability could be better studied by the augmentation of proteomics in different organisms. Most importantly, our novel

finding as the role of higher *Nti* frequency in FFD codons over TFD codons needs to be substantiated in other organisms.

- Temporary exposure of double-stranded DNA during essential processes such as DNA replication and transcription increases its susceptibility to base substitution variations. Our findings in chapter 4 and 5 demonstrates the impact of such fundamental processes on co-transcribed genes and the impact of mutation and/or selection on different intergenic regions. The Intra-operon IGRs shows the lower possibility of mutations due to its regulatory role in gene expression. However, such findings itself are intriguing and invokes to understand the complex eukaryotic systems in depth. Although different fundamental processes vary between the prokaryotic and eukaryotic systems, it will be easy to differentiate between the contributions of some complicated concepts like genetic drift, selection, and context dependent mutations even in eukaryotic systems.

- In conclusion, if mutation is a random event, selection gives it the correct direction for the sustenance of an organism. The study of SNVs within the genetic code table, facilitated by the codon substitution model, has gained popularity and widespread acceptance due to its comprehensive coverage of mutation and selection. The *ti* bias has always been the key point of research. The predominant presence of *Nti* in FFD codons particularly gives us a new direction to study the protein folding pattern. This could ultimately answer the quest for the amino acid's assignment to the codons from its evolutionary prospects.

- The possibility of codon context cannot be avoided. It is still unknown regarding its uniform nature in bacteria except for pyrimidine dimer, where the cell possesses nucleotide excision repair and photoreactivation against it. Considering a wide range of CUB in bacterial species, and most of the coding sequences fall under low expression, the context dependent mutation seems more bacterial specific. In that case it is not easy to differentiate between context dependent mutation and selection. More research is required in the future to overcome the dilemma.

  Our findings primarily address some fundamental questions in the field of molecular evolution, including the long-standing issue of whether codons are assigned to amino acids or if amino acids are assigned to codons. Additionally, the distribution of amino

acids in specific boxes and the assignment of codons across different organisms, from prokaryotes to complex eukaryotes, remains a perplexing phenomenon. This assignment underpins the basic rules of life. The non-uniformity in codon degeneracy also remains unresolved. As demonstrated in this thesis, the impact of codon degeneracy is clearly illustrated, and we anticipate that future research should extend to other species beyond *E. coli*.

Our study is based on theoretical grounds and based on datasets available in the public database. We did not have an extensive experimental study in terms of the outcome of the work. here may be potential sequencing errors, as the data comes from different laboratories and was submitted at different times. Exploring more distantly related species could introduce additional complexity to the results, as codon usage bias (CUB) and mutational patterns are often species-specific phenomena. Additionally, conducting separate analyses of genes located on the leading and lagging strands could yield valuable insights in prokaryotes, given that many essential genes reside on the leading strand to prevent collisions between DNAP and RNAP. Interestingly, our findings on the lower amino acid exchangeability in the SB of Ser and Arg than the FB of both of them needs further investigations in other organisms to understand the mystery behind the presence of SFD codons. We delved more into the TFD codons and FFD codons and tried to understand their role and importance from many aspects of evolution. However, the functional groups of the amino acids encoded by the FFD codons are usually smaller than the functional groups of the amino acids encoded by the TFD codons. Also, amino acids encoded by the TFD codons confer many vital functions in the cells ubiquitously. The preliminary knowledge on such evolutionary categorization between two prominent degeneracy classes is thought provoking. Therefore, much remains to be understood regarding the structure and organization of the genetic code table. Future research will unravel many insights about it.