Contents

1 Introduction **1**

1.1 Codon usage bias in bacterial genomes 4

1.2 Single nucleotide polymorphism in bacterial genomes 7

1.3 RNA secondary structure estimation 8

1.3.1 Thermodynamic features considered in RNA secondary structure prediction algorithms 10

1.4 Challenges 12

1.5 Objective of the thesis 13

1.6 Organization of the thesis 14

2 Codon usage bias in bacterial genomes **16**

2.1 Introduction 16

2.2 Importance of the codons towards classifying high and low expression genes: a machine learning-based analysis 20

2.2.1 Machine learning-based estimation of the importance of the codons in classifying high and low expression genes 22

2.2.2 Machine learning-based analysis of CUB between high and low expression genes . . . 23

2.2.3 Results and Discussion . . . 26

2.2.3.1 RSCU values of the codons do not contribute equally towards classifying high and low expression genes . . . 26

2.2.3.2 Codons with higher degeneracy are more important toward gene expression prediction . . . 30

2.2.3.3 Codon importance features of 683 bacteria species . 32

2.3 Improved Implementation of CAI . . . 35

2.3.1 Limitations in existing implementation of CAI . . . 35

2.3.1.1 Considering default *E. coli* reference set may generate erroneous result . . . 36

2.3.2 Improved Codon Adaptation Index web portal . . . 37

2.3.2.1 Reference set of high expression genes available in the web portal . . . 37

2.4 Conclusion . . . 39

3 Single nucleotide polymorphism in bacterial genomes 41

3.1 Introduction . . . 41

3.2 Materials and Methods . . . 44

3.2.1 Segregating the leading and the lagging strands in bacterial chromosomes . . . 44

3.2.2 Polymorphism analysis at IRs and at FFS in CDS of the bacterial chromosomes 46

3.3 Results 51

3.3.1 Similar pattern of single nucleotide polymorphism at intergenic regions across these bacteria 51

3.3.2 The polymorphism pattern at the four-fold degenerate site (FFS) is different from that at IRs 56

3.3.3 The polymorphism at the four-fold degenerate site coincides with codon usage bias 61

3.3.4 The polymorphism spectra at the four-fold degenerate sites in the high expression genes is different from that in the four-fold degenerate sites in the whole genome as well as IRs 65

3.4 Discussion 65

3.5 Conclusion 69

4 RNA secondary structure estimation 71

4.1 Introduction 71

4.1.1 RNA secondary structure representation format 72

4.1.1.1 Dot-bracket notation 72

4.1.1.2 Connectivity Table (CT) format 73

4.1.1.3 BPSEQ format 73

4.1.2 RNA secondary structure visualizing tool 74

4.1.3 Computational methods used for RNA secondary structure prediction . . . 74

4.2 Improved method to predict RNA secondary structure based on MIS 76

4.2.1 Materials and Method . . . 76

4.2.1.1 Algorithm RNA structure estimation . . . 79

4.2.1.2 Description of how to use the TU web server: . . . 83

4.2.1.3 Performance Measurement . . . 83

4.2.2 Results and Discussion . . . 84

4.3 A review on RNA secondary structure prediction using deep learning methods . . . 86

4.3.1 Computational methods that predict pseudoknot-free structures . . . 86

4.3.1.1 Learning based RNA folding methods . . . 86

4.3.1.2 Deep-learning based RNA folding methods . . . 86

4.3.2 Computational Methods that predict pseudoknotted structures 87

4.3.2.1 Machine learning methods to predict pseudoknotted structures . . . 87

4.3.2.2 Deep learning methods that can predict RNA secondary structure with pseudo-knots . . . 87

4.3.3 Comparative Results and Discussion . . . 95

4.3.3.1 RNA secondary structure information . . . 95

4.3.3.2 Performance measurement . . . 95

4.3.3.3 Performance metrices 96

4.3.3.4 Results 97

4.3.4 Conclusion 99

5 Base substitutions in tRNA gene secondary structure motifs 100

5.1 Introduction 100

5.2 Materials and Methods 103

5.2.1 Extracting intergenic regions, tRNA genes and segregating loop and stem regions 103

5.2.2 Segregating compensatory and non-compensatory substitutions in stem regions 106

5.2.3 Finding substitutions from the sequence alignments 106

5.2.4 Visualization of 2-D and 3-D structures of tRNA genes . . . 107

5.3 Results 107

5.3.1 Higher transition to transversion ratio in tRNA genes than intergenic regions 107

5.3.2 Higher transition to transversion ratio at the stem regions than the loop regions within tRNA genes 108

5.3.3 Biased transition substitution towards keto bases in the stem region of tRNA genes 110

5.4 Discussion 112

5.5 Conclusion 114

6 Conclusion and future direction 116

6.1 Codon usage bias in bacterial genomes 116

6.1.1 Future directions 118

6.2 Single nucleotide polymorphism in bacterial genomes 118

6.2.1 Possible future directions 119

6.3 RNA Secondary structure estimation and base substitutions in tRNA gene secondary structure motifs 120

6.3.1 Possible future directions 121

A Data related to Experiments **122**

A.1 Codon Usage Bias in bacterial genomes Supplementary Data 122

A.1.1 Importance of the codons estimated in randomized gene sets of *E. coli* genome 122

A.1.2 List of high and low expression genes considered in this study 123

A.1.3 Details of 683 bacteria species 123

A.1.4 Important codons (marked as 'C') of 683 bacteria species . . 124

A.2 Single nucleotide polymorphism in bacterial genomes Supplementary Data 124

A.2.1 Finding single nucleotide polymorphism from the sequence alignments 124

A.2.2 GC and AT skews in chromosomes 125

A.2.3 Phylogeny of the five bacteria *Ec, Kp, Se, Sa* and *Sp* constructed using *rpoB* and *rpoC* gene sequence 125

A.2.4 Polymorphism frequency distribution among strains of the bacteria using rpoB and rpoC gene sequences 126

A.2.5 (G+C)% in whole genome sequence (WGS), at IRs and FFS of the five amino acids in genome, as well as in the high expression genes (HEGs) of LeS and LaS 126

A.2.6 Polymorphism spectra at FFS of five amino acids in the leading and the lagging strands of five bacteria 126

A.2.7 Comparison of polymorphism frequency at WGS, HEGs and IRs 126

A.2.8 Correlation of genome size with genome G+C% in bacterial groups 127

A.2.9 Intra-species polymorphism in *rpoB* and *rpoC* genes and inter-species substitution frequency in *rpoB* and *rpoC* comparison between *Ec*, *Kp*, and *Se* as well as between *Sa* and *Sp* 127

A.2.10 Nucleotide composition at FFS of different amino acids in the high expression genes (HEGs) of five bacteria 127

A.3 RNA Secondary Structure Estimation Supplementary Data 127

A.3.1 Details of 25 Pseudoknotted Structure 127

A.3.2 Gene wise performance metrices 128

A.3.3 Count of RNAs, mean, standard deviation, minimum value, 25% or Q1, 50% or Q2, 75% or Q3, maximum value, Interquartile range (IQR) of Accuracy, Precision, Recall, Specificity and F1 Score of the 25 pseudoknotted structures . . . 128

A.4 Base substitutions in tRNA genes Supplementary Data 129

A.4.1 Substitutions in predicted tRNA Secondary structure 129

A.4.2 Base substitution frequency in non-compensatory and compensatory stem 130

A.4.3 G:T mispairing and A:C mispairing energy calculation study 131

A.4.4 Amino(A/C) → Keto (G/T), Keto (G/T) → Amino (A/C) polymorphism frequencies IRs of five bacteria 132

A.4.5 Details of tRNA studied in dot bracket notation 132

A.4.6 List of tRNA genes of five bacteria 132

A.4.7 Proportionate fold increase of ti as well as tv in the IRs than tRNA 133

A.4.8 Energy calculation of different base pairing 133

A.4.9 Isoacceptor Details 133

B Publication list **134**