

*Dedicated to my parents and brother*

# Declaration

I certify that

- The work contained in the dissertation is original and has been done by myself under the general supervision of my supervisors.
- The work has not been submitted to any other institute for any degree or diploma.
- I have followed the guidelines provided by Tezpur University in writing the thesis.
- I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the university.
- Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the dissertation and giving their details in the references.

*Piyali Sen*  
25/02/2023

**Piyali Sen**



Department of Computer Science & Engineering  
Tezpur University  
Napaam, Tezpur- 784028, Assam, India.

Dr. Siddhartha Sankar Satapathy  
Associate Professor

Phone: +91-3712-275117  
Fax: +91-3712-267005/6  
E-Mail: ssankar@tezu.ernet.in

### Certificate of Supervisor

This is to certify that the thesis entitled “**Computational Analysis of Codon Usage Bias, Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences**” submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Piyali Sen** under my supervision and guidance.

All helps received by her from various sources have been duly acknowledged. No part of this thesis has been submitted else where for award of any other degree.

*S. S. Satapathy*  
29/3/2023  
Signature of Supervisor

(Siddhartha Sankar Satapathy)

Associate Professor

Department of Computer Science and Engineering

Tezpur University

Assam, India-784028



Department of Molecular Biology & Biotechnology  
Tezpur University  
Napaam, Tezpur- 784028, Assam, India.

Dr. Suvendra Kr. Ray  
Professor

Phone: +91-3712-275406  
E-Mail: [suven@tezu.ernet.in](mailto:suven@tezu.ernet.in)

### Certificate of Co-Supervisor

This is to certify that the thesis entitled “**Computational Analysis of Codon Usage Bias, Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences**” submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Piyali Sen** under my personal co-supervision and guidance.

All helps received by her from various sources have been duly acknowledged. No part of this thesis has been submitted else where for award of any other degree.

  
Signature of Co-Supervisor

(Suvendra Kr. Ray)

Professor

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

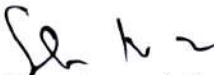


## Certificate

This is to certify that the thesis entitled “**Computational Analysis of Codon Usage Bias, Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences**” submitted by **Piyali Sen** to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering has been examined by us on 29/03/2023..... and found to be satisfactory.

The Committee recommends for award of the degree of Doctor of Philosophy.

  
Signature of Supervisor

  
Signature of Co-Supervisor

  
Signature of External Examiner

# Acknowledgment

First and foremost, I offer my deepest gratitude to my supervisor, Dr. Siddhartha Sankar Satapathy, and my co-supervisor, Prof. Suvendra Kumar Ray of the Molecular Biology and Bio-technology Department, who introduced me to the research. They taught me the basics of molecular biology, the RNA structure, and also to develop applications and computational tools based on these concepts. Without their constant guidance, support, monitoring, inspiring advise, suggestions, thoughtful discussion, and sensible review, this research could not have been completed.

I would also like to express my appreciation to the members of my doctoral committee, Prof. Utpal Sarma, Dr. Rosy Sarmah, Prof. Suvendra Kumar Ray, the head of the department, the doctoral research committee, Prof. N. Sarma, Dr. Arindam Karmakar and all the faculty members of the department of Computer Science and Engineering, Tezpur University, for their constructive comments on my thesis.

I am also deeply indebted to Prof. Edward J. Feil, University of Bath, whose wisdom, advises and critical review smoothened the path of my Ph.D. I would also like to acknowledge Dr. Harry Thorpe for providing the sequence data of bacterial species. I sincerely thank Prof. Ramesh Deka, Dr. Nima Dondu Namsa and Dr. Aditya Kumar for their constant support and valuable comments on my research papers. I am also thankful to UGC, Government of India, for providing me the financial support of UGC NET-JRF fellowship. I would also like

---

to thank all the anonymous reviewers of the published works and the thesis for their constructive comments.

I got a lot of support and input from my fellow research scholars and friends in the Departments of Computer Sc. and Engineering as well as Molecular Biology and Biotechnology. I would like to thank Annushree Kurmi, Ruksana Aziz, Pratyush Kumar Beura, and Kristi Kabyashree. I would also like to thank Debapriya, Abdul, Saumita, Arundhati, Carynthia, Deena, Kaushal da, Nirmal da, Mampi di, Jyoti, Vijay, Priyanjana, Tapas, Tamal, Kunal, Muzakkir, Santanu and others who have directly or indirectly helped and encouraged me throughout the Ph.D journey and with whom I have spent many unforgettable times together.

I would like to pay my regards to my parents, my younger brother for their love, constant support, encouragement and belief in me. I would like to thank the almighty lord for outpouring positive energy in my life.

*Piyali Sen*  
25/02/2023

**Piyali Sen**

# List of Figures

1-1	Different substitution mutations in the genome . . . . .	8
1-2	Different faces in a RNA secondary structure . . . . .	10
2-1	Distribution of CAI values in <i>Bradyrhizobium japonicum</i> . . . . .	21
2-2	Distribution of CAI values in <i>Staphylococcus aureus</i> . . . . .	22
2-3	Codon importance estimation using Boruta Algorithm with the help of a hypothetical set of genes and RSCU values of a few codons . . .	24
2-4	Workflow of Boruta Algorithm . . . . .	24
2-5	Confusion Matrix along with Precision, TPR or Recall, FPR, and Accuracy . . . . .	25
2-6	Importance of the codons in <i>E. coli</i> genome . . . . .	27
2-7	ROC curve of confirmed and rejected features using RF classifier . .	28
2-8	ROC curve of FFS and TFS features using RF and XGB . . . . .	31
2-9	Correlation of CAI and gene expression based on the composition of FFS and TFS in <i>E. coli</i> . . . . .	32
2-10	Importance of the codons in <i>Wigglesworthia glossinidia</i> genome . .	33



## List of Figures

---

2-11	Importance of the codons in <i>Anaeromyxobacter dehalogenans</i> genome	34
2-12	Frequency distribution of importance value in 683 organisms . . . . .	35
3-1	A schematic view of the distribution of IRs, CDS, tRNA and rRNA in the leading and lagging strands in double stranded DNA . . . . .	45
3-2	Inter-species phylogeny of 12 different bacteria species constructed using <i>rpoB</i> gene sequences . . . . .	47
3-3	Inter-species phylogeny of 12 different bacteria species constructed using <i>rpoC</i> gene sequences . . . . .	47
3-4	Inter-species phylogeny of 12 different bacteria species constructed using <i>rpoB</i> gene sequences . . . . .	48
3-5	Intra-species phylogeny of 12 strains of <i>E.coli</i> bacteria using <i>rpoB</i> gene sequences . . . . .	48
3-6	Intra-species phylogeny of 12 strains of <i>E.coli</i> bacteria using <i>rpoC</i> gene sequences . . . . .	49
3-7	Intra-species phylogeny of 12 strains of <i>E.coli</i> bacteria using <i>dnaK</i> gene sequences . . . . .	49
3-8	Interspecies and intraspecies pairwise difference distribution . . . . .	50
3-9	Difference between complementary transition polymorphisms in the LeS and the LaS at IRs . . . . .	55
3-10	Polymorphism frequency at FFS across five amino acids in five bacteria	58
3-11	Difference between complementary transition polymorphisms in the LeS and the LaS at FFS . . . . .	60
4-1	RNA Secondary Structure with Stem and Hairpin loop . . . . .	76

## List of Figures

---

4-2	Steps followed to detect RNA Secondary Structure . . . . .	78
4-3	Box-plot of Accuracy, Precision and F1 Score for pseudoknotted structures . . . . .	98
5-1	Different substitution mutations in the genome . . . . .	101
5-2	Effect of substitutions on secondary structure in a hypothetical RNA sequence . . . . .	102
5-3	Compensatory or non-compensatory polymorphisms in secondary structure of a hypothetical RNA sequence . . . . .	106
5-4	Ratio of $t_i$ to $t_v$ in loop and non-compensatory stem regions in tRNA of five bacteria . . . . .	109
5-5	Polymorphism frequencies in stem region of tRNA genes among Amino(A/C) $\rightarrow$ Keto (G/T), Keto (G/T) $\rightarrow$ Amino (A/C) . . . . .	111
5-6	Amino(A/C) $\rightarrow$ Keto (G/T), Keto (G/T) $\rightarrow$ Amino (A/C) polymorphism frequencies in loop region of tRNA genes . . . . .	111
A-1	Importance of the codons estimated in randomized gene sets of <i>E. coli</i> genome . . . . .	123
A-2	Determining single nucleotide polymorphism from the sequence alignments . . . . .	125
A-3	Substitutions in predicted tRNA Secondary structure of <i>Ec</i> Gln tRNA with UUG anti-codon . . . . .	129
A-4	Base substitution frequency in non-compensatory and compensatory stem . . . . .	130
A-5	G:T mispairing and A:C mispairing energy calculation study . . . . .	131

**List of Figures**

---

A-6 Amino(A/C)  $\rightarrow$  Keto (G/T), Keto (G/T)  $\rightarrow$  Amino (A/C) poly-  
morphism frequencies IRs of five bacteria . . . . . 132

# List of Tables

1.1	Genetic code table . . . . .	5
2.1	RSCU values in high and low expression genes in <i>E. coli</i> . . . . .	29
2.2	Confusion matrix matrices of Random Forest (RF) and XGBoost (XGB) Classifier . . . . .	31
2.3	Details of bacteria whose reference set of high expression genes available in the web portal . . . . .	38
3.1	Compositional features of IRs in LeS and LaS in five bacteria . . . . .	53
3.2	Polymorphism spectra at IRs . . . . .	54
3.3	Comparison between transition-transversion polymorphism at FFS of five bacteria . . . . .	59
3.4	Nucleotide frequency at FFS . . . . .	62
3.5	Polymorphism spectra at FFS of five amino acids in the leading and the lagging strands of five bacteria . . . . .	63

4.1	Stacking Energy:- The table presents stacking energy, where the leftmost column represents the current base pair and the topmost row represents the next base pair in the stack. For example, value in row 2, column 1 represent the energy when C/G is followed by A/U [221]. . . . .	80
4.2	Tinoco's Stability Number . . . . .	81
4.3	<b>Comarative Result.</b> L: Sequence Length (number of bases), SS: Sensitivity, SP: Specificity and CC: Correlation coefficient. . . . .	84
4.4	Deep learning architectures that predict pseudoknotted RNA Secondary Structure . . . . .	89
4.5	Confusion Matrix . . . . .	95
4.6	Anomalous Representation . . . . .	96
5.1	ti/tv ratio in tRNA genes and Intergenic Regions (IRs) . . . . .	105

# Glossary of Terms

A	Adenine
ADJ	Adjacency Graph
AUC	Area Under Curve
BF	Bifurcation loop
BL	Bulge Loop
BLSTM	Bi-directional Long Short Term Memory
BP	Base Pair
BPM	Base Pair Matrix
C	Cytosine
CAI	Codon Adaptation Index
CC	Correlation Coefficient
CDS	Coding sequence
CFG	Context Free Grammar
CG	Circle Graph
CHG	Circle Graph with Hairpin Loop
CNN	Convolutional Neural Network
CQS	Count consecutive Stacking region
CS	Count Stacking region
CT	Connectivity Table
CUB	Codon Usage Bias
DDBJ	DNA Data Bank of Japan
DNA	Deoxy-ribonucleic Acid
<i>Ec</i>	<i>Escherichia coli</i>
EMBL	European Molecular Biology Laboratory
FCL	Fully Connected Layer
FFS	Four-Fold degenerate Sites
FN	False Negative
FP	False Positive
FPR	False Positive Rate
G	Guanine

GRU	Gated Recurrent Unit
H	Hairpin loop
HEG	High Expression Gene
I	Internal Loop
IBPMP	Improved Base Pair Maximization Principle
IQR	Inter-quartile Range
IRs	Intergenic Regions
K	Keto
<i>Kp</i>	<i>Klebsiella pneumoniae</i>
L	Length of RNA Sequence
LaS	Lagging Strand
LEG	Low Expression Gene
LeS	Leading Strand
M	Amino
MIS	Maximum Independent Set
MPSA	Maximum Probability Sum Algorithm
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
OC	Optimal Codon
ORF	Open Reading Frame
PACO	Parallel Ant Colony Optimization
PCA	Principal Component Analysis
PDB	Protein Data Bank
PK	PseudoKnot
R	Purine
ResNets	Residual Networks
RF	Random Forest
RNA	Ribonucleic Acid
RNN	Recurrent Neural Network
ROC	Receivers Operating Curve
RSCU	Relative Synonymous Codon Usage
S	Strong base
<i>Sa</i>	<i>Staphylococcus aureus</i>
SCFG	Stochastic Context Free Grammar
<i>Se</i>	<i>Salmonella enterica</i>
SE	Stacking Energy
SMD	Selection Mutation Drift
SNP	Single Nucleotide Polymorphism
<i>Sp</i>	<i>Streptococcus pneumoniae</i>

SP	Specificity
SRP	Signal Recognition Particle
SS	Sensitivity
SVM	Support Vector Machine
T	Thymine
<i>ti</i>	transition
TN	True Negative
TP	True Positive
TPR	True Positive Rate
tRNA	transfer Ribonucleic Acid
TSN	Tinoco's Stability Number
<i>tv</i>	transversion
U	Uracil
W	Weak base
XGBoost	Extreme Gradient Boosting
Y	Pyrimidine