Computational Analysis of Codon Usage Bias, Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Piyali Sen

Enrollment No. CSP17001

Registration No. TZ189834 of 2018

Department of Computer Science and Engineering

School of Engineering, Tezpur University

Tezpur, Assam, India - 784028

November, 2022

Chapter 6

Conclusion and future direction

Genomes of organisms are continuously evolving under the influence of different selection and mutation factors. The role of these factors in the evolution of different functional regions in a genome is not yet fully understood. Scientists across disciplines are analyzing genome sequences to understand various aspects of evolution at the molecular level. In this thesis report, we have analyzed a large volume of bacterial genome sequences available in the public domain to understand codon usage bias, single nucleotide polymorphism in intergenic regions (IRs) and four-fold degenerate sites (FFS) of the genes, RNA secondary structures and base substitution in stem and loop motifs. In our study, we have used computational methods developed in this study, including machine learning algorithms. We have also developed software tools for measuring codon usage bias and predicting RNA secondary structure and made them available in web portals for research purposes. Our study is summarized in the following subsections.

6.1 Codon usage bias in bacterial genomes

Translational selection for optimum codon recognition is a factor influencing codon usage bias in genomes. For example, Leucine codon CUG frequency is higher in high expression genes (HEG) than the low expression genes (LEG) in *Escherichia*

coli genome, because of higher cognate tRNA gene abundance for CUG codon facilitating faster translation. Other similar notable codons are the Glycine codon GGU and the Arginine codon CGU. Apart from these positively selected codons, specific codons such as Glycine codon GGG are used in lower frequency in HEG compared to LEG, possibly because of any retarding effect on optimum gene expression. Using the machine learning approach, we have identified prominent codons from different degenerate groups influencing gene expression in bacteria. The findings of our study are:

- In *Escherichia coli* genome, the usage of codons such as UCU (Ser), CUG (Leu), GGG (Gly), CGG (Arg) etc., were influenced maximum by the gene-expression. In the other hand, Cys codons, such as UGU and UGC, and Lys codons such as AAA and AAG, were least influenced by gene expression.
- In general, codons with higher degeneracy were more important toward classifying high and low-expression genes in the *Escherichia coli* genome.
- The codon adaptation index (CAI) correlated better with gene expression in Escherichia coli for the genes rich in four-fold degenerate amino acid codons compared to the genes rich in two-fold degenerate amino acid codons.
- Among 683 bacterial species, we observed that the Cys (UGU/UGC) and Ser (AGU/AGC) codons were the least different between the two groups of genes across these bacterial species. Codons such as CGA, CUG, GGG, GCC, ACC, AUA, and AUC were identified to be influenced by the gene expression across the majority of these species.
- Our study demonstrated a commonality among bacteria regarding behavior of certain codons with regard to gene expression.

We also have made available a web portal for measuring Codon Adaptation Index (CAI) online, which is one of the widely used methods to measure

codon usage bias (CUB). A reference set of organism-specific high-expression genes is a compulsory requirement for calculating CAI values. The earlier software were providing reference sets for only a limited number of organisms. We have developed the webserver to calculate CAI value where we have provided reference sequences of high expression genes for more than 600 bacteria, yeast, and human. The web server is available at http://14.139.219.242:8003/cai.

6.1.1 Future directions

Future work lies in further critical analysis of the common observation on codon usage across bacterial species. It has been reported in research literature that the extent of codon usage bias corresponds to the lifestyle of the organism and also influenced by phylogeny. Organisms surviving in a wide range of habitats exhibit great extents of codon usage bias consistent with their need to adapt efficiently to different environments. Therefore, it will be interesting to do a comparative study on codon usage between pathogens/non-pathogens, thermophiles/non-thermophiles etc.

6.2 Single nucleotide polymorphism in bacterial genomes

Coding regions of a genome are under different evolutionary pressure compared to the intergenic regions (IRs) owing to the difference in associated functions. As base substitution at four-fold degenerate sites (FFS) in the coding region do not alter the amino acid, these regions are believed to be evolving under near neutrality. Also, the non-regulatory portions of larger intergenic regions are believed to be neutral. In this thesis report, we have done a comparative analysis of nucleotide polymorphism spectra at IRs and FFS in hundreds of strains of three γ-Proteobacteria, namely *Escherichia coli* (*Ec*), *Klebsiella pneumoniae* (*Kp*), *Salmonella enterica* (*Se*) and

two Firmicutes, namely *Staphylococcus aureus* (*Sa*), *Streptococcus pneumoniae* (*Sp*). Major findings on the base substitutions in the genomes of these bacteria are as follows:

- The patterns of transitions were alike between the leading and the lagging strands, whereas transversions were variable in the IRs.
- Contrasting trends of complementary polymorphisms such as C→T vs G→A as well as A→G vs T→C were observed in the IRs. These results vindicated similar replication-associated strand asymmetry regarding cytosine and adenine deamination, respectively. The polymorphism pattern at FFS was different from that of the IRs, and its frequency was always more than the IRs. The polymorphism patterns within a bacterium were inconsistent across the five amino acids, which neither the replication nor the transcription-associated mutations could explain. The polymorphism at FFS coincided with amino acid-specific codon usage bias in the five bacteria. Strand asymmetry in nucleotide composition could be explained by the polymorphism at FFS but not at the IRs.

6.2.1 Possible future directions

Polymorphisms at FFS observed in our study suggest that the base substitutions at this site might not be treated as nearly neutral, unlike in IRs. Base substitution at FFS can be further analyzed to understand the impact of any selection pressure on synonymous codon changes.

Out of the 16 boxes in the genetic code table, there are five split boxes (SBs) where the two NNY codons are assigned to one amino acid, and the other two NNR codons are assigned to another amino acid. Base substitutions NNY→NNR as well as NNR→NNY can be analyzed for understanding codons to amino acids assignment in genetic code table.

6.3 RNA Secondary structure estimation and base substitutions in tRNA gene secondary structure motifs

Bases in RNA primary sequence have a tendency to attain a definite shape called secondary structure, important for its structure and function. In order to supplement scientific methods for finding RNA secondary structure, structure computational algorithms have been developed since several years. In this thesis report we have done a systematic review on these computational methods. Some of the contributions from this thesis report on RNA secondary structure as follows:

- We have also reviewed a wide range of algorithms proposed to predict RNA secondary structure, focusing on machine learning and deep learning methods.
- We have developed a graph-based algorithm to predict RNA secondary structure. The web-based implementation of the algorithm for small RNA sequences is available at the website http://14.139.219.242:8003/rna_struct.

Transition and transversion base substitutions have a different impact on the stem and loop regions in the RNA secondary structure. In this thesis report, we have done a systematic study on base substitution in the well-defined secondary structure of transfer RNA genes in five bacterial species *Escherichia coli, Klebsiella pneumoniae, Salmonella enterica, Staphylococcus aureus, and Streptococcus pneumoniae*. Some of the findings on the influence of base substitutions on tRNA gene secondary structure are as follows:

- Transition to transversion ratio in the stem region was observed to be higher than that in the loop region.
- Non-compensatory substitutions were observed more frequently than compensatory substitutions in stem regions of tRNA genes.

- Transition substitutions from amino to keto bases were found to be significantly higher than that from keto to amino bases in the stem regions across the five bacteria.

6.3.1 Possible future directions

Keeping large volumes of information on RNA secondary in the public domain, computationally efficient algorithms can be developed using machine learning/deep learning algorithms and also considering that the RNA folding is co-transcriptional. Base substitution study in stem-loop motifs of tRNA genes can be further extended to secondary structures of ribosomal RNA genes, rho-independent transcription termination sites in prokaryotes.