Prediction of small RNA and its Targets in Bacteria

A dissertation submitted in partial fulfillment of the requirements for the M.Tech.(Computer Science) degree of Indian Statistical Institute

By

Laxman A
Roll No. CS0917

Under the supervision of

Prof. Sanghamitra Bandyopadhyay
Machine Intelligence Unit

INDIAN STATISTICAL INSTITUTE
203, Barrackpore Trunk Road
Kolkata-700108

Indian Statistical Institute

CERTIFICATE

This is to certify that the thesis entitled '*prediction of small RNA and its targets in bacteria* ' is submitted in the partial fulfillment of the degree of M.Tech. in Computer Science at Indian Statistical Institute, Kolkata.

The work out by Laxman A under my supervision and guidance is adequate, in scope and quality as a dissertation for the required degree. It is further certified that no part of this thesis has been submitted to any other university or institute for the award of any degree or diploma.

Prof. Sanghamitra Bandyopadhyay
(Supervisor)

Countersigned
(External Examiner)
Date: of July 2011

Acknowledgment

I take this opportunity to thank Prof. Sanghamitra Bandyopadhyay, Machine Intelligence Unit, ISI Kolkata for her valuable guidance, inspiration. Her encouraging words have always kept me sprits up. I am grateful to her for providing me to work under her supervision.

I would like to thank Dr Saikat Chakrabarti, IICB Kolkata, for giving opportunity to work with him. Finally I would like to thank Avijit (IICB), Malay Bhattacharya, Ramkrishna Mitra, Ramachandra Murthy, Tapas Bhadra, and my wife Chandrakala for their support and motivation.

Laxman A
M.Tech. (CS)

Contents

1 Introduction 1

1.1 Basic Molecular Biology 1

1.1.1 DNA 2

1.1.2 Bases 2

1.1.3 Nucleotides 3

1.1.4 RNA 3

1.1.5 Chromosome 4

1.1.6 Gene 4

1.1.7 Genome 4

1.2 Central Dogma of Molecular Biology 5

1.2.1 DNA Replication 5

1.2.2 Transcription 6

1.2.3 Translation 7

1.3 Introduction to MicroRNAs (miRNA) 8

1.4 Introduction to small RNAs(sRNA) 8

1.5 Scope of dissertation 10

2 Analysis of sRNAs 11

2.1 Introduction 11

2.2 Comparison Between small RNA (sRNA) and MicroRNA (miRNA) 11

2.2.1 Data Collection 11

2.2.2 Feature Extraction 11

2.2.3 t-Test Results 12

2.2.4 Clustering by Single Linkage Algorithm 13

2.3 Comparison between Different Types of sRNAs 14

2.3.1 Comparison between Cis-encoded sRNA and Trans-encoded sRNA 14

2.2.2 Comparison between Cis-encoded sRNA and Protein Binding sRNA 15

2.2.3 Comparison between Trans-encoded sRNA and Protein Binding sRNA 17

2.4 Conclusions 18

3 sRNA Prediction 19

3.1 Introduction 19

3.2 Data Collection 21

3.3 Features Considered 21

3.3.1 Primary Nucleotide Features 21

3.3.2 Secondary Structure Features

3.4 Classification using Support Vector Machine (SVM) 23
3.5 Classification Results 24
3.6 Feature Selection and Classification 28
3.7 Conclusions 33

4 Target Identification of sRNAs 34
4.1 Introduction 34
4.2 Data Collection 35
4.3 Features Extraction 35
4.3.1 Binding Site Features 36
4.3.2 sRNA Secondary Structure Features 36
4.4 Classification Using SVM 36
4.5 Classification Results 36
4.6 Conclusions 38

5 Discussions and Conclusions 39
5.1 Discussions 39
5.2 Scope of Future Work 39

References 40

Chapter 1

Introduction

1.1 Basic Molecular Biology

The main actors in the chemistry of life are molecules are called proteins and nucleic acids. Proteins are responsible for what a living being is and does in a physical sense. Nucleic acids encode information necessary to produce proteins and are responsible for passing along this recipe to subsequent generations. A protein is a chain of simpler molecules called amino acid residues [1]. This sequence is known as its primary structure.

Every amino acid has one central carbon atom known as α-carbon. To the C_α (α-carbon) atoms are attached a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH) and a side chain. In nature, there are twenty different kinds of amino acids. Examples of amino acids are depicted in Fig 1.1.

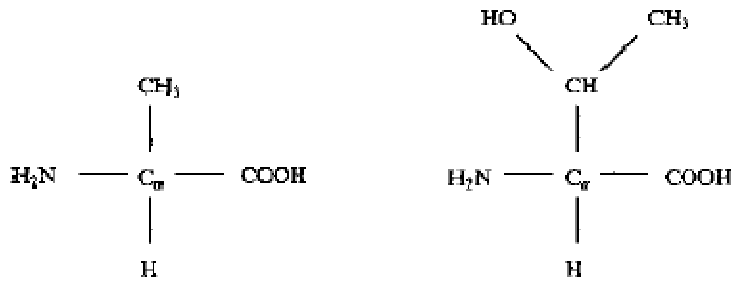


Fig 1.1 Examples of two amino acids: Alanine (left) and Threonine (right)

A protein is not just a linear sequence of residues. Proteins actually fold in three dimensions, presenting secondary, tertiary and quaternary structures. Proteins are produced in a cell structure called ribosome. In ribosome, the component amino acids of a protein are assembled one by one according to information contained in an important molecule called messenger ribonucleic acid [1].

Living organisms contain two kinds of nucleic acids namely ribonucleic acids (RNA) and deoxyribonucleic acids (DNA).

1.1.1 DNA

Like protein, DNA is a chain of simpler molecules. Actually it is a double chain which is called the strand. This strand has a backbone consisting of repetitions of the same basic unit. This unit is formed by a sugar molecule called 2'-deoxyribose attached to phosphate residue. Typical structure of 2'-deoxyribose is depicted in Fig 1.2. For details refer to [1]. The sugar molecule contains five carbon atoms, and they are labeled 1' through 5'. The bond that creates the backbone is between the 3' carbon and the 5' carbon of the next unit. This sequence is denoted in the order 5'-3'. The information necessary to build each protein or RNA found in an organism is encoded in the DNA molecules [1].

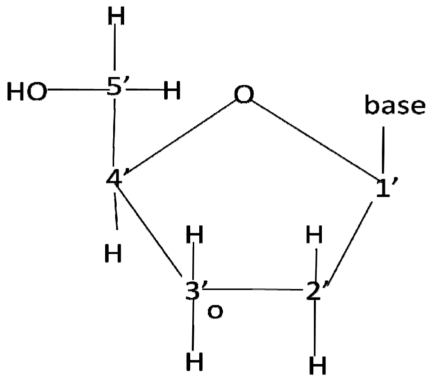


Fig 1.2 Structure of 2'-deoxyribose

1.1.2 Bases

Attached to each 1' carbon in the backbone are other molecules called bases. There are four kinds of bases, Adenine (A), Guanine (G), Cytosine(C), and Thymine (T). Bases 'A' and 'G' belong to a larger group of substances called purines, where as 'C' and 'T' belong to the pyrimidines [1].

Base A is always paired with base T, and C is always paired with G. Bases A and T are said to be a pair of complementary bases. Similarly C and G are complements to each other. These pairs are known as Watson-Crick base pairs. Base pairs provide the unit of length most used when referring to DNA molecules abbreviated to 'bp'. We consider DNA as a string of letters, each letter representing a base [1].

1.1.3 Nucleotides

When we see the basic unit of a DNA molecule as consisting of the sugar, the phosphate and its base, we call it nucleotide. DNA molecule having a few (tens of) nucleotides is referred to as an Oligonucleotide. DNA molecules in nature are very long, much longer than proteins. In human cell, DNA molecules are about three billion bp long. Typical double strand structure of DNA is given in Fig 1.3 [1].

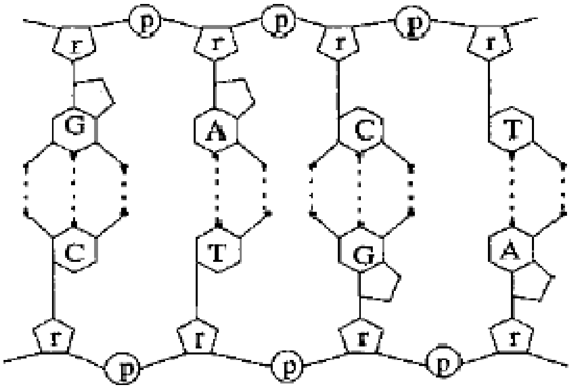


Fig 1.3 a schematic molecular structure view of a double strand of DNA

1.1.4 RNA

RNA molecules are much like DNA molecules with the following basic compositional and structural differences.

1. In RNA the sugar is ribose instead of 2'-deoxyribose. The figure for ribose is given in

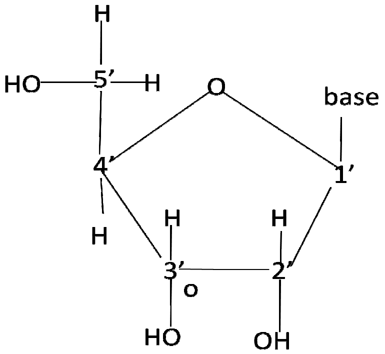


Fig 1.4 Structure of ribose

2. In RNA we do not find Thymine (T), instead, Uracil (U) is present. Uracil also binds with Adenine like Thymine does.

3. RNA does not form a double helix, it is single stranded structure.

There are several types of RNA present in the cell. These RNAs can be broadly categorized into coding (that are able to code for proteins, ex: messenger RNA (mRNA)) and non-coding (that are not able to code for proteins ex: small RNA (sRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), etc) [1].

1.1.5 Chromosome

Each cell of an organism has a few, very long, DNA molecules. Each such molecule is called a chromosome. Certain contiguous stretches along it encode information for building proteins but others do not. To each different kind of protein in an organism there usually corresponds one and only one contiguous stretch along the DNA. This stretch is known as a gene [1].

1.1.6 Gene

A gene is a continuous stretch of DNA that contains the information necessary to build a protein or an RNA molecule. In the case of human a gene may have something like 10000 base pairs (bps) [1].

1.1.7 Genome

The complete set of chromosomes inside a cell is called genome. The number of chromosomes in a genome is characteristic of a species. Prokaryotes generally have just one chromosome where as eukaryote chromosomes appear in pairs. The human genome consists of 23 chromosome pairs [1].

The two chromosomes that form a pair are called homologes and a gene in one member of pair corresponds to a gene in the other. Cells that carry only one member of each pair of chromosomes are called haploid. The haploid human genome consists of 3 billion DNA base pairs whereas most bacteria have a single circular chromosome that can range in size from only 160,000 base pairs to 12,200,000 base pairs [3, 4].

1.2 Central Dogma of Molecular Biology

The process of transcription followed by translation that results in the formation of a protein from a gene forms the basis of the central dogma of molecular biology. It can be stated as 'DNA makes RNA makes protein'**.** The typical flow in Central Dogma of Molecular Biology is depicted in Fig 1.5.

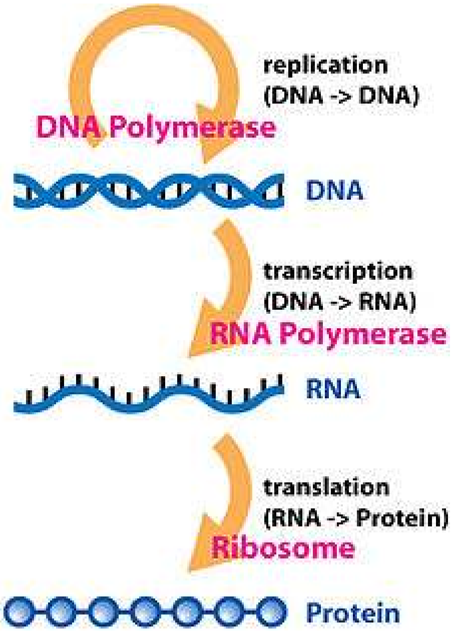


Fig 1.5 Typical flow in central dogma of molecular biology

1.2.1 DNA Replication

DNA replication is a biological process that makes copies of the DNA. It occurs in all living organisms and is the basis for biological inheritance. The process starts with one double-stranded DNA molecule and produces two identical copies of the molecule. Each strand of the original double-stranded DNA molecule serves as template for the production of the complementary strand [1].

1.2.2 Transcription

A cell mechanism recognizes the beginning of a gene or a gene cluster thanks to the promoter. Promoter is a region before each gene in the DNA that serves as an indication to the cellular mechanism that a gene is ahead. Having recognized the beginning of a gene or gene cluster, a copy of the gene is made as an RNA molecule. This resulting RNA is the messenger RNA (mRNA). mRNA will have exactly the same sequence as one of the strands of the gene but substituting ‘U’ for ‘T’. This process is called transcription [1].

In DNA , the strand that looks like the mRNA is called the antisense or coding strand , and the other one is the sense or anticoding or template strand. The template strand is the one that is actually transcribed, because the mRNA is composed by binding together ribonucleotides complementary to this strand. The process always builds mRNA molecules from their 5’ end to their 3’ end, where as the template strand is read from 3’ to 5’ [1].

Transcription as described is valid for organisms categorized as prokaryotes. These organisms have their DNA free in the cell, as they lack a nuclear membrane. Other organisms, categorized as eukaryotes, have a nucleus separated from the rest of cell by a nuclear membrane and their DNA is kept in the nucleus. In these organisms genetic transcription is much more complex. Many eukaryotic genes are composed of alternating parts called introns and exons [1].

After transcription, the introns are removed from the mRNA. This means that introns are part of a gene not used in protein synthesis. After introns are removed, the shorted mRNA, containing copies of only the spliced exons plus regulatory regions in the beginning and end leaves the nucleus. For details refer to [1].

The entire gene as found in the chromosome is called genomic DNA. The introns, exons, transcription start site, transcription end site and promoter region are depicted in Fig 1.6

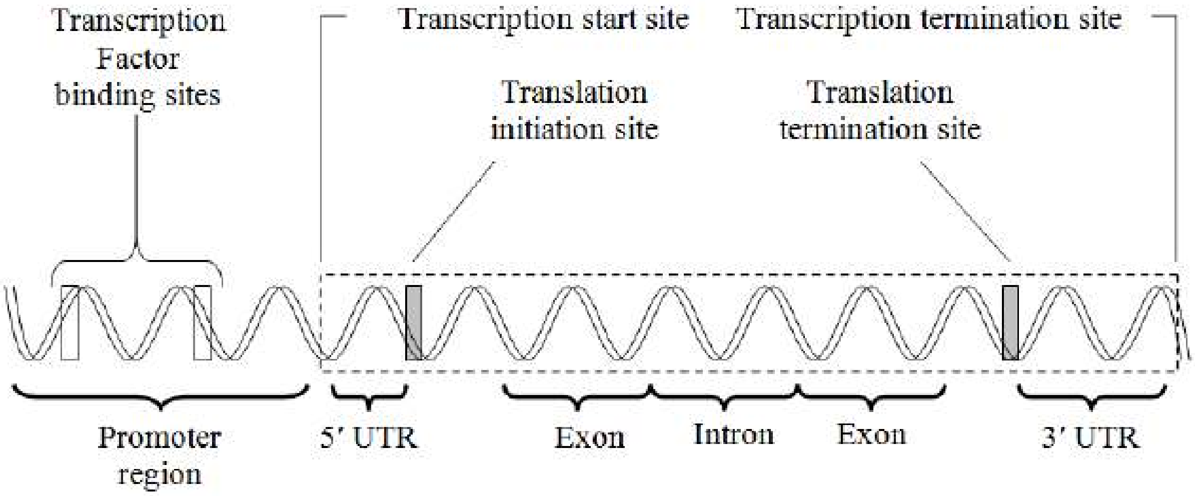

Transcription Factor binding sites
Transcription start site
Transcription termination site
Translation initiation site
Translation termination site
Promoter region
5′ UTR
Exon
Intron
Exon
3′ UTR

Fig 1.6 Structure of an eukaryotic gene

1.2.3 Translation

Protein synthesis takes place inside a cellular structure called Ribosome. Ribosomes are made of protein and a form of RNA called ribosomal RNA (rRNA). In molecular biology and genetics, translation is the third stage of protein biosynthesis. In translation, messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific chain of amino acid, or polypeptide, that will later fold into an active protein. As can be seen from Fig 1.6, there are two regions of the mRNA called 5′ and 3′ untranslated regions (UTRs). These appear at the 5′ and 3′ ends of the mRNA (the two ends of any sequence of nucleotides are distinguished as 5′ and 3′ end), respectively, and do not participate in translation [1].

Transfer RNAs (tRNAs) are the molecules that actually implement the genetic code in translation. They make the connection between a codon (a nucleotide triplet) and a specific amino acid this codon codes for. As the messenger RNA passes through the interior of the ribosome, a tRNA matching the current codon – the codon in the mRNA currently inside the ribosome, bind to it, bringing along the corresponding amino acid. For details refer to [1].

1.3 Introduction to Micro RNAs (miRNA)

In addition to mRNAs, transcription also produces several non-coding RNAs that do not code for any proteins. MicroRNAs (miRNAs) which are of length 22 nucleotides (nt) belong to this category. These single stranded small molecules bind to the 3′ untranslated region (UTR) of a specific target mRNA and regulate its expression at post-transcriptional level either by degrading the target mRNA or translationally repressing its protein production. As is well-known, 3′ UTR of mRNA may contain sequences that regulate translation efficiency, mRNA stability, and polyadenylation signals. Therefore when a miRNA binds to the 3′ UTR of its target mRNA, then these functionalities get affected, thereby repressing the production of the corresponding protein [2].

1.4 Introduction to small RNAs (sRNA)

Bacterial sRNAs are a class of non-coding RNAs found in bacteria with their lengths typically varying from 40 to 200 nucleotides (nt) which play a variety of important roles in many biological processes through binding to their mRNA or protein targets. They are involved in many biological processes, such as posttranscriptional regulation of gene expression, RNA processing, mRNA stability and translation and protein degradation etc. Furthermore, one main role of sRNA is to regulate gene expression through non-perfect complementary matches between sRNA and 5’ untranslated region (UTR) of its target mRNAs and most often closes the ribosome-binding site , i.e.,, they bind to the ribosome binding site [4].

For the few characterized sRNA-mRNA interactions, the inhibition of ribosome binding is the main contributor to reduce protein levels. However, sRNAs can also activate expression of their target mRNAs through an anti-antisense mechanism whereby base pairing of the sRNA disrupts an inhibitory secondary structure, which separates the ribosome-binding site [4].

Bacterial sRNA can be divided into two classes according to their mode of action. The first class binds to protein targets and thereby modifies the activity of the target protein, while the second class binds to the mRNA targets and regulates expression or stability of their target genes at the post-transcriptional level [4].

In contrast to the protein-binding sRNAs, most characterized sRNAs regulate gene expression by base pairing with mRNAs and fall into two broad classes: those having extensive potential for base pairing with their target RNA and those with more limited complementarity. The former are called cis-encoded sRNAs and later trans-encoded sRNAs [4].

Cis-encoded sRNAs are encoded in cis (nearby) on the DNA strand, opposite to the target RNA and share extended regions of complete complementarity with their targets, often 75 nt or more. Although the two transcripts are encoded in the same region of DNA, they are transcribed from opposite strands.

Another class of base pairing sRNAs is the trans-encoded sRNAs, which, in contrast to the cis-encoded antisense sRNA, share only limited complementarity with their target mRNAs. These sRNAs regulate the translation and stability of target mRNAs and are, in many respects, functionally analogous to eukaryotic miRNAs. The majority of the regulation by the known trans-encoded sRNAs is negative. Base pairing between the sRNA and its target mRNA usually leads to repression of protein levels through translational inhibition, mRNA degradation, or both [4].

For trans-encoded sRNAs there is little correlation between the chromosomal location of the sRNA gene and the target mRNA gene. In fact, each trans-encoded sRNA typically base pairs with multiple mRNAs. The capacity for multiple base pairing interactions results from the fact that trans-encoded sRNAs make more limited contacts with their target mRNAs in discontinuous patches, rather than extended stretches of perfect complementarity, as for cis-encoded antisense sRNAs. The region of potential base pairing between trans-encoded sRNAs and target mRNAs typically includes ~10-25 nts [4].

Due to divergence in functions, sequences and structures, there are no common identifiers for bacterial sRNAs. Even widely accepted characteristics, such as small size (<200 nt) and lacking coding capacity, do not always apply, as is demonstrated by the example of *Staphylococcus aureus* RNAIII, which clearly is a regulatory RNA, but is 514 nt in size and does code for a short peptide as well [5].

1.5 Scope of Dissertation

As already discussed sRNAs are regulators of mRNAs; either they stop the production of bacterial proteins or indirectly aid in their production. Several bacterial proteins are harmful for human beings. To stop the production of protein we should know the sRNA which represses the corresponding mRNA or the protein. For this we need to know the models for prediction of sRNA and their targets. This is the main focus of this work.

In this dissertation, the primary goal is to develop algorithms for analyzing bacterial sRNAs. In particular a method for predicting sRNA, given stretches of the bacterial genomes, is developed. Thereafter a technique for predicting the target mRNAs of a given bacterial sRNA is described. Results are provided to demonstrate the effectiveness of the proposed approach.

In Chapter 2, we compare the similarities between the miRNA and sRNA and also among different types of sRNAs. In Chapter 3, we develop three models for predicting sRNA. More precisely, if a sequence of nucleotides is given then the task is to decide whether it falls in the category of sRNA or not. In Chapter 4, we have developed a model for predicting targets of sRNA. In other words, if a pair of sRNA and mRNA is given, then the problem is to decide whether the mRNA is a possible target of the given sRNA or not. Finally, in Chapter 5, we discuss some issues and future work in the related field.

Chapter 2

Analysis of sRNAs

2.1 Introduction

In this chapter, we compare sRNA and microRNAs using some primary nucleotide features and then tried to cluster using single linkage clustering algorithm. Comparison among the different categories of sRNAs, as mentioned in Chapter 1, is also reported.

2.2 Comparison between small RNA (sRNA) and MicroRNA (miRNA)

As the functions of miRNA and sRNAs are similar in nature, here we tried to compare both of them by using several sequence level features. The data set used is described first, followed by the computation of the features and the comparative results.

2.2.1 Data collection

950 sRNAs are collected from sRNAMap [6], CD-HIT [7] is applied for reducing the redundancy among the sRNA sequences. We have considered two sequences to be similar if they have more than 90% matches after alignment. For all similar sequences, CD-HIT returns only one sequence. By applying CD-HIT, the number of sRNAs was reduced to 370. Among the sRNAs collected, minimum length of sRNA is 51 (dicF4), maximum 1496 (SokB) and on average the length is 156. 1582 miRNAs used in TargerMiner [2] are taken for comparison purpose.

2.2.2 Feature Extraction

To compare sRNA and miRNA total 88 primary nucleotide features are extracted. These are composed of

- Single nucleotide frequencies: 4 features
- Dinucleotide frequencies: 16 features
- Trinucleotide frequencies: 64 and
- four Quad-nucleotide frequencies “aaaa, cccc, gggg, ttttt”.

These features have been extracted for the 370 sRNAs and 1582 miRNAs. All the above features are normalized, that is divided by their sequence length.

2.2.3 t-Test Results

The average values of all the above mentioned features are plotted in Fig 2.1a. The figure shows that sRNA and miRNA are different in almost all the features. To find the significance in the difference, t-test has been applied. The t-test results shows that sRNA and miRNA are significantly different in 54 of the 88 features namely aa, ag, ca, cc, cg, cu, ga, gc, gu, ug, uu, aaa, aac, aag, aau, aca, acg, acu, agu, auu, caa, cac, ccc, ccg, cga, cgc, cgg, cgu, cua, cug, gaa, gac, gau, gca, gcc, gcg, gug, guu, uaa, uac, uag, ucg, ucu, ugc, ugg, ugu, uua, uuc, uug, uuu, aaaa, cccc, gggg, uuuu with 5% significance level. For other features p-values are greater than or equal to 0.08. For the first two significantly different features 'aa' and 'ag', the box plot is shown in Fig 2.1b. By box plot, it can be observed that sRNA and miRNAs are different in these features.

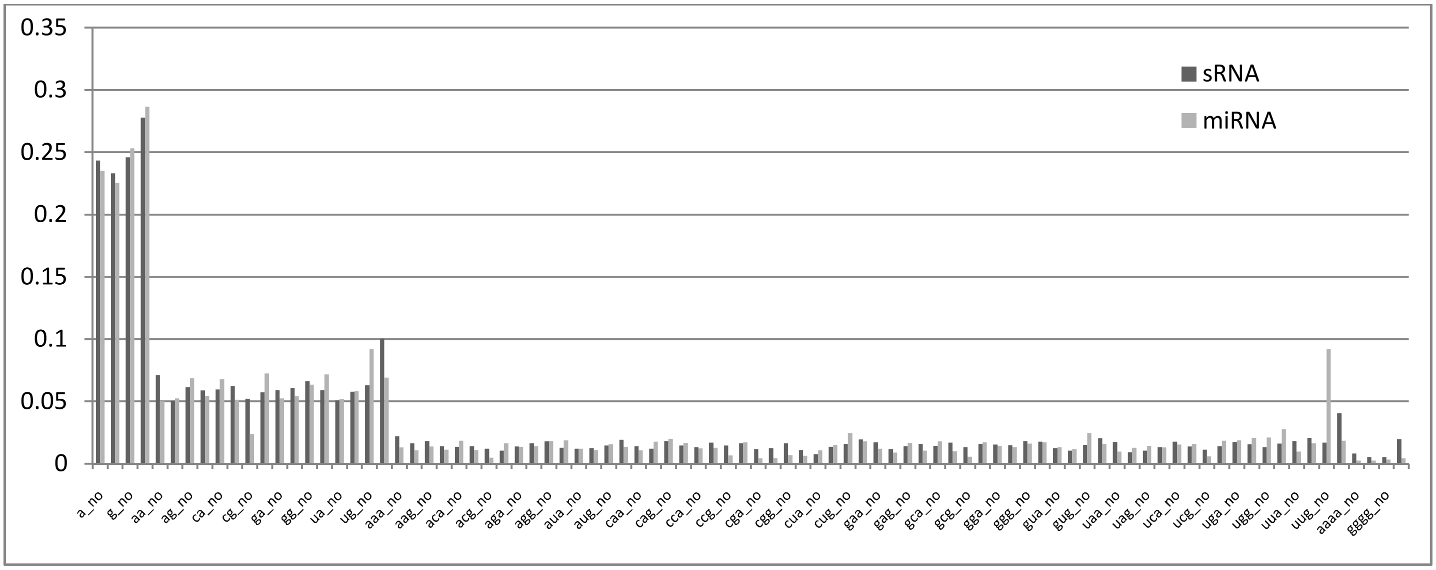


Fig 2.1a Average values of normalized 88 nucleotide features of miRNA and sRNA

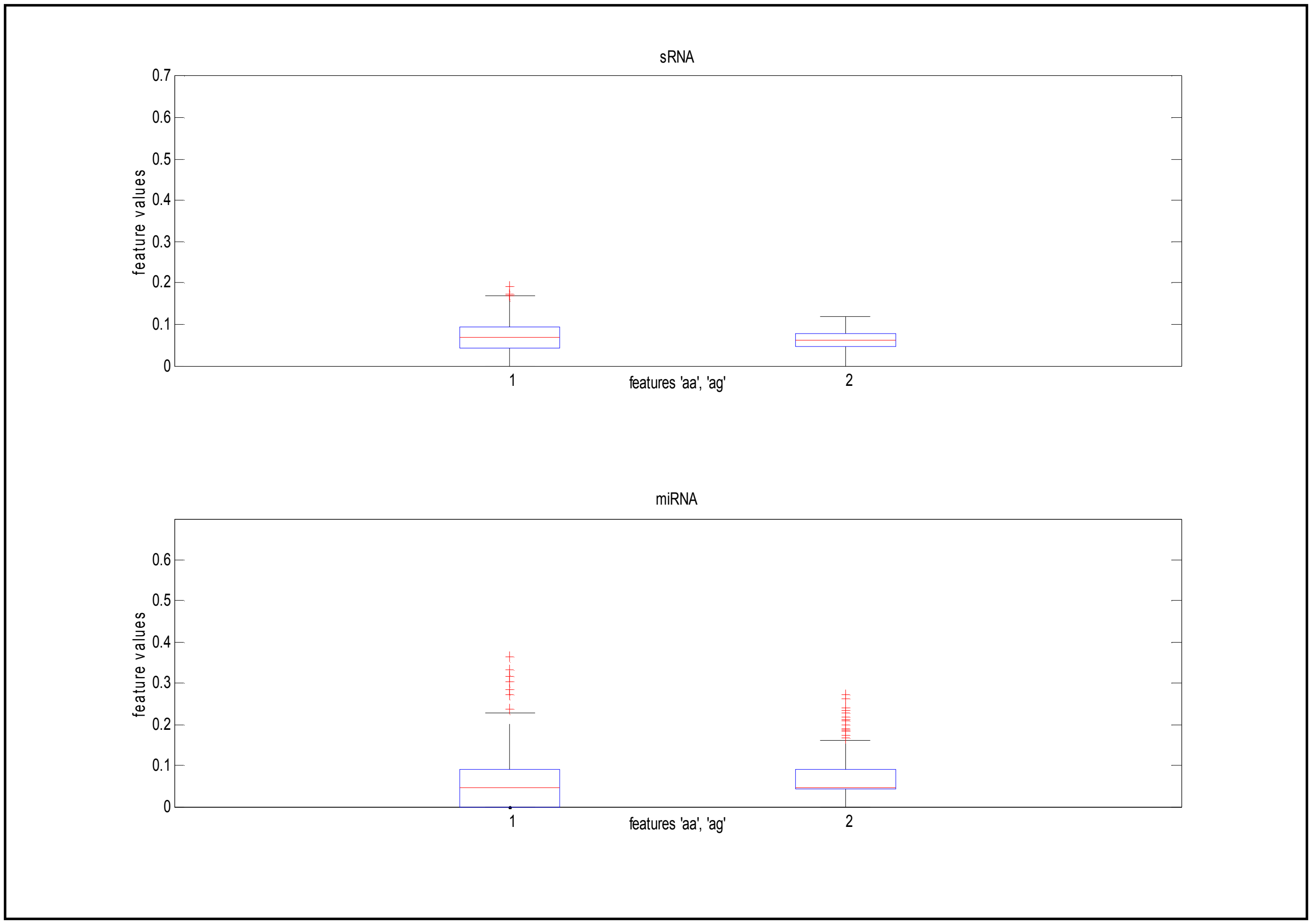


Fig 2.1b Box plot for the features ‘aa’ and ‘ag’ of sRNA and miRNA

2.2.4 Clustering by Single Linkage Algorithm

To find whether based on the features, it is possible to segregate the miRNAs and sRNAs, Single Linkage Clustering algorithm has been applied on the data set comprising 1952 sequences (370 sRNAs + 1582 miRNAs). Here similarity between two sequences is defined by the Euclidian distance between the above mentioned 88 features of the sequences. Initially each data point is considered to be a cluster of its own: that is, there are 1952 clusters each consisting of either a single sRNA or a single miRNA. In each iteration, the closest two clusters are merged, and the total number of clusters is reduced by one. Ideally, the algorithm should provide two clusters at the last but one level, one consisting of all sRNAs and the other consists of all miRNAs. It has been observed that the merging of the clusters is done properly up to 841 iterations (i.e., clusters with sRNA are merged into sRNA clusters and mRNA clusters are merged into mRNA clusters). At this stage a maximum cluster size of 360 of sRNAs is formed while the other clusters consisted of miRNAs. Only 10 sRNAs are merged with the other clusters. Thus out of 370 sRNAs, 360 sRNAs could be distinctly separated out from the miRNAs. This shows that sRNAs can be separated from miRNAs with 97% accuracy.

2.3 Comparison between Different Types of sRNAs

As already mentioned, according to the targets of sRNAs, they can be classified into two types namely mRNA binding sRNAs and protein binding sRNAs. mRNA binding sRNAs can be further categorized into two types cis-encoded, and trans-encoded sRNAs. In literature [8], there exist 9 cis-encoded, 120 trans-encoded and 9 protein binding sRNAs.

2.3.1 Comparison between Cis-encoded sRNA and Trans-encoded sRNA

For each sRNA, a total of 89 (88 previously mentioned features and one total length feature) nucleotide features extracted and average has been taken. For cis-encoded sRNA and trans-encoded sRNA the results are given in Fig 2.2a.

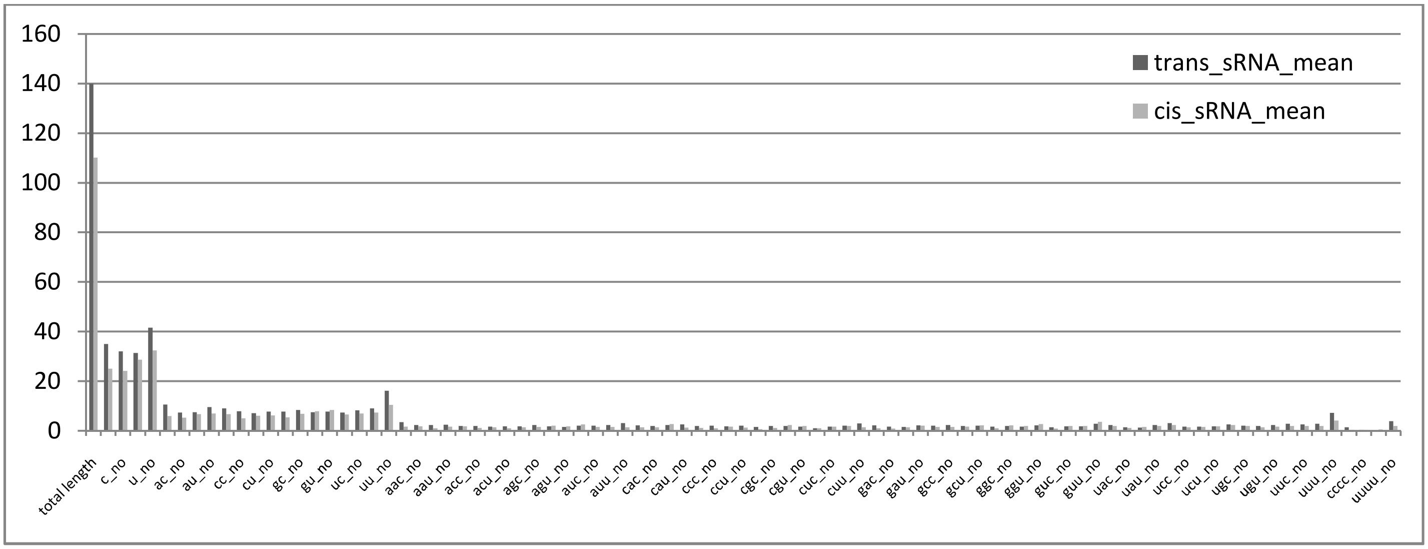


Fig 2.2a 89 nucleotide features of trans and cis-encoded sRNA

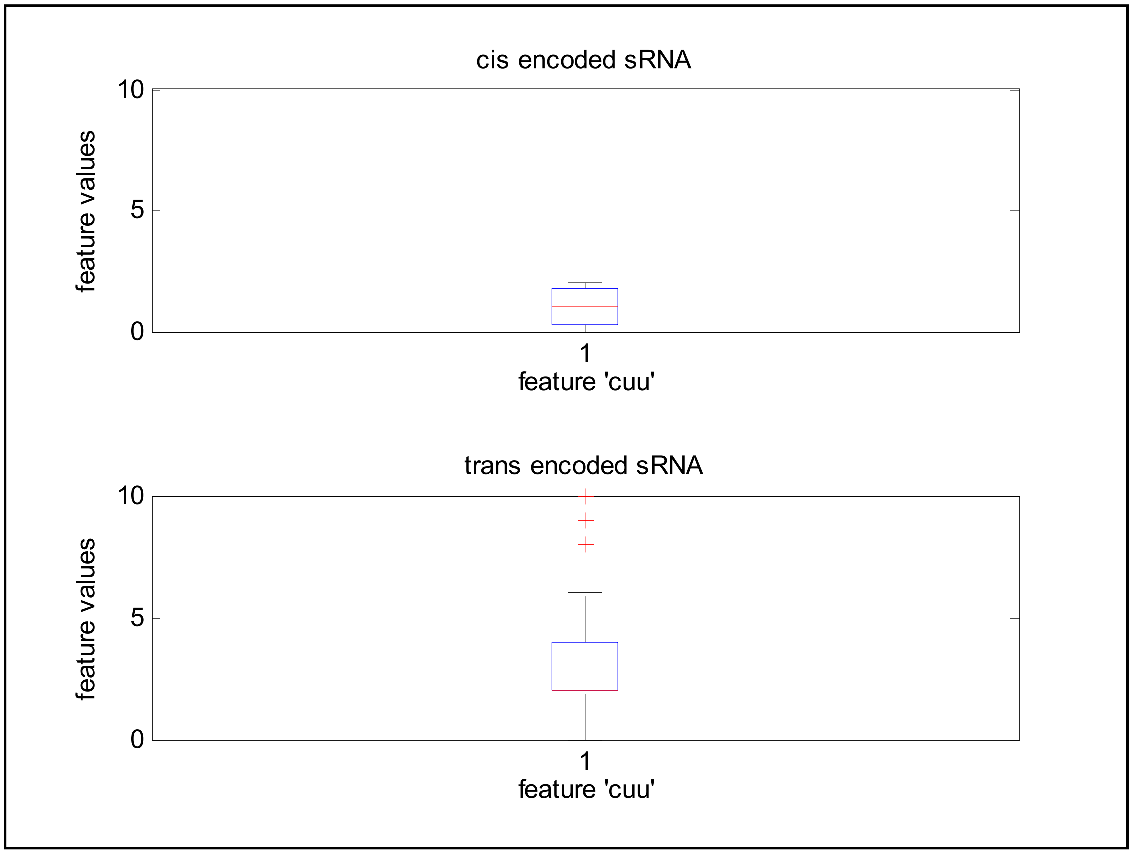


Fig 2.2b Box plot for feature ‘cuu’ of both cis and trans encoded sRNAs

For cis-encoded sRNAs usually binding site length is usually more than 70nt, where as for trans-encoded it varies from 5 to 50nt [4]. Though the above figure shows cis and trans-encoded sRNAs are differ in some features, by applying the t-test to above data it is observed that only in single feature ‘cuu’, they are significantly different with the significance level 5%. For all other features the p value is greater than or equal to 0.054. Box plot for the feature ‘cuu’ for both cis and tras-encoded sRNAs is shown in Fig 2.2b. From the figure it can be observed that they are different in this feature.

2.3.2 Comparison between Cis-encoded sRNA and Protein Binding sRNA

A similar comparison is conducted between cis-encoded sRNA and protein binding sRNA. The plot is given in Fig 2.3a. The t-test results shows that they are significantly different in 16 features namely g, aa, ga, gg, aaa, aag, agg, aug, cgc, gaa, gac, gga, ggg, uac, ugg, aaaa at the significance level of 5%. For other features the p-value in the t-test is at least 0.057. For the first significantly different features ‘g’ and ‘aa’ box plot is shown in Figure 2.3b. From the box plot it can be observed that cis and protein binding sRNAs are really different in these features.

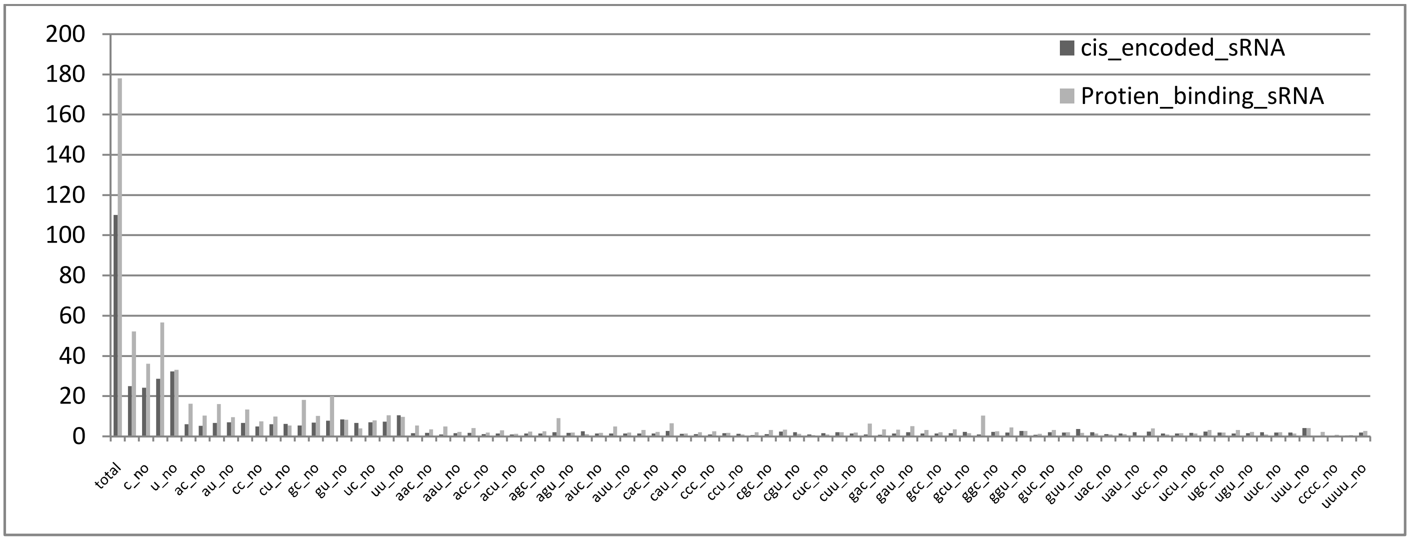


Fig 2.3a 89 nucleotide features of protein binding sRNA and cis-encoded sRNA

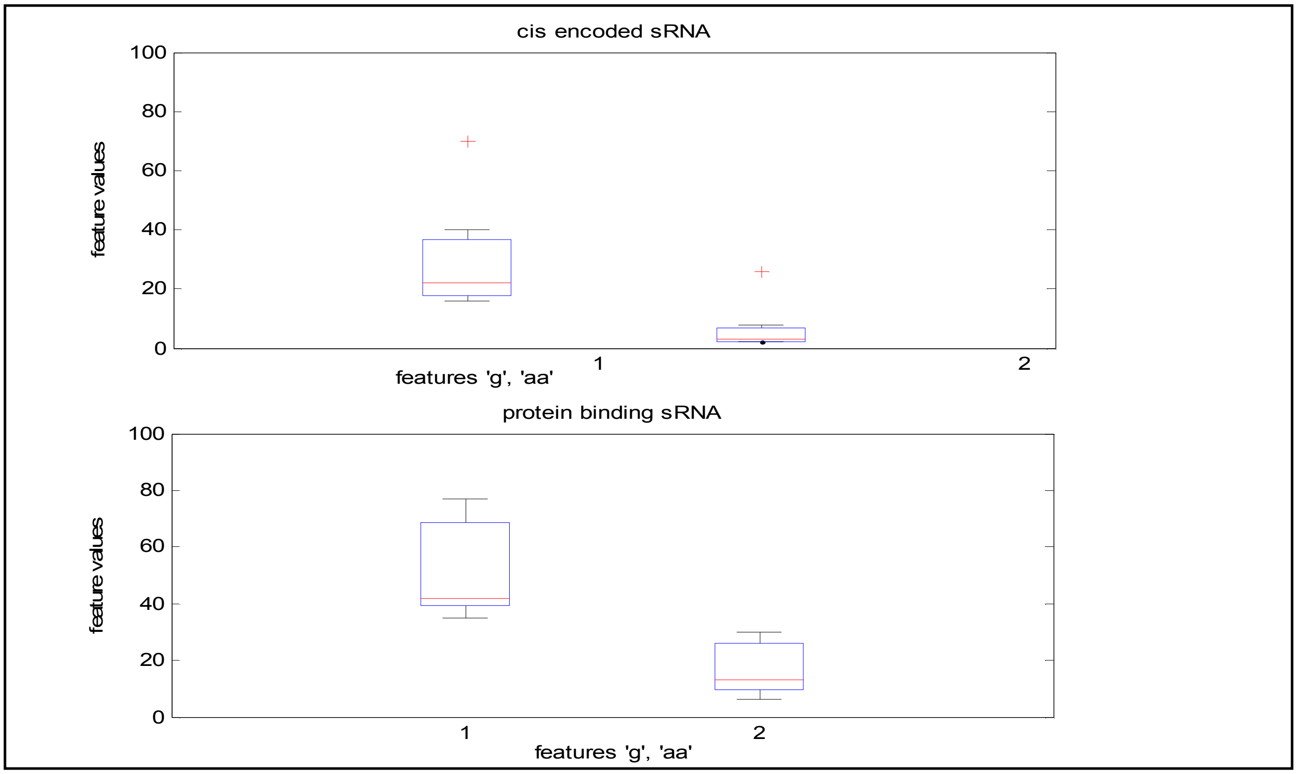


Fig 2.3b Box plot for features ‘g’ and ‘aa’ of both cis and protein binding sRNAs

2.3.3 Comparison between Trans-encoded sRNA and Protein Binding sRNA

Comparison between trans-encoded sRNA and protein binding sRNA are as given in Fig 2.4a. The t-test results show that they are significantly different in 19 features namely g, ag, ga, gg, aag, aca, acg, agg, aug, cag cgg, gaa, gac, gag, gau, gga, ggg, guc, uau with 5% significant. For the remaining features the p-value in t-test is at least 0.072. For the first significantly different features ‘g’ and ‘ag’ the box plot is shown in Fig 2.4b. From the figure it can be observed that trans-encoded and protein binding sRNAs are different in these features.

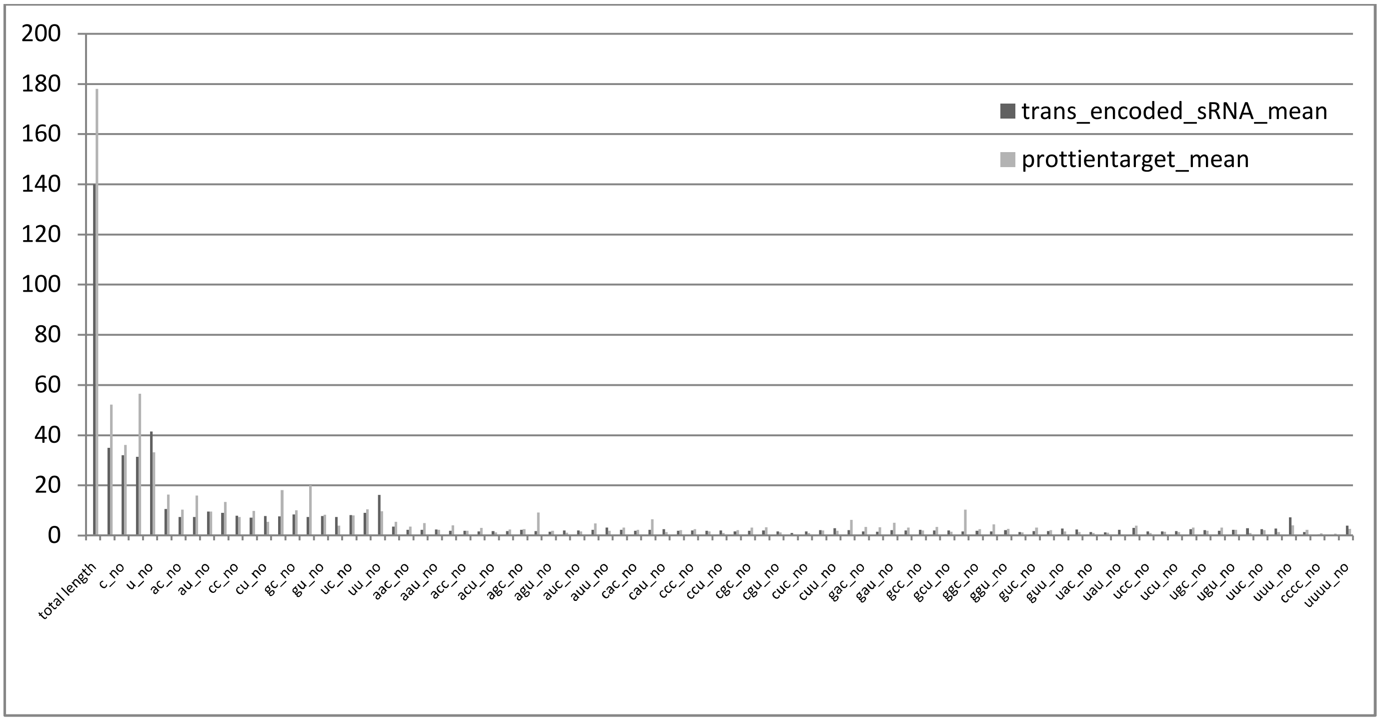


Fig 2.4.1 89 nucleotide features of trans-encoded sRNA and protein binding sRNA

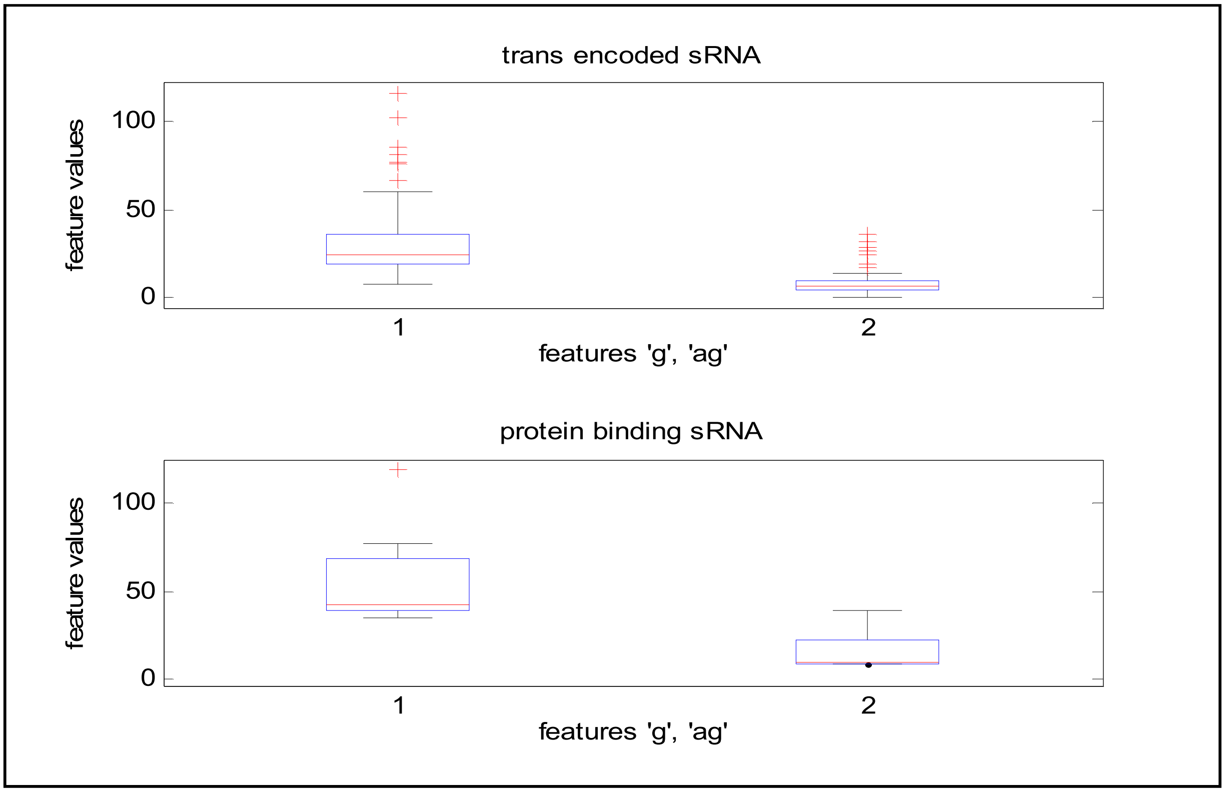


Fig 2.4.2 Box plot for features ‘g’ and ‘ag’ of both trans and protein binding sRNAs

2.4 Conclusions

In this chapter we tried to find the similarity between miRNA and sRNA and also analyzed the different types of sRNAs. For miRNAs and sRNAs we found that out of 88 features, in 54 features they are significantly different at the 5% significance level. We also verified that the single linkage clustering algorithm can separate them with 97% accuracy.

Consequently we believe that the prediction models existing of miRNAs cannot be applied for prediction of sRNAs and their targets. This would require development of techniques specific for sRNAs.

Chapter 3
sRNA Prediction

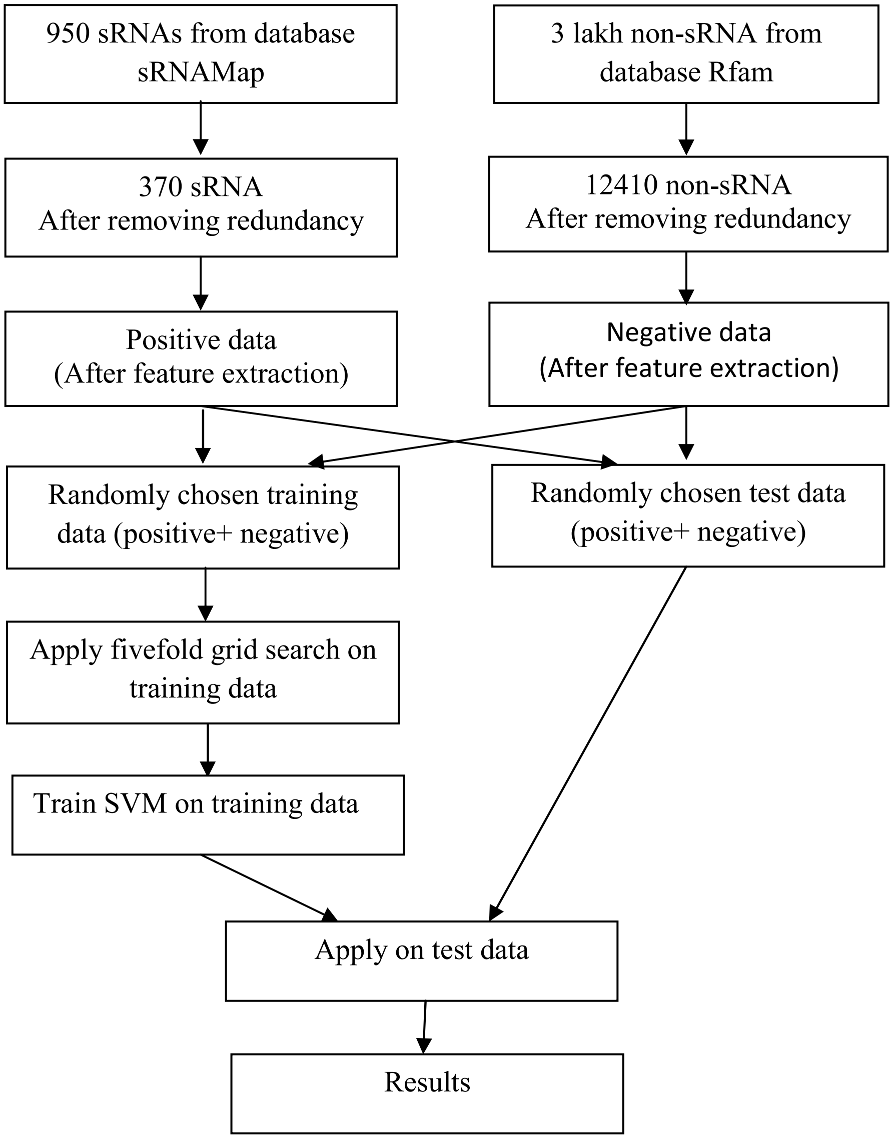
3.1 Introduction

As already mentioned in Chapter 1, bacterial sRNAs are a class of non-coding RNAs found in bacteria with their lengths varying from 40 to 200 which play a variety of important roles in many biological processes through binding to their mRNA or protein targets. They are involved in many biological processes, such as posttranscriptional regulation of gene expression, RNA processing, mRNA stability and translation, and protein degradation etc. The main role of sRNA is to regulate gene expression through non-perfect complementary matches between sRNA and 5' untranslated region (UTR) of its target mRNAs, which plays an important role in the interaction between bacterial and environments [4].

In view that sRNAs have tremendous heterogeneity in sizes, structures and functions, and are usually not translated into proteins (there are some exceptions), the methods combining bioinformatics prediction and experimental validation are often used to find new sRNAs. Therefore, it is very important to develop models for prediction of bacterial sRNAs.

Up to the present, several models [9-17] for prediction of bacterial sRNAs had been developed. These methods are generally classified into three categories, namely, comparative genomics-based methods, transcription units-based methods and machine learning-based prediction methods. Machine learning-based prediction models provide a general scheme for identification of sRNA genes in specificic bacterial genomes. However, the application of machine learning methods in detection of sRNAs for some particular bacterial genome is often limited because of the limit number of available validated sRNAs in that genome. For example, there are only about 81 sRNAs found in extensively studied E. coli (NCBI code: NC_000913.2). For other bacteria, the number of validated sRNAs is still smaller [9].

For addressing the above issue here we have developed a machine learning model for predicting the sRNAs in a more general sense. We do not restrict ourselves to specific bacterial species; rather we combine the available one and develop a more general bacterial sRNA prediction model. Moreover, we also include some new secondary structure features that have not been considered in past studies. For classification we have used support vector machine classifier. Here, initially we have developed three models by considering different ratios of the positive and negative data in the training. For the existing method [9] the accuracy is 64% for their test set. Though we are not comparing with their result (as the test set is different), as our method is giving good results on our test set, we believe that our work provides support for experimental identification of bacterial sRNA. The typical flow of sRNA prediction algorithm is given by the following flow chart.


950 sRNAs from database sRNAMap
3 lakh non-sRNA from database Rfam
370 sRNA After removing redundancy
12410 non-sRNA After removing redundancy
Positive data (After feature extraction)
Negative data (After feature extraction)
Randomly chosen training data (positive+ negative)
Randomly chosen test data (positive+ negative)
Apply fivefold grid search on training data
Train SVM on training data
Apply on test data
Results

Thereafter we have used the simple F-score measure to select the top few features in the three models. This results in simpler models with a small decrease in the classification performance, but the time is reduced by almost half. The details are provided below.

3.2 Data Collection

Total 950 sRNAs are collected from the database sRNAMap [6]. CD-HIT [7] is applied on the data to reduce 90% of the redundancy by which the number of sRNAs are reduced to 370. Theses sequences are taken as positive data. Total approximately three lakh bacterial non sRNAs are taken from Rfam [18]. We again applied CD-HIT to reduce 90% of the redundancy. The total number of non-sRNAs is reduced to 12410. This set is taken as negative data for the classifier. Both the data sets are taken in FASTA format.

3.3 Features Considered

Two types of features are extracted from both the positive and negative data, namely, primary nucleotide features and RNA secondary structure features. Moreover these secondary features have not been considered in past studies.

3.3.1 Primary Nucleotide Features

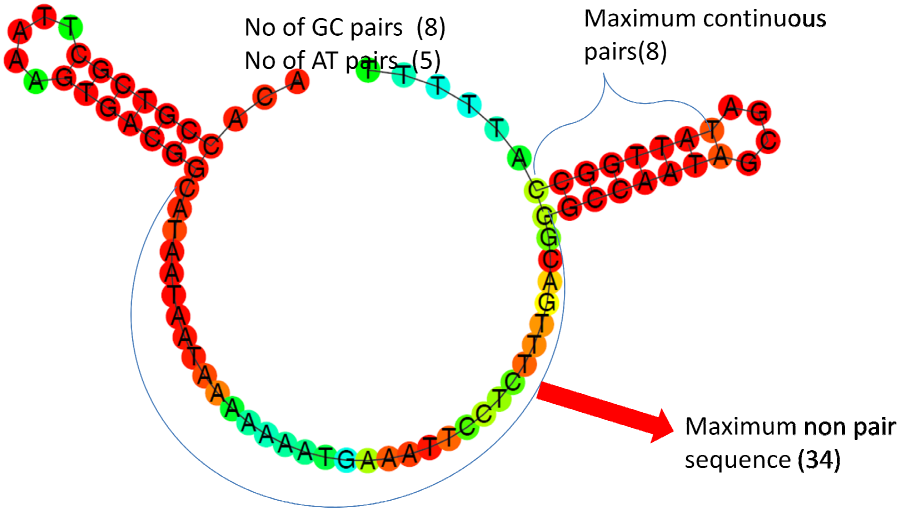
As discussed in Chapter 2, four single nucleotide features, sixteen di-nucleotide features, sixty four tri-nucleotide features and four Quad-nucleotide features namely aaaa, cccc, gggg, uuuu are extracted. In addition the G+C count and A+T count features are also computed. Thus we have a total of 90 nucleotide features.

3.3.2 Secondary Structure Features

To extract secondary structure features, secondary structure of RNAs are obtained by RNAfold [20]. In total six secondary features are considered namely,

1. Number of CG bonds (pairs) in the structure.
2. Number of AT bonds (pairs) in the structure.
3. Length of maximum continuous bonds
4. Length of maximum non pair sequence (non-bonding in between).
5. Average continuous bonds
6. Average continuous non bonds

The typical secondary structure of RNA obtained from the Vienna RNA package is as shown in the Fig 3.1. The first four secondary structure features considered are also shown in the figure.



ChipX sRNA predicted Secondary Structure

Fig 3.1 Typical secondary structure of a sRNA obtained from Vienna Package

3.4 Classification using Support Vector Machine (SVM)

Support vector machine (SVM) a promising tool for data classification. Its basic idea is to map data into a high dimensional space and find a separating hyper plane with the maximal margin. Given training vectors $x_k \in R^n$, $k = 1, 2 \ldots m$ in two classes, and a vector of labels $y \in R^m$ such that $y_k \in \{1, -1\}$, SVM solves a quadratic optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C\sum_{k=1}^{m} \xi_k \ , \\ \text{subject to} \quad & y_k(w^T \phi(x_k) + b) \geq 1 - \xi_k, \\ & \xi_k \geq 0, k = 1, \ldots, m, \end{aligned} \qquad (3.1)$$

where training data are mapped to a higher dimensional space by the function ϕ, and C is a penalty parameter on the training error, w is the coefficient vector of the hyper plane to be determined and ξ_k $k=1,2,\ldots,m$ are the slack variables which measures the degree of misclassification of point x_k. For any testing instance x, the decision function (predictor) is

$$f(x) = \operatorname{sgn}\left(w^T \phi(x) + b\right) \qquad (3.2)$$

Practically, we need only $k(x, x') = \phi(x)^T \phi(x')$ the kernel function, to train the SVM. The RBF kernel is used in our experiments:

$$k(x, x') = \exp(-\gamma\|x - x'\|^2) \qquad (3.3)$$

With the RBF kernel (3), there are two parameters to be determined in the SVM model: C and γ.

For applying SVM we have used LIBSVM. LIBSVM – a library for Support Vector Machine is software for support vector classification, regression and distribution estimation [19]. To get good generalization ability, we conduct a grid search and cross validation process to decide parameters. To apply this software the above considered nucleotide features and secondary structures features are combined and converted into the following format.

<class index> <feature label: feature value> <feature label: feature value> …<feature label: feature value>

3.5 Classification Results

After the features are extracted, randomly chosen 200 sRNAs and 2000 nonsRNAs (1:10 ratio) are taken as the training data, i.e., training data consists of 2200 patterns (200 from positive data and 2000 from negative data). The remaining data (170 sRNAs, 10410 nonsRNAs) is taken as the test data. Optimum values for SVM parameters (C and γ in equation (3.1) and (3.3)) are chosen by a grid search with a 5-fold cross validation on training dataset. After the training model is created LIBSVM is applied on both training data and test data for classification. The above process is repeated five times, each time using randomly chosen training data. In each case sensitivity and specificity are calculated. These are defined as follows:

Sensitivity = (True positive/ (True positive + False negative))
Specificity = (True negative/ (True negative + False positive)).

Here 'True positive' is the number of patterns belongs to positive class and classified as positive. Similarly 'True negative' is the number of negative patterns classified as negatives. 'False positive' is the number of negatives patterns classified as positive and 'False negative' is the number of positive patterns classified as negative. The ROC curve also been plotted. The classification results are given in Table 3.1.

For every case, for training data, only one or two patterns are misclassified, and for test data, on average 97.96% accuracy, 60.58% sensitivity and 98.57% specificity have been obtained.

In order to investigate the effectiveness of using the secondary structure features, the same experiment was carried out with only the nucleotide features (i.e., without the secondary structure features). The following results were obtained: accuracy 98%, sensitivity 48%, and specificity 99%. As can be seen, while the accuracy and specificity did not change much, the sensitivity reduced drastically by about 12%. This underlines the importance of using the secondary structure features.

Run	Accuracy	Sensitivity	Specificity
1	98.156898	62.941176	98.731988
2	98.421547	53.529412	99.154659
3	97.561440	67.647059	98.049952
4	97.930054	59.411765	98.559078
5	97.759926	59.411765	98.386167
AVG	97.965973	60.5882354	98.5763688

Table 3.1: Classification results on test data with 200 sRNA and 2000 non sRNA for training

In another scenario, we considered 1:2 ratio of the positive and negative samples in training data, that is 200 sRNA(positive data) 400 nonsRNAs (negative data). Here the remaining data (i.e., 170 sRNAs and 12010 non sRNAs) is taken as the test data. The above process is executed for five times. The classification results are given in Table 3.2. On average the results are 75.65% sensitivity, 93.38% specificity and 93.25% accuracy.

Run	accuracy	Sensitivity	Specificity
1	92.639467	72.41176	92.364532
2	89.129723	82.941176	89.217319
3	94.646965	74.705882	94.929226
4	94.868637	70.058824	95.212323
5	94.942528	77.058824	95.195670
AVG	93.245464	75.647586	93.383814

Table 3.2: Classification results on test data with 200 sRNA and 400 non sRNA for training

By increasing the positive data in training set (i.e., increasing the sRNAs) to 300 and taking 600 non sRNAs, the same procedure has been repeated. Here the test data is the remaining sRNAs (70) and remaining non sRNAs (9810). The classification results are shown in Table 3.3. On average 93.50% accuracy, 81.71% sensitivity and 93.50% specificity are obtained. The corresponding ROC plots are shown in Fig 3.2, Fig 3.3, and Fig 3.4. The curves also show area under the curve (AUC). The larger the AUC the better is the performance. The curves show that the minimum AUC is 0.84 for testing data which shows that our classifier is giving good results.

Run	accuracy	Sensitivity	Specificity
1	95.479797	78.571429	95.580017
2	95.454544	78.571429	95.554615
3	92.180138	88.571429	92.201524
4	90.850166	81.428571	90.906012
5	93.560608	81.428571	93.3632515
AVG	93.5050506	81.7142858	93.5749366

Table3.3: Classification results on test data with 300 sRNA and 600 non sRNA for training

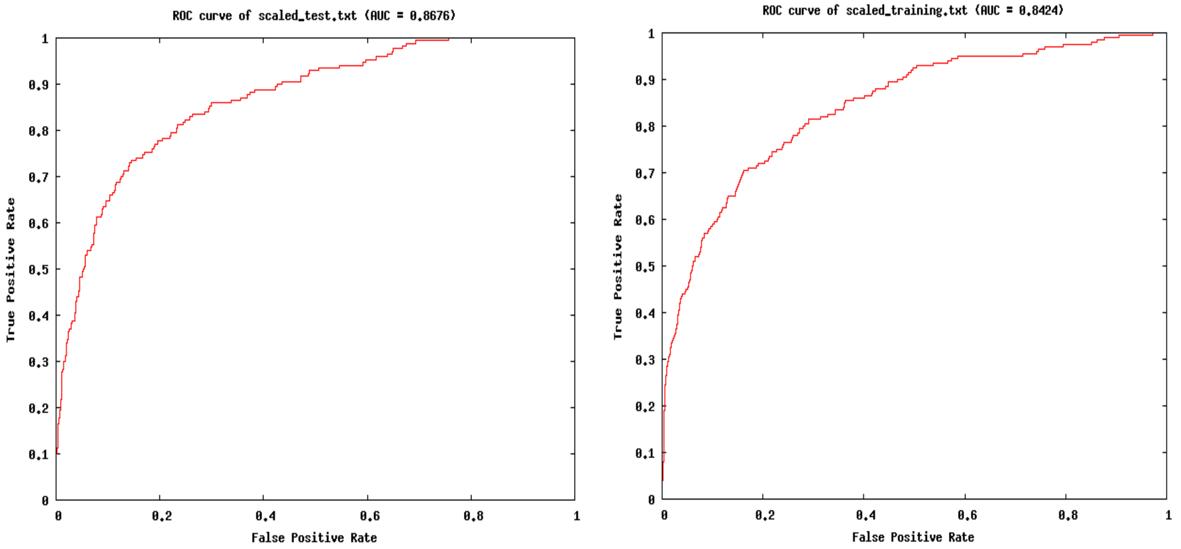


Fig 3.2: ROC curves with 200 sRNA and 2000 non sRNA for training

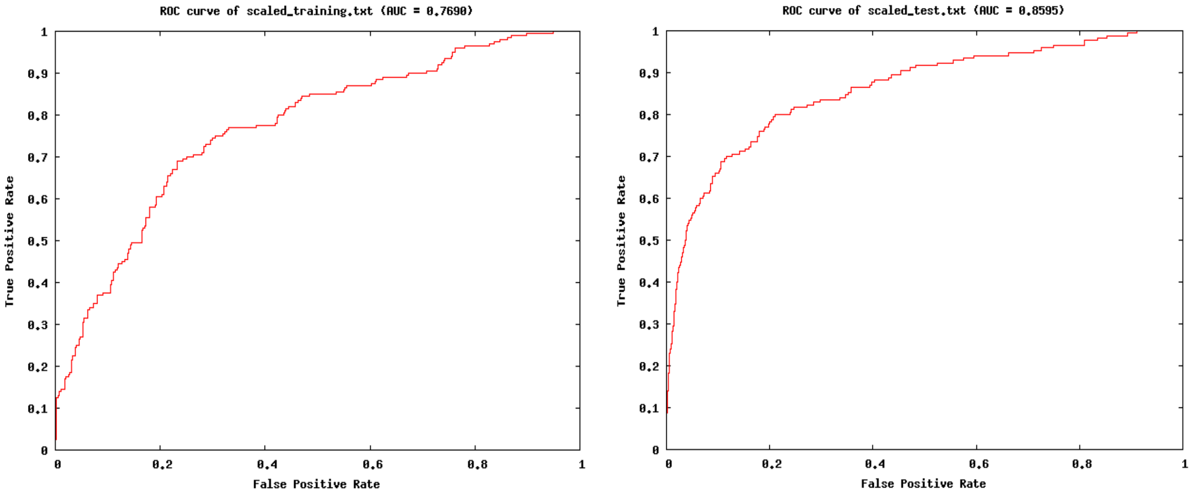


Fig 3.3: ROC curves with 200 sRNA and 400 non sRNA for training

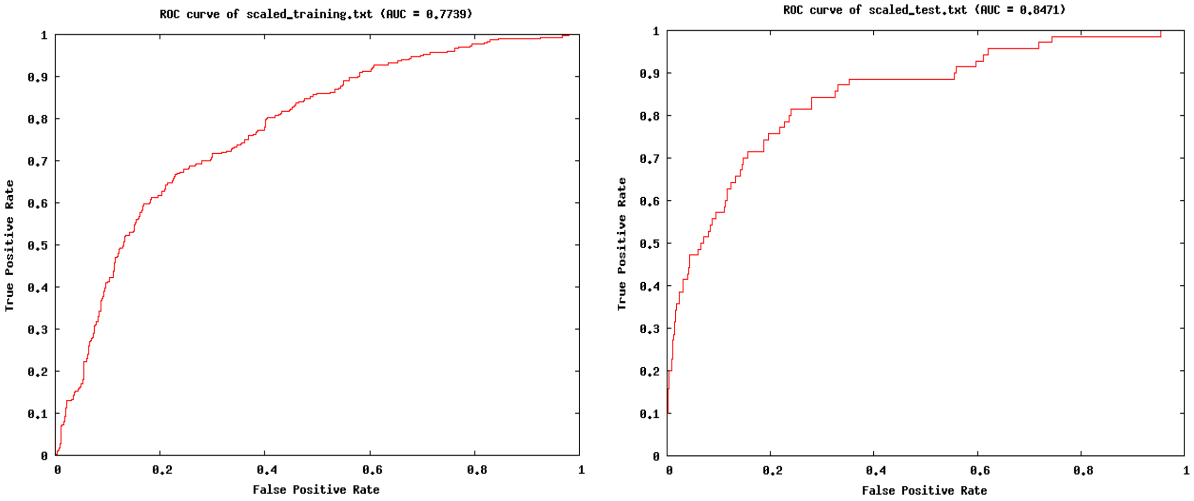


Fig 3.4: ROC curves with 200 sRNA and 2000 non sRNA for training

3.6 Feature Selection and Classification

In general, selection of a more informative subset of features leads to a simpler model that often results in a better performance. Here, we have simply computed the *F*-score of each feature as follows: For a set of training vectors x_k , *k*=1, 2,…, *m*, if the number of positive and negative instances are *n*+ and *n*−, respectively, then the *F*-score of the *i*-th feature is defined as

$$F(i) = \frac{\left(\bar{x}_i{}^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i{}^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i{}^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i{}^{(-)}\right)^2} \qquad (4)$$

To make a simpler model LIBSVM feature selection algorithm (F-score (4)) is applied and top 50% features having maximum F-score are considered. Plot for F-scores is given in Fig 3.5 and the F-score values are given in Table 3.4. The secondary structures are highlighted in the table. With these features and the above three cases were repeated. The results are shown in the Table 3.5, Table 3.6, Table 3.7 respectively. The ROC curves in all above three cases are given in Fig 3.6 through Fig 3.8. Though the results show that there is decrease in performance in terms of sensitivity, specificity and accuracy it improves in terms of computation and time. The times in different cases are shown in Table 3.8. From the Table 3.8, it can be observed that the time is reduced by almost half. It can be observed that AUC of the curves are not changed much which shows that performance is not changed significantly.

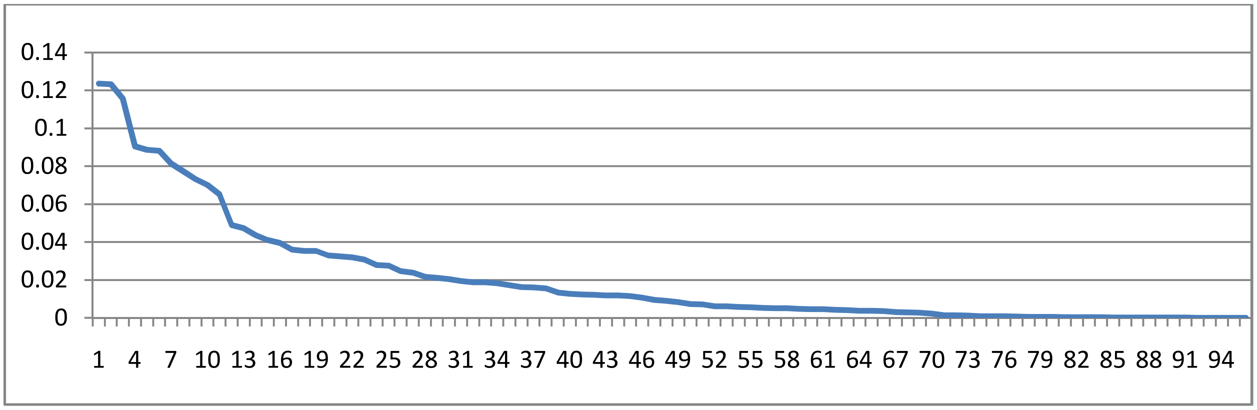


Fig 3.5: plot showing F-score values in descending order

No	Feature no	F-score	No	Feature no	F-score	No	Feature no	F-score
1	20	0.123527	33	81	0.018715	65	49	0.00372
2	82	0.123225	34	79	0.018215	66	44	0.00357
3	4	0.115518	35	58	0.017258	67	51	0.00312
4	88	0.090442	36	62	0.016253	68	67	0.00278
5	84	0.088621	37	59	0.016007	69	17	0.00272
6	3	0.088144	38	72	0.0156	70	56	0.00222
7	73	0.081283	39	25	0.013329	71	35	0.00133
8	53	0.077299	40	14	0.012623	72	30	0.00128
9	13	0.073108	41	5	0.012272	73	65	0.00116
10	52	0.070064	42	39	0.012181	74	**95**	**0.00088**
11	18	0.06509	43	46	0.01189	75	2	0.00085
12	43	0.04888	44	31	0.011813	76	75	0.00082
13	29	0.047301	45	**92**	**0.011482**	77	77	0.00062
14	36	0.043539	46	45	0.010568	78	16	0.00051
15	15	0.040975	47	63	0.009516	79	87	0.00044
16	40	0.039535	48	50	0.009026	80	37	0.00043
17	55	0.035914	49	1	0.008237	81	41	0.00031
18	90	0.035266	50	66	0.007245	82	70	0.00031
19	89	0.035207	51	34	0.007033	83	74	0.00029
20	7	0.03288	52	26	0.006148	84	**96**	**0.00027**
21	9	0.03242	53	54	0.006023	85	28	0.00022
22	11	0.031895	54	71	0.005687	86	33	0.00016
23	23	0.030658	55	10	0.005611	87	57	0.00012
24	**94**	**0.02777**	56	60	0.005171	88	6	0.00010
25	76	0.027546	57	69	0.005085	89	78	0.00009
26	61	0.024572	58	**93**	**0.005016**	90	42	0.00009
27	12	0.023864	59	64	0.004737	91	24	0.00009
28	8	0.02154	60	38	0.004543	92	22	0.00006
29	47	0.021111	61	83	0.004494	93	32	0.00002
30	68	0.02041	62	21	0.004232	94	19	0.00001
31	**91**	**0.019427**	63	85	0.004122	95	27	0.00001
32	80	0.01875	64	48	0.003727	96	86	0

Table 3.4 F-score values for the features

Run	Accuracy	Sensitivity	Specificity
1	98.336487	55.882353	99.029779
2	97.618149	56.470588	98.290106
3	97.533081	56.470588	98.203650
4	98.610588	58235294	99.269933
5	98.478264	57.058824	99.154659
AVG	98.1153138	56.8235294	98.7896254

Table 3.5: classification results on test data with 47 features with 200 sRNA and 2000 non sRNA for training

Run	Accuracy	Sensitivity	Specificity
1	93.530380	74.117647	93.805162
2	88.275864	84.117647	88.334721
3	89.499176	65.882353	89.833472
4	91.551727	75.294118	91.781848
5	91.584564	78.235294	91.773522
AVG	90.8883422	75.5294118	91.105745

Table 3.6: classification results on test data with 47 features with 200 sRNA and 400 non sRNA for training

Run	Accuracy	sensitivity	Specificity
1	92.634682	75.714286	92.734970
2	92.281143	81.428571	92.345470
3	97.222221	70.00000	97.383573
4	90.639732	88.571429	90.651990
5	94.511787	67.142.857	94.674005
AVG	93.457913	76.5714286	93.5582016

Table 3.7: classification results on test data with 47 features with 300 sRNA and 600 non sRNA for training

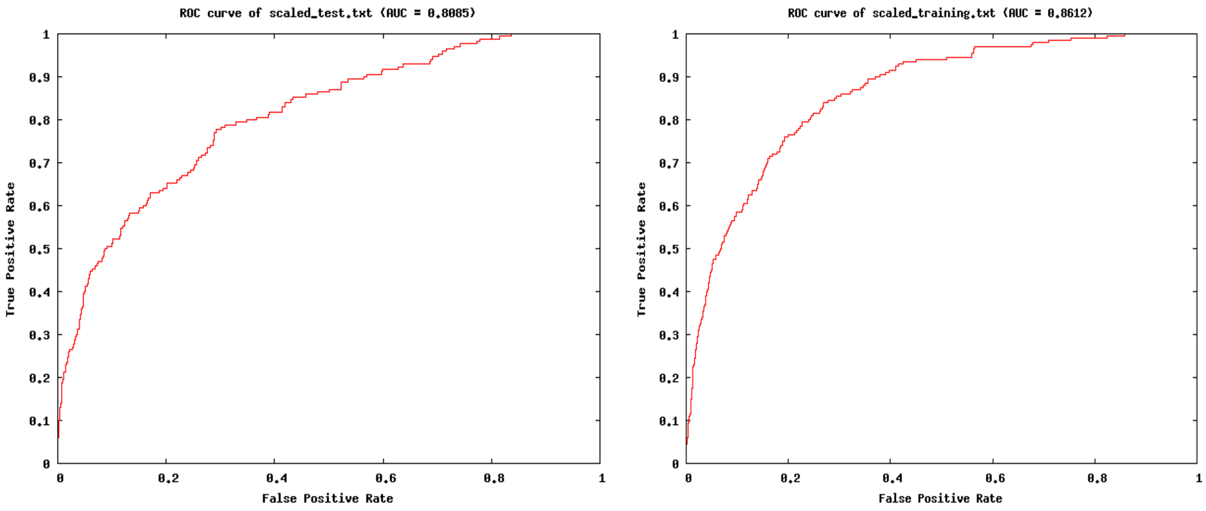


Fig 3.5: ROC curves with 200 sRNA and 2000 non sRNA for training with 47 features

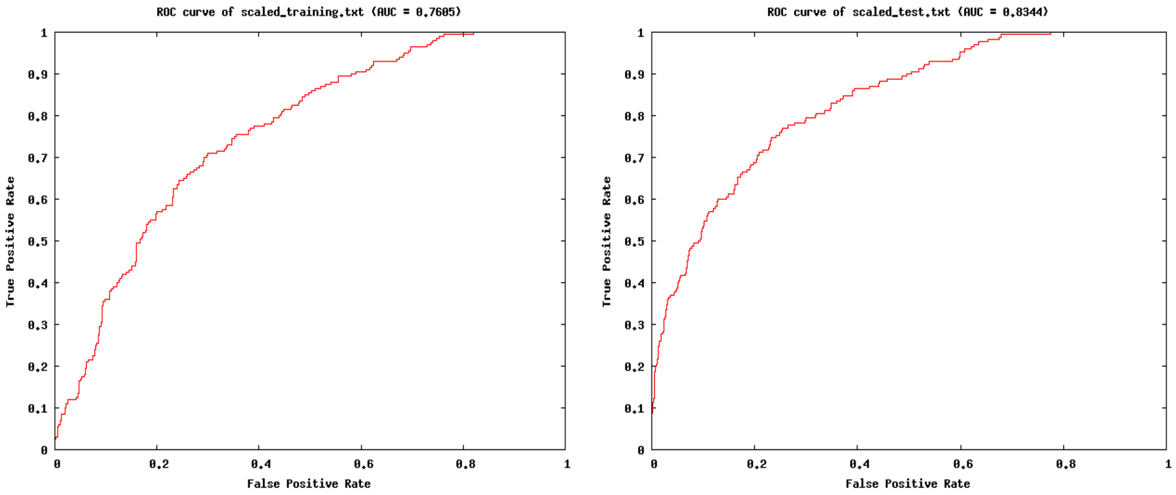


Fig 3.6: curves ROC curves with 200 sRNA and 400 non sRNA for training with 47 features

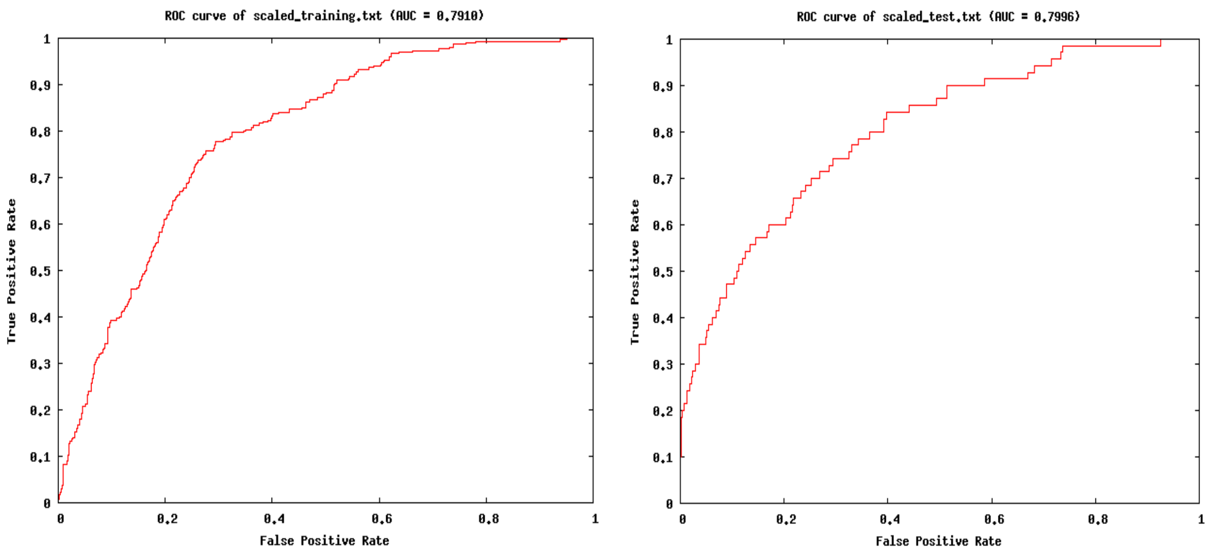


Fig 3.7: ROC curves with 300 sRNA and 600 non sRNA for training with 47 features

	With 96 features			With 50% of feaures		
	Real	User	System	Real	User	System
Grid search	2m 43.372s	1m 33.395s	0m 1.557s	1m 42.224s	0m 44.303s	0m 1.108s
Training	0m 0.309s	0m 0.295s	0m0.012s	0m0.150s	0m 0.145s	0m 0.006
Testing	0m 14.586s	0m 4.388s	0m0.039	0m 7.492s	0m 1.608s	0m 0.023s

Table 3.8 Different times for different functions with different features

3.7 Conclusions

In this chapter we have developed three models for sRNA prediction using support vector machine classifier. Several nucleotide based features as well as some novel ones based on RNA secondary structure were used for this purpose. Initially we applied SVM for the data consisting of 90 primary nucleotide features and later secondary structure features. By adding the secondary structure features the sensitivity is increased by 12% which shows the importance of secondary structure features considered. We have considered sRNAs and non-sRNA from different genomes. We got good accuracy on our test sets. Although we cannot compare our results with the existing ones (as the test sets are different), as we got good results on our test sets, and these tests are randomly taken,we believe that our work provides support for experimental identification of bacterial sRNA. For creating simpler models we have applied F-score feature selection algorithm and considered the top 50% features with maximum F-score. It is found that while the models become simple, the classification performance is not changed significantly. Time for grid search, training and testing with 96 features and the top 50% of the features with maximum F-score is shown in Table 3.8. It can be observed that with the 50% of top features, the time is reduced by almost half.

Chapter 4
Target Identification of sRNAs

4.1 Introduction

In this chapter we describe a model for predicting the possible targets of sRNA. The problem is as follows: given a pair of sRNA and mRNA, predict whether the sRNA is targeting the given mRNA or not. As already discussed in earlier chapters, sRNAs are non-coding RNAs in bacteria typically 40-200 nt in length. Most of them function as posttranscriptional regulators of gene expression through binding to the translation initiation region (TIR) of their target mRNAs. In positive regulation, the binding region is generally located 50-150nt upstream of the start codon, and the role is to activate expression of target genes. In negative regulation, the binding region is near the Shine-Dalgarno (SD sequence) sequence, and the role is to block ribosome binding [22]. SD sequence is a ribosomal binding site in the mRNA, generally located 8 base pairs upstream of the start codon AUG. The Shine-Dalgarno sequence exists only in prokaryotes. The six-base consensus sequence is AGGAGG; in *E. coli*, for example, the sequence is AGGAGGU. This sequence helps recruit the ribosome to the mRNA to initiate protein synthesis by aligning it with the start codon.

Up to now, there are various models existing for sRNA target prediction [22-29]. Most of the methods consider a very few data set for training and testing. Some of them considered entire mRNA sequence for alignment purpose which is unnecessary as sRNA mostly binding to the TIR region. Here, we have developed a model by considering only the 200nt upstream region to 100nt downstream region of start codon AUG of mRNA. For this we have utilized a large set of training and test data from [8] and consider our own features.

4.2 Data Collection

In [8], 390 pairs of sRNA and mRNAs are provided which are biologically verified for targeting activity. Among them some are found to be protein binding pairs, some are mRNA binding pairs and the remaining are no interaction pairs. Here from the interacting pairs, we have removed all protein binding pairs and also some of mRNA binding pairs for which the sRNA binding site is not mentioned. Thereafter, we have considered 95 pairs of sequences for which the exact sRNA binding site and binding alignments are mentioned. From these alignments we have extracted some of the features given below. This is taken as positive data.

From [8], 250 no interaction pairs are also obtained. These are coming from 17 different genomes. We have downloaded all these 17 genomes. As in the positive data, 95% sRNAs are binding to the mRNAs in their TIR regions, more specifically; this region is between 200nt upstream of start codon to 100nt downstream region of the start codon. So we have extracted the 200nt upstream region to 100nt downstream region for mRNAs of negative data. We made the complements of these sequences for applying BLAST [30]. Out of these 250 sequences for 5 of them no BLAST alignments were found, so we have taken 245 into consideration for further work as negative data.

4.3 Features Extraction

Here we have considered two types of features, namely feature from binding sites and features from secondary structure of sRNAs. For positive data, in database [8], the exact binding sites of both sRNA and mRNA are mentioned. By considering them and complementing the mRNA regions we have extracted the first category of features. For negative data, as already mentioned, we have considered the complements of the regions considered for mRNA (200nt upstream region to 100nt downstream region to the start codon AUG) and applied BLAST [29], and extracted the corresponding features. The secondary structure features of sRNAs are extracted from the secondary structure of sRNA given by RNAfold[20].

4.3.1 Binding Site Features

From the binding sites, we have extracted all possible 24 nucleotide frequency based features from the sRNA and mRNA alignments. These are AA , AC, AG, AU, A-, CA, CC, CG, CU, C-, GA, GC, GG, GU, G-, UA, UC, UG, UU, U-, -A, -C, -G, -U, where '-' indicates the gap between the sequences. Note that these are considered in order, that is, feature 'AG' (representing an 'A' in the sRNA aligned with a 'G' in the mRNA) is different from feature 'GA' (representing a 'G' in the sRNA aligned with an 'A' in mRNA).

4.3.2 sRNA Secondary Structure Features

These are the first four secondary structure features which we already considered for the sRNA prediction in Chapter 3. For this we used the RNAfold [20] for extracting the secondary structures of all sRNA(positive sRNAs and negative sRNAs) and from then we extracted the features namely,

1. Number of CG bonds (pairs) in the structure.
2. Number of AT bonds (pairs) in the structure.
3. Length of maximum continuous bonds
4. Length of maximum non pair sequence (non-bonding in between).

4.4 Classification Using SVM

After extracting the features, we have converted them into the following format to apply LIBSVM [19].

<class index> <feature label: feature value> <feature label: feature value> …<feature label: feature value>

We have 60 positive and 120 negative in training data (1:2 ratio) and remaining 160 (35 positive and125 negative) in test data. We have applied support vector machine with RBF kernel for classification. For choosing the optimum parameters (C and γ in equations 3.1 and 3.3) we have applied grid search with a 5-fold cross validation on the training dataset.

4.5 Classification Results

After the training model is created, LIBSVM is applied on both training data and test data for classification. Each of the above cases is repeated manually four times. Each time randomly chosen training data and remaining data for test data has been considered. In each case sensitivity and specificity are calculated and the ROC curves have been plotted.

Initially we considered only the 24 bindings site features and applied the above the process. The classification results are given in Table 4.1. On test data, on average 86.42% sensitivity, 99.8% specificity and 96.875% accuracy have been obtained. Later we added the secondary structure features of sRNA and applied the above process. The results are given in Table 4.2. On test data, on average 91.42% sensitivity, 99.2% specificity and 97.5% accuracy have been obtained. By observation, it shows that the accuracy and sensitivity are improved which indicates the importance of the secondary structure features considered.

Run	Sensitivity	Specificity	Accuracy
1	85.714286	100.0000	96.87500
2	88.571429	99.2000	96.87500
3	88.571429	100.0000	97.50000
4	82.857143	100.0000	96.25000
AVG	86.42857	99.8000	96.8750

Table 4.1: Classification results with 24 features

Run	Sensitivity	Specificity	Accuracy
1	94.285714	97.6000	96.8750
2	91.428571	100.000	98.1250
3	94.285714	100.000	98.7500
4	85.714286	99.200	96.2500
AVG	91.428573	99.200	97.500

Table 4.2: Classification results with 28 features

4.6 Conclusions

In this chapter we have developed two models for predicting targets of bacterial sRNA. In the first model we considered 24 features and in the second we added 4 more features (secondary structure features of sRNA). SVM with RBF kernel is used for classification. We developed two models. The existing methods like IntaRNA looks for probable complementarity between the sRNA and mRNA and based on matching length they give score based analysis. They are not considering the fact that most of the sRNAs bind to the TIR region. As we incorporated this fact in out model, we hope our model useful for predicting the targets of bacterial sRANs. As we have not extended our model to genome wide, we could not compare with the tools developed for genome wide. We extend our work over genome wide in future.

Chapter 5

Discussions and Conclusions

5.1 Discussions

Small RNAs play important regulatory roles in bacteria by targeting the mRNAs at a post transcriptional level. In this dissertation we develop some computational approaches for analyzing bacterial sRNA. First of all, we establish that they have distinct properties from miRNAs. We also establish that the different classes of sRNAs have distinct features properties. Thereafter we have developed some models for prediction of sRNAs by posing the problem as one of classification. SVM is used as the underlying classification model. New secondary structure features of sRNA have incorporated in this regard. Finally by using the fact the sRNA binds to the translation initiation region of mRNA, we developed two models for prediction of mRNA targets.

5.2 Scope of Future Work

We have developed different models of prediction of sRNA and its targets. Here according to our model, given a region of a bacterial genome, the method will be able to predict whether it is a sRNA or not. The work can be extended to genome wise search for sRNAs. We have not included any thermodynamic features in our models. In future it would interest to see the effect of incorporating such features for sRNA prediction.

We also developed two models for prediction of sRNA targets. Here according to our model, given a sRNA sequence and translation initiation region of an mRNA the method will be able to predict whether it is a probable target of sRNA or not. The work can be extended to genome wide search for probable targets of sRNAs. Incorporation of new features also needs to be studied in this regard. Finally, more sophisticated feature selection techniques need to be applied for improving the performance of the prediction models.

References

1. Joao Setubal, Joao Meidanis, *Introduction to Computational Molecular Biology*, PWS,1999 .

2. Sanghamitra Bandyopadhyay and Ramkrishna Mitra, *TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples*, Bioinformatics, Volume25, Issue20.

3. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar H, Moran N, Hattori M (2006), *The 160-kilobase genome of the bacterial Endosymbiont Carsonella*. Science 314 (5797).

4. Lauren S. Waters and Gisela Storz, *Regulatory RNAs in Bacteria, Cell* 136, 615-625, Feb, 2009.

5. Rolf Backofen and Wolfgang R. Hess, *Computational prediction of sRNAs and their targets in bacteria*, RNA Biology 7:1, 33-42; Jan/Feb 2010.

6. Huang H.Y., Chang H.Y., Chou C.H., Tseng C.P., Ho S.Y., Yang C.D., Ju Y.W., and Huang H.D., *sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes.* Nucleic Acids Research, (2009), Vol 37, D150-D154

7. Weizhong Li & Adam Godzik.*Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*, Bioinformatics (2006) 22:1658-1659

8. Yuan Cao, JiayaoWu, Quian Liu, *sRNATarBase: A comprehensive database of bacterial sRNA targets verified by experiments*, RNA 2010 16:2051-2057.

9. WANG Li-gui, Ying Xiao-ming,CAO Yuan, ZHA Lei, Wu-ju, sRNASVM: *a model for prediction of small non-coding RNAs in E.coli using support vector machines,* acta biophysica sinica Vol.25 No.4, Aug 2009.

10. Rivas E, Klein RJ, Jones TA, Eddy SR, *Computational identification of non-coding RNAs in E.coli by comparative genomics*, Curr Biol,2001,11(17);1369-1373.

11. Rivas E, Eddy SR, *Non-coding RNA gene detection using comparative sequence analysis*, BMC Bioinformatics, 2001,2:8.

12. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EH, Margalit H, Altuvia S. *Novel small RNA-encoding genes in the intergenic regions of Escherichia coli*. Curr Biol, 2001, 11(12):941~950.

13. Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ, Blyn LB. *A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome*. Biosystems, 2002,65(2~3):157~177.

14. Livny J, Fogel MA, Davis BM, Waldor MK. *sRNAPredict an integrative computational approach to identify sRNAs in bacterial genomes,* Nucleic Acids Res, 2005,33(13):4096~4105.

15. Carter RJ, Dubchak I, Holbrook SR. *A computational approach to identify genes for functional RNAs in genomic sequences,* Nucleic Acids Res, 2001,29(19):3928~3938.

16. Saetrom P, Sneve R, Kristiansen KI, Snove O, Grunfeld JT, Rognes T, Seeberg E. *Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming*. Nucleic Acids Res, 2005,33(10):3263~3270.

17. Yachie N, Numata K, Saito R, Kanai A, Tomita M. *Prediction of non-coding and antisense RNA genes in Escherichia coli with Gapped Markov Model*. Gene, 2006,372 (1):171~181.

18. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR and Bateman A, *Rfam: annotating non-coding RNAs in complete genomes*, Nucleic acids research 2005;33;Database issue;D121-4.

19. Chih Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011.

20. Hofacker I.L., Fontana W., Stadler P.F., Bonhoeffer S., Tacker M., Schuster P. *Fast Folding and Comparison of RNA Secondary Structures*. Monatshefte f. Chemie 125: 167-188, (1994).

21. Rolf Backofen and Wolfgang R. Hess, *Computational prediction of sRNAs and their targets in bacteria*, RNA Biology 7:1, 33-42; Jan/Feb 2010.

22. Yalin Zhao, Hua Li, Yanyan Hou, Lei Cha, Yuan Cao, Ligui Wang, Xiaomin Ying, Wuju Li, *Construction of two mathematical models for prediction of bacterial sRNA targets*, Bi0chemical and biophysical Research Communication37 (2008) 346-350.

23. Yuan Cao, Yalin Zhao, Lei Cha, Xiaomin Ying, Ligui Wang, Ningshen shao, Wuju Li, *sRNATarget: aweb server for prediction of sRNA targets,* Bioinformation 3(9): 364-366(2009).

24. Zhang Y., Sun S., Wu T., Wang J., Liu C., Chen L., Zhu X., Zhao Y., Zhang Z., Shi B., Lu H., Chen R., *Identifying Hfq-binding small RNA targets in Escherichia coli*, Biochem. Biophys. Res. Commun. 343 (2006) 950–955.

25. Vogel J., Wagner E.G., *Target identification of small non-coding RNAs in bacteria*, Curr. Opin. Microbiol. 10 (2007) 262–270.

26. Tjaden B., Goodwin S.S., Opdyke J.A., Guillier M., Fu D.X., Gottesman S., Storz G., *Target prediction for small, non-coding RNAs in bacteria*, Nucleic Acids Res. 34 (2006) 2791–2802.

27. Brian Tjaden, *TargetRNA: a tool for predicting targets of small RNA action in bacteria,* Nucleic Acids Research, 2008, Vol.36.

28. Christophe Pichon and Brice Felden, *Small RNA gene identification and mRNA target predictions in bacteria,* Bioinformatics, Vol. 24 no. 24 2008, pages 2807-2813.

29. Anke Busch, Andreas S, Richter and Rolf Backofen, *IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions*, Bioinformatics, Vol. 24 no. 24 2008, pages 2849-2856.

30. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman "*Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*", (1997), Nucleic Acids Res. 25:3389-3402.