FABIO IVAN REINOSO VILCA

PROPOSAL OF A DATA MINING PIPELINE TO IMPROVE AB INITIO PREDICTION OF BACTERIAL SMALL RNA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
R373p
2018

Reinoso Vilca, Fabio Ivan, 1988-
Proposal of a data mining pipeline to improve ab initio prediction of bacterial small rna / Fabio Ivan Reinoso Vilca. – Viçosa, MG, 2018.
xiv, 64 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Fábio Ribeiro Cerqueira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 58-64.

1. Mineração de dados (Computação). 2. Bioinformática. 3. RNA. 4. Aprendizado do computador. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. II. Título.

CDD 22. ed. 006.312

FABIO IVAN REINOSO VILCA

PROPOSAL OF A DATA MINING PIPELINE TO IMPROVE AB INITIO PREDICTION OF BACTERIAL SMALL RNA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 23 de fevereiro de 2018.

Marcos Henrique Fonseca Ribeiro

Denise Mara Soares Bazzolli

Fabio Ribeiro Cerqueira
(Orientador)

I dedicate this work to my beloved mom and sister, they are my inspiration and support to go forward through everything in this world.

Acknowledgments

First of all, I want to thank God, I passed many difficulties since my first year in Brazil, but always He gives me the strength to overcome all these problems.

Thanks to my professors of DPI, for the patience and dedication in their labor. I liked a lot to learn many different subjects that I can't in my country, and for understanding my portunhol in the tests. A special thanks to my advisor Fabio for his patience, teaching, to give me a big challenge to do this research.

To Kenneth, thanks to clear my doubts about the sRNA and for making a great project of Ecogene.

To the Organización de los Estados Americanos (OEA), Grupo Coimbra de Universidades Brasileiras (GCUB), CAPES, CNPq, and Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), and Universidade Federal de Viçosa (UFV), for giving me the opportunity to study in this beautiful city and I've had the best experience of my life until now.

To my colleagues Thales and Cleysinho for all the help especially at the beginning of my thesis.

To all my friends in Brazil, thanks for all the support and for being my family in this country, in especial to Guidson, Vinicio, Roberta, Jonatas, Daniel, Camila, Hector, Vinicius, Joana, Liliane, Aly, and Gilson.

To my friend Marco, without you, I never would have passed the first semester here, thanks for everything, and for your invaluable contribution to this research. By the way thanks for give me the best days here in Brazil, finally, I admit it.

To my family in Panama, Peru and Australia, especially to my aunts Mary and Vicky, my uncle Luis, my brother in law Timmy and my dad Fabio for the constant support.

Finally, to the most important people in my life my sister Jessica and my mom Irma thanks for everything you're my force, my inspiration my world ..., without you, I would never have been able to finish my research.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ACC:	*Accuracy*
ARFF:	*Attribute-Relation File Format*
AUC:	*Area Under the Curve*
AUROC:	*Area Under the ROC curve*
BSRD:	*Bacterial Small Regulatory RNA Database*
CAPES:	*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*
CDS:	*Coding Sequences*
CFS:	*Correlation-based Feature Selection*
CSV:	*Comma-separated Values*
DNA:	*Deoxyribonucleic Acid*
FAPEMIG:	*Fundação de Amparo a Pesquisa do Estado de Minas Gerais*
FN:	*False Negatives*
FP:	*False Positives*
ID:	*Identifier*
IG:	*InfoGain*
IGR:	*Intergenic Region*
MCC:	*Mathews Correlation Coeficient*
ML:	*Machine Learning*
MLP:	*Multilayer Perceptron*
mRNA:	*messenger RNA*
ncRNA:	*Non-coding RNA*
nt:	*nucleotides*
Pr:	*Precision*

RNA:	*Ribonucleic Acid*
ROC:	*Receiver Operating Characteristic*
SMO:	*Sequential Minimal Optimization*
Sn:	*Sensitivity*
SPC:	*Specificity*
sRNA:	*Small RNA*
SVM:	*Support Vector Machine*
TN:	*True Negatives*
TNR:	*True Negative Rate*
TP:	*True Positive*
TPR:	*True Positive Ratio*
UFV:	*Universidade Federal de Viçosa*
WEKA:	*Waikato Enviromment for Knowledge Analysis*

Abstract

REINOSO VILCA, Fabio Ivan, M.Sc., Universidade Federal de Viçosa, February, 2018. **Proposal of a Data Mining Pipeline to improve ab initio prediction of Bacterial Small RNA**. Adviser: Fábio Ribeiro Cerqueira. Co-Adviser: Sabrina de Azevedo Silveira.

Bacterial small RNAs (sRNAs) are usually non-coding RNAs (ncRNAs) with a size of 50–500 nucleotides, and act mainly as post-transcriptional regulators. Prediction of sRNAs is a challenging issue in bioinformatics. The current computational tools deliver a high number of false positives. Hence, the development of more precise predictive methods is of fundamental importance to narrow the number of costly and time-consuming sequence validations on the laboratory workbench. In this work, we collected a series of features from the existent computational tools for ncRNA prediction in order to select the best ones for classifying putative bacterial sRNA sequences. Out of the 264 initially-chosen features, 22 relevant and non-redundant features could be selected by using feature-selection algorithms. To validate this proposal we used a dataset built with only experimentally-validated sRNAs from different bacteria sub-strains, considered as model organisms in genetics, as well as non-sRNA sequences. Finally, a Random Forest algorithm was applied for the classification task. Our first validation experiment of this proposal covered the single sequence prediction task, using 6 testing sets. Our pipeline presented better results than the only *ab initio* method we could find in literature. The differentiating characteristics of our method are the lower computational cost, the dimensionality reduction and the analytic power analysis due to the single 22 features selected. Our approach could reach an average of 80% of *Accuracy*, 71.28% of *Precision*, 82.11% of *Specificity* and an area under the ROC curve of 0.879. Furthermore, we presented a Genome-wide *framework* to sRNA prediction, obtaining a 39% lower False Positive Ratio and the double of Specificity than the above-mentioned *ab initio* method.

Resumo

REINOSO VILCA, Fabio Ivan, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Proposta de uma Pipeline de Mineração de Dados para melhorar a predição ab initio de Pequenos RNAs Bacterianos**. Orientador: Fábio Ribeiro Cerqueira. Coorientadora: Sabrina de Azevedo Silveira.

Pequenos RNAs (sRNAs) são RNAs não codificantes (ncRNAs) com um tamanho de 50 a 500 nucleótidos e atuam principalmente como reguladores pós-transcrição. A predição de sRNAs é um problema aberto na bioinformática. As ferramentas computacionais atuais fornecem um alto número de falsos positivos. Desta forma, o desenvolvimento de métodos preditivos computacionais são de grande importância para reduzir o número de sequências putativas que implicam altos custos e tempos de validação em laboratório. Neste trabalho, reunimos uma série de atributos utilizados em métodos prévios, baseados em aprendizado de máquina para a predição de ncRNA, a fim de selecionar os melhores para classificar sequências putativas bacterianas de sRNA. Dos 264 atributos coletados inicialmente, 22 atributos relevantes e não redundantes foram selecionados usando algoritmos de seleção de atributos. Para validar esta proposta, foi usado um conjunto de dados construído com sRNAs validados experimentalmente de diferentes sub-cepas de bactérias consideradas como organismos modelo em genética, assim como seqüências não-sRNA. Finalmente, o algoritmo de Random Forest foi usado com a finalidade de realizar a tarefa de classificação. A primeira validação da abordagem aqui proposta foi em sequências completas de sRNA em 6 conjuntos de testes. A abordagem proposta, apresentou melhores resultados do que a única ferramenta *ab initio* que pudemos encontrar na literatura. As características diferenciais do método proposto são o baixo custo computacional, redução de dimensionalidade e análise de poder analítico devido aos 22 atributos selecionados. Nossa abordagem atinge uma média de 80% de *Precisão*, 71,28% de *Precisão*, 82,11% de *Especificidade* e uma área sob a curva ROC de 0,879.

Além disso, apresentamos um *framework* para a predição em genoma bacterianos de sRNAs, que apresenta uma taxa 39% menor de Falsos Positivos e o dobro da Especificidade do que o método *ab initio* acima mencionado.

Chapter 1

Introduction

Bacterial small RNAs (sRNAs) comprise a class of RNAs with a size ranging from 50 to 500 nucleotides (nt) (Huang et al., 2009). A typical bacterial genome contains from 100 to 200 sRNAs. These small molecules usually act as gene expression regulators by binding to their target proteins or mRNAs, and are involved in many fundamental activities such as: iron homeostasis (Massé et al., 2005, 2007), expression of outer membrane proteins (Guillier and Gottesman, 2006; Valentin-Hansen et al., 2007), quorum sensing (Feng et al., 2015; Lenz et al., 2004), bacterial virulence(Pitman and Cho, 2015) and other important functions(Michaux et al., 2014). sRNAs can be divided into two major groups: (i) mRNA-binding antisense sRNAs and (ii) protein-binding sRNAs (Li et al., 2013). The first group is the focus of our work. This group can be categorized as cis-encoded antisense sRNAs, which are completely complementary to their target; and trans-encoded antisense sRNAs, which are only partially complementary to their target. The sRNA–mRNA interaction in both categories comprises a significant part of the gene regulation network.

There exist two main types of prediction tools: (i) Methods based on homology: These methods use different approaches like the search of sequences highly conserved in intergenic regions in bacterial genomes phylogenetically close, as well as the conservation of RNA secondary structure and the use of genomic signals like promoters and terminators. The main disadvantage of these methods is obtaining good results only with genes highly related to the genes that are already known and not with new families(Tran et al., 2009). (ii) De novo or ab initio methods: These methods do not depend on any previous knowledge about the organism. The disadvantage of these methods is the high number of False Positives in their results (Tran et al., 2009).

The present study has an objective the development of new specific ab initio

tool for sRNA prediction, looking for improving the balance between the precision and sensitivity of the final predictions that is the main disadvantage of the current methods. Also, we want to present a Framework to sRNA Genome-wide prediction in prokaryotic genomes.

1.1 Problem and its importance

Although the number of sRNAs identified in recent years has been growing due to advancements in bioinformatics and sequencing technologies, the current computational and experimental methods still await significant amelioration, specially regarding the minimization of false positives. The difficulty is due to challenging characteristics of sRNAs such as: (i) varying size (50 - 500 nt); (ii) no common secondary structure; (iii) sRNAs do not show any statistically distinguishable nucleotide biases; and (iv) there is a poor conservation between sRNAs of distantly-related genomes (Sridhar and Gunasekaran, 2013).

In general, the current ab initio methods cannot provide good sensitivity and precision at the same time. If the parameters are set to favor sensitivity, then precision usually decreases. The same happens the other way around, i.e., when the parameters are optimized for a high precision, sensitivity is drastically reduced (Arnedo et al., 2014). In addition, the prediction accuracy of ab initio approaches is highly variable. Also, it is a complex task to compare them with each other due to a lack of a suitable benchmark, and because their results are reported using different statistical measures. Furthermore, some standard statistical measures are not reported. The area under the Receiver Operating Characteristic (ROC) curve, for example, is omitted by most papers that describe the current ab initio methods. Even accuracy, sensitivity, and precision are often absent from the results.

The significant importance about the biological functions in bacteria and the key role of the sRNAs in their regulations makes a necessity to improve the predictions of the sRNAs to avoid the cost and time of the experimental validations.

1.2 Justification

The quasi-inexistent ab initio tools for sRNA discovery, and the disadvantages of the homology tools mainly in search of new kind of families of sRNA. Lead to the necessity to develop a new approach in this field. Furthermore, the increase of information about families of sRNA and the new features of ncRNA published in

the last years give us an opportunity to apply machine learning methods to create a better and robust predictive model.

1.3 Hypotheses

It is possible to develop a new robust method applying machine learning techniques to obtain high performance sRNA prediction, not only for single sequence classification, but also for genome-wide prediction, taking advantage of all publicly available information about sRNAs.

1.4 Objectives

1.4.1 General Objective

Establish a robust machine learning (ML) pipeline to overcome the issues of current ab initio approaches, with a special attention in keeping sensitivity as well as precision at valuable levels.

1.4.2 Specific Objectives

- To collect reliable data to compose our training sets with the most representative sequences of sRNA.
- To collect significant features related to ncRNA prediction.
- To define a machine learning classification algorithm to be applied.
- To develop a new Framework of Genome-wide prediction.
- To demonstrate the effectiveness of the new method with the current prediction tools, measuring sensitivity, precision, and accuracy.

1.5 Outline

This thesis is organized as follows. In Chapters 2 and 3 we introduce some background information about basic biological concepts related to the present study as well as the two different approaches in sRNA prediction. In Chapter 4 we describe the machine learning algorithms and the measures used in our experiments. Chapter 5 presents all the steps applied in our methodology as well as the description

of all the data and features used in our prediction model. Also in this chapter, it is described our Genome-Wide prediction Framework. In Chapters 6 and 7, we provided our results and conclusions, respectively.

Chapter 2

Basic Concepts about Bacteria and Molecular Biology

2.1 Bacteria

Bacteria are prokaryotes microorganism with a size between 0.5-5 micrometers and were the first forms of life to appear on Earth. They consist of a single cell, but unlike eukaryotic cells, bacteria do not have organelles and nucleus. Their genetic material consists of a circular chromosome of DNA in the cytoplasm. They are found in every habitat on Earth even in the more extreme environment, such as Arctic sea-ice at subzero temperature (Junge et al., 2002) or in regions with temperatures around 100°C (Stetter, 2006).

Bacteria can be beneficial as well as harmful to humans. They can be pathogenic and cause many diseases like pneumonia, meningitis, tuberculosis, cholera, dysentery, etc. Also, there are helpful bacteria involved in nutrient absorption, food digestion, vitamin production, etc (Sears, 2005).

2.1.1 Classification

There are many classifications for bacteria, but the most used is based on the nature of their cell walls.

The Gram stain test is used to classify bacteria into two categories according to the structural characteristic of their cell walls after been dyed with a crystal violet. There are: (i) Gram-positive bacteria: They take up the crystal violet stain and appear with a purple color. (ii) Gram Negative bacteria: They cannot retain the violet and appear colored red or pink, this is due to the difference in the structure

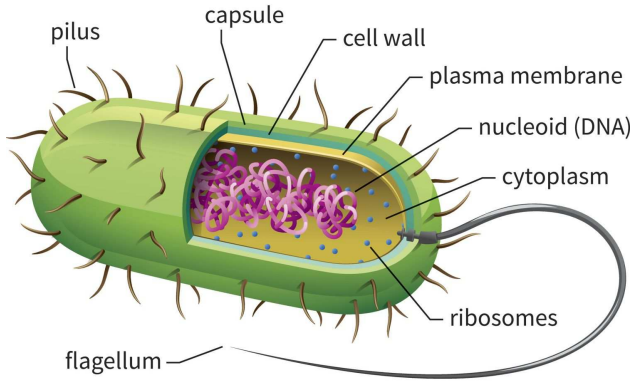


Figure 2.1. Prokaryotic cell. Source (Jack0m Getty Images, 2017)

of their bacterial cell wall, Gram-positive does not have an outer cell membrane in contrast to the Gram-negatives. Because of this characteristic, the Gram- Positives are more receptive to antibiotics than the Gram-Negatives.

2.2 Genome

The Genome is the genetic material of an organism, which consists of DNA (Deoxyribonucleic acid). Most bacteria contain their genetic material in a single circle chromosome. In comparison with eukaryotes, bacteria have less genetic information. For example, human has about 25,000 genes in a 3,234.83 Mb (Mega-basepairs) per haploid genome and *Escherichia coli* K12 MG1655 (bacterium used in this study) has around 4000 genes in a genome of 4.498 Mb.

DNA is a macromolecule composed of a double strand sequence of nucleotides. Each nucleotide is composed of a 5-carbon sugar and a nitrogenous base, i.e., Cytosine (C), Guanine (G), Adenine (A) and Thymine (T). Both strands of the DNA are bonded by base pairs (bp), there are rules of base pairing: Adenine (A) always pairs with Thymine (T) and Cytosine (C) with Guanine (G).

In this study, we use two terms *forward* or *plus strand* and *reverse* or *minus strand* to identify the two strands of the genome. In bacteria, the forward direction is the direction that replication happens. As shown in Figure 2.2, the reverse strand is complementary to the forward strand.

Figure 2.2. Forward and reverse strands.

As mentioned above, DNA contains all the genetic information. DNA encodes information need to a process called *transcription*, where RNAs (Ribonucleic acid) are produced. An RNA is a single-stranded molecule with the same bases of DNA only replacing the Thymine with Uracil (U). Differences between both molecules are shown in Figure 2.3.

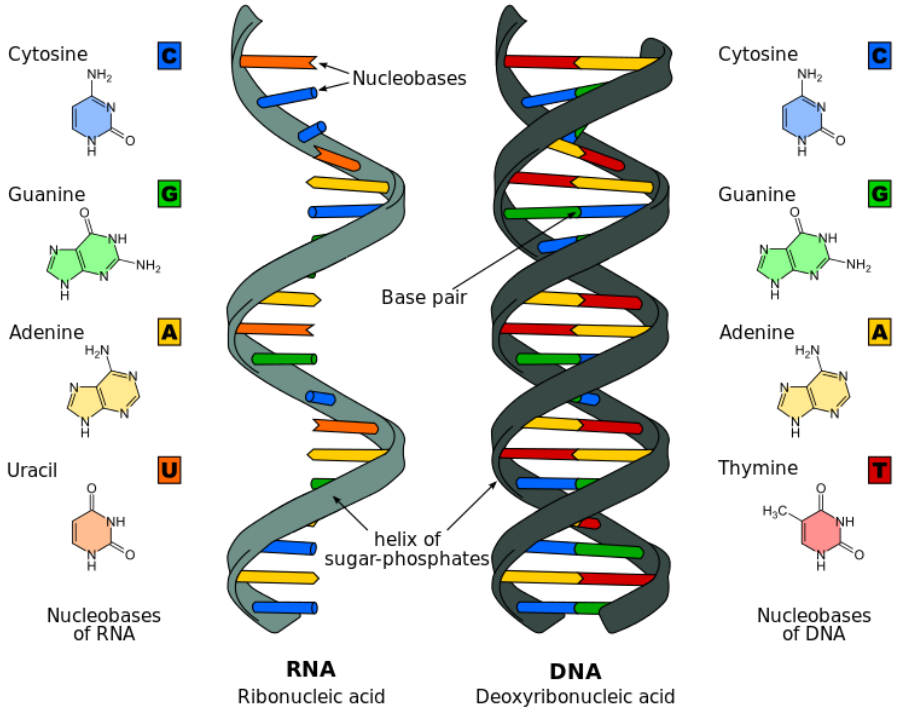


Figure 2.3. DNA vs RNA. Source (Sponk Wikimedia Commons, 2010)

2.2.1 Central Dogma

The Central Dogma is a framework to explain the flux of information between the three biopolymers: DNA, RNA, and Protein. This flux has three main processes: (i) DNA replication: The process in which a cell makes an identical copy of their entire genome. (ii) Transcription: Process in which the enzyme RNA polymerase reads a DNA segment and produces an RNA molecule out of complementary nucleotides. This RNA transcript has the same information of the DNA, except by containing

the base Uracil (U) instead of Thymine (T). (iii) Translation: It is the process in which the Messenger RNA (mRNA) is decoded to a protein.

As shown in the Figure 2.4, there exist exceptions to the central dogma, like the transcription of RNA to DNA called reverse transcription in some retroviruses or the RNA replication that is the copy of RNA to a new RNA.

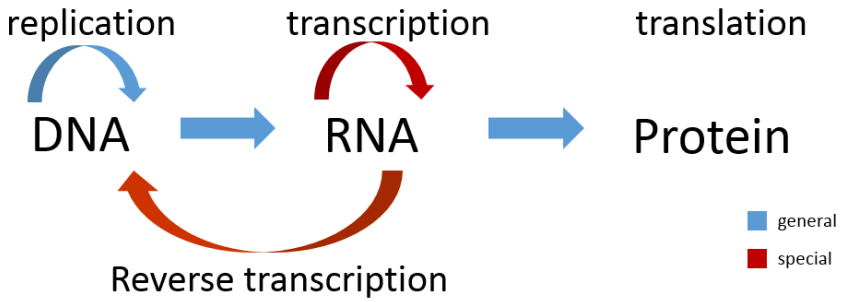


Figure 2.4. Central Dogma.

2.2.2 The RNA

RNAs can be divided into two types, RNAs that can be translated into Proteins called Coding RNAs and those that do not, called non-coding RNAs (ncRNAs).

During a long time, a significant part of the studies in genetics was focused on the search of protein-coding RNAs, until the discovery of the first ncRNAs, i.e., the transfer RNA (tRNA) in 1968 and the ribosomal RNA (rRNA) in the early 1980s, both of them involved in essential roles in protein synthesis. In 1998, a new kind of ncRNA was discovered. That ncRNA could switch of certain mRNAs, i.e., it can be involved in important functions like the gene regulation and of course, change the perspective of the Central Dogma. Currently, there is a critical necessity for a better understanding all the functions of ncRNAs, because they are involved in important roles, i.e., genetic diseases, neurological disorder, cancer, etc (Panwar et al., 2014).

RNAs can be classified by their function as shown in figure 2.5 and the focus of this study are the sRNAs.

2.2.3 Transcription signals

There are two transcription signal mentioned in this study (i.e., promoters and terminators), as shown in Figure 2.6 it can be observed their positions in relation with the region to be transcribed.

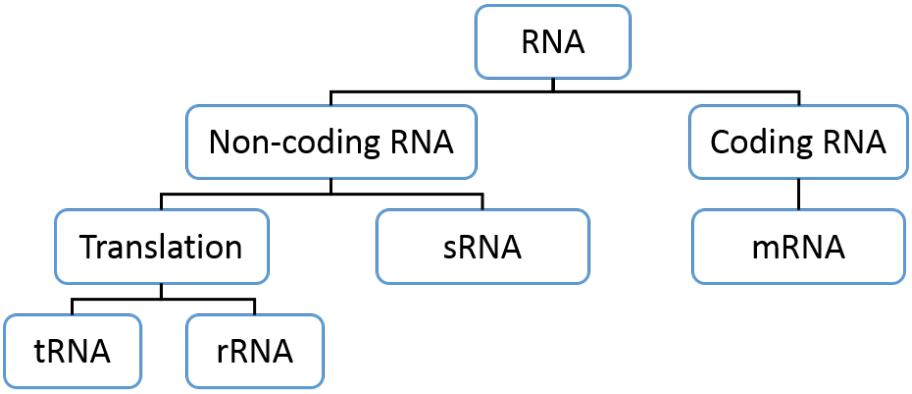


Figure 2.5. Classes of RNA molecules according to their functions.

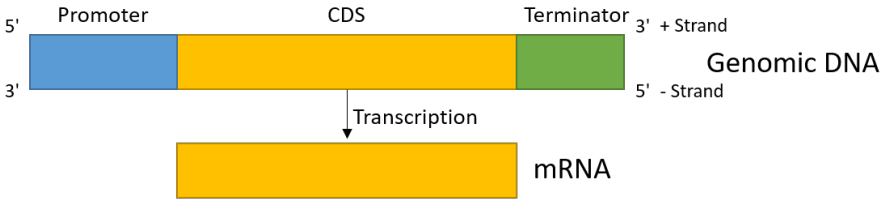


Figure 2.6. Typical bacterial transcription unit and its product.

- Promoter: It is a region in DNA where starts the transcription process of a gene. This region typically consists of two short sequences, which are separated by a defined number of bases, i.e., -10 and -35 boxes. These two sequences are located approximately 10 and 35 bases upstream respectively from the start position of transcription.

- Terminator: It is a section of the DNA that marks the end of transcription. In prokaryotes exist two types of terminator, i.e., Rho-dependent and Rho-independent.

Chapter 3

Bacterial Small RNAs

3.1 Definition

Regulatory non-coding RNA commonly known as Small RNA (sRNA) is a type of RNA that acts as a functional element in a prokaryotic cell, with a size between 50-500 nucleotides (nt) (Huang et al., 2009). It acts on independently expressed targets. Even when the first sRNA was discovered by the year of 1981 (Tomizawa et al., 1981), it was not until by the years 2001-2002 with the development of new bioinformatics methods based on homology that starts the discovery of many new sRNAs in *Escherichia coli* and closely related bacterial species (Azhikina et al., 2016).

The sRNAs are very important because they are involved in many fundamental activities in the bacteria such as: iron homeostasis (Massé et al., 2005, 2007), expression of outer membrane proteins (Guillier and Gottesman, 2006; Valentin-Hansen et al., 2007), quorum sensing (Feng et al., 2015; Lenz et al., 2004), bacterial virulence (Pitman and Cho, 2015) and other important functions (Michaux et al., 2014).

Currently, the exact number of sRNAs present in a bacterial genome is still unknown. For example, in *Escherichia coli* the most studied and well-understood organism, only have been confirmed around of 80 sRNA, only 2% in comparison with the other genes identified (Waters and Storz, 2009), but it seems like the numbers of sRNA in a bacteria it is around a few hundred instead of thousands of sRNAs (Gottesman and Storz, 2011).

3.2 Types

The sRNAs can be classified into three major groups:

- The Antisense: This type of sRNA is the most characterized, this regulates the target mRNA by base pairing. It is categorized into two groups:
  - Cis-encoded antisense RNAs: This RNA is fully complementary and encoded at the same locus but in the opposite strand of their target mRNA. This sRNA is involved in important regulatory processes such as translation, transcription, initiation of replication, plasmid conjugation, transposition, and mRNA degradation. In contrast with Trans-encoded, the physiological roles of the cis-encoded anti-sense are insufficiently studied (Azhikina et al., 2016; Waters and Storz, 2009).
  - Trans-encoded antisense RNAs: This sRNA is located in genome regions distant from their target mRNA. Furthermore, the base-pairing with their target is too short, i.e., 10 to 25 base pairs. Because of this, many of these sRNAs required a protein chaperone Hfq to facilitate the RNA-RNA interaction between the sRNA and their target. This type of sRNA is involved in response to environmental stimuli or stress situations, e.g., nutrient starvation, quorum sensing, membrane stress, oxidative stress, and SOS response to DNA damage (Beisel and Storz, 2011).

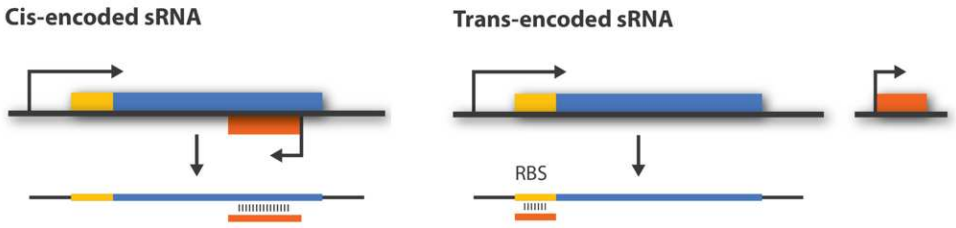


Figure 3.1. Types of Antisense sRNA. As shown in the figure, it can be observed the difference of the genomic locations between the cis- and trans-encoded sRNA. The blue boxes represent the target mRNA, yellow boxes the ribosome binding site and the orange boxes the sRNA.

- Protein Regulators: This type of sRNA binds to a protein to inhibit or modify the protein's activity, and usually has many binding sites for protein recognition. Some examples are 6S RNA and CsrB/c (Javayel et al., 2011).

- House-keeping RNAs: This group includes three ubiquitous sRNAs involved in the control of many processes in the prokaryotic cell, i.e., transfer-messenger RNA (tmRNA), 4.5 sRNA and the catalytic subunit of the tRNA-processing RNase P (Javayel et al., 2011).

3.3 Secondary Structure

The secondary structure is the second level of the Nucleic acid structure. After the transcription process in the genome, start a base-pairing of complementary nucleotides in the RNA and form an structure highly related to their function.

A single molecule of RNA can adopt many different structures due to existing many ways of pairing between their bases. The collection of all possibles structures its called ensemble structure. Usually, the predicted secondary structure is the structure with the minimum free energy (mfe), because it is thermodynamically the most stable structure.

There exist two common representation of secondary structures: (i) The bracket notation: each dot represents an unpaired base, open parentheses indicates that the base is paired to another base ahead of it and closed parentheses denotes that a base is paired to another base behind it. (ii) Bi-dimensional representation: It is a graphical representation of the structure including in some cases the bases and is easier to recognize different elements of the secondary structure like the helix, loops, hairpin loops, etc.

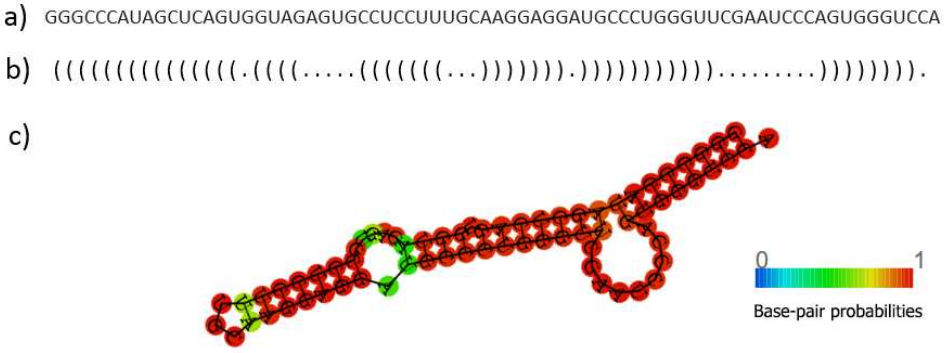


Figure 3.2. Typical RNA secondary structure. As shown in the figure the primary sequence representation of the RNA (a), the bracket notation (b) and the bi-dimensional representation of the secondary structure (c) generated using RNAfold (Lorenz et al., 2011).

3.4 Common characteristics

- Location: The major part of trans-encoded sRNAs genes were found in intergenic regions (IGR) (i.e., region of noncoding DNA between two coding genes) in the plus strand of the genome (5' to 3'), in the most of cases in opposite sense of the surrounding genes (Storz et al., 2011).
- Transcription signals: sRNAs commonly have Rho-independent terminators and promoter sigma 70 (σ^{70}). But, they are not easily recognizable and thereby are unreliable indicators.
- Structure: The evolutionary process of sRNA is unique, due to the occurrence of many base substitutions in the genes but preserving the secondary structure. The conservation of the primary sequence is only with a phylogenetically close organisms. Furthermore, some sRNA acquires a secondary structure with mfe lower than the random sequence of the same size with the same mono and dinucleotide frequencies (Clote and Bayegan, 2015).

3.5 Identification methods and prediction of sRNAs

The typical flow to discover new sRNAs is to select a group of candidates using different computational prediction tools, followed by experimental validation of the putative sequences.

3.5.1 Laboratory identification methods

There exist many experimental methods to validate sRNAs, i.e., direct sequencing of ncRNAs, shotgun cloning of small-sized ncRNAs (cDNA libraries), microarray analysis, genomic SELEX and Northern blots (Hüttenhofer and Vogel, 2006).

Even with many experimental methods available, up until now, it is very difficult to detect novel sRNAs, because they are expressed in particular conditions, depending on factors like growing phases or related to cellular stress. Because of these circumstances, many sRNAs are not discovered yet (Altuvia, 2007). Unfortunately, the experimental validation of the putative sequences is a costly process. Therefore, it is necessary to use computational methods to reduce the number of putative sequences (Altuvia, 2007).

3.5.2 Computational prediction methods

There are many computational tools used to predict sRNA, most of them are not specific to this type of RNA. They can be divided into two categories: Homology methods and ab initio methods.

3.5.2.1 Homology based methods

Comparative genomics-based models or homology methods are the most commonly used to find new sRNAs. The basic assumption of these methods is that exist conservation of sequence and secondary structure of the sRNAs among bacterial genomes phylogenetically close. An example of this was applied to the genome of *Escherichia coli* and selecting close bacterial genomes like *Salmonella enterica* serovar Typhi, *Salmonella enterica* serovar Paratyphi and *Salmonella enterica* serovar Typhimurium. As a result, 24 putative genes of sRNA were identified (Argaman et al., 2001).

This method usually has some similar steps: (i) Find or choose the closely-related genomes to the target genome. (ii) Extract the intergenic regions (IGR) of the genomes selected and apply BLAST (Altschul et al., 1990) to compare these regions. Then, the pair wise BLAST (Altschul et al., 1990) hits are grouped into clusters of two or more sequences, and then, are aligned using ClustalW (Thompson et al., 1994) or ncDNAalign (Rose et al., 2008). (iv) Then the results are scored with RNAz (Gruber et al., 2010) or EvoFold (Pedersen et al., 2006). (v) It is applied a structural conservation analysis for the IGR that means for some positions in each sequence the base pairing information is kept. (vi) The final step is to predict or find transcription signals of the putative sequences, like promoter, transcription factor binding sites or Rho-independent terminator (Li et al., 2012).

The methodology described above comprises (partially or totally) the following computational methods: QRNA (Rivas and Eddy, 2001), RNAz (Gruber et al., 2010), EvoFold (Pedersen et al., 2006), SIPHT (Livny et al., 2008), and sRNAPredict (Livny, 2005).

There exist some disadvantages of this type of method: (i) There is no rule to choose the best genomes to be included in comparative genomic-based prediction. (ii) These methods only can be applied to the discovery of evolutionarily-conserved sRNA and not to find new unique genes for a given genome.

3.5.2.2 Ab initio methods

Ab initio is a term for any method that makes predictions about biological features using machine learning approaches without previous knowledge about the microorganism to be studied and sometimes interchangeable with the term *de novo*. These methods can discover new families of sRNA.

This methodology can be resumed in the following steps: (i) First step is to construct a training dataset composed of positive and negative samples. The positives samples are sRNA genes, and the negative samples are another kind of non-sRNA sequences (e.g., random sequences, coding sequences, shuffle sequences from the positive samples, etc.) (ii) The second step is to extract features from both types of samples. Have the best features to describe the data is a key part to obtain robust models. (iii) Third, machine learning methods are applied to develop a prediction model using the training set. (iv) The model is applied to genome-wide discovery or single sequence prediction for experimental validation (Li et al., 2012).

The main disadvantage of these methods is the high quantity of False Positives (FP) in genome-wide prediction. Usually, many filters are applied using transcription signals to reduce the FP. Some studies using this method are from Tran et al. (2009) and Barman et al. (2017).

1. Tran Approach: This was the first ab initio approach of ncRNA prediction in Prokaryotic related to sRNAs. Tran et al. (2009) extract from many sources a high quantity of sequences of ncRNAs and only filtered the tRNA and rRNA, and create its negative set, shuffling each of the ncRNA sequences while preserving the mono and di-nucleotide frequencies. This approach uses a neural network to do a genome-wide prediction in *Escherichia coli*, recovering 84/93 ncRNA genes documented by Tran et al. (2009), but obtaining a high quantity of false positives. This work introduces new features in ncRNA identification as well as a methodology of genome-wide prediction with the use of many filters to reduce the search space. There is some problem with this work, they described two results in genome-wide prediction, the first one reaches a Precision of 0.56% with 16,571 putative sequences and after applying some filters to this sequences obtained a Precision 6.32 % with 601 putative sequences. Only the 16,571 sequences were published and not exist good documentation of the filters applied.

2. Barman Approach: The more recent work on sRNA prediction. Barman et al. (2017) use a set of 182 sequences of sRNA from *Salmonella enterica* serovar

Typhimurium LT2 (SLT2) and a negative set of two random sequences from each sequence of sRNA. This approach uses as features tri-nucleotides frequencies in a combination of a support vector machine to do the classification process. There is some weakness in this study: (i) Barman et al. (2017) compare their results with other prediction tools but using their training set, maybe due to the complexity to do a test with a different type of approaches and the nonexistence of other ab initio tool. (ii) Their model is only tested in two different strains of *Salmonella enterica* and one phylogenetically close genome from other species (i.e., *Escherichia coli*). (iii) They did not publish real measures in genome-wide prediction, just the number of sRNA recovered. (iv) They confuse the coordinates between the genome version used and the sRNAs in *Escherichia coli* K12. Despite these problems, the computational tool provided by Barman et al. (2017) is the only one we could find ready for use and test.

Chapter 4

Machine Learning

Machine Learning is a subfield of Computer Science in the branch of artificial intelligence with the objective to develop new computational techniques focused on the construction of systems with an automatic acquisition of knowledge from experience (Langley and Simon, 1995).

A more formal definition is "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 2006).

Machine learning picks ideas from different disciplines like probability, statistics, neurobiology, computational complexity, etc (Mitchell et al., 1997). Furthermore machine learning has been successfully applying in many areas: Game playing (e.g., chess, poker, etc.), image recognition, medical diagnosis, agriculture, music, natural language processing and many more (Ayodele and Zhang, 2010).

The machine learning algorithms are often classified into two categories:

- Supervised Machine Learning: This type of machine learning is the more commonly used. The process starts with an input variable X and output variable Y. Supervised algorithms aim to learn a mapping function from input X to output Y, i.e., $Y = f(X)$. With the objective to predict an output Y correctly with new input data X. In other words, the input data are pre-categorized or labeled with a class, and this information (i.e., training dataset) is used to train a model to classify new input data. The supervised machine learning can be divided into: (i) Classification: It is when the output is a class or category (e.g., "black" or "white" and "male" or "female"). (ii) Regression: It is when the output is a real value (e.g., "temperature" or "distance").

- Unsupervised Machine Learning The unsupervised machine learning is closer to be real intelligence, due to that it does not need any supervision. The algorithms are left to identify patterns in the data without human guidance (i.e., no labeled input). Unsupervised learning is categorized into two types: (i) Clustering: Find the inherent groupings in the data (e.g., advertising platform using purchasing habits of customers). (ii) Association: Find rules that describe a significant portion of the data.

4.1 Classification

Classification is a task to assign an object to one predefined category or class. This process can be resumed into two steps: (i) Training or learning step: A classifier model is constructed by using a classification algorithm with an input data of series of records called Training set. Each of these records is characterized by a tuple (X,Y), where X is an n-dimensional vector of attributes $X_n = (x_1,x_2,...,x_n)$, and Y is the respective belonging class. The goal of a classifier is to learn a target function $Y = f(X)$ that maps each of attributes of X to one of the predefined labels Y, this target function is known as the classification model. (ii) Classification step: The classification model is used to predict labels of a new given data, usually called as Testing set.

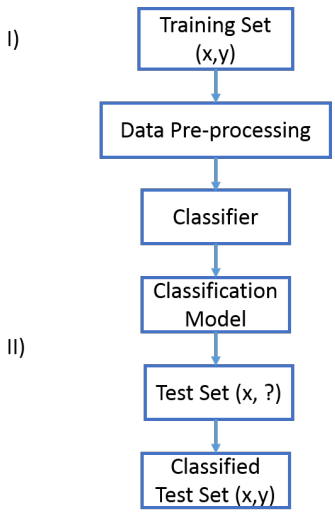


Figure 4.1. Classification process.

4.2 Machine Learning Algorithms

Below, we described each of the supervised machine learning algorithms used to construct our classification models in this work.

4.2.1 Decision tree C4.5

Decision trees are non-parametric supervised learning methods used for classification. These algorithms aim to create a prediction model by learning simple decision rules inferred from a training data (J.Han, 2012).

In decision trees, the internal nodes represent a test on a specific attribute, each branch of the nodes represents the result of the test, and each leaf node represents a final class.

The decision tree C4.5 algorithm proposed in 1993 by Quinlan (1993) is an improvement of the ID3 algorithm also developed by him. This algorithm generates a decision tree used for classification.

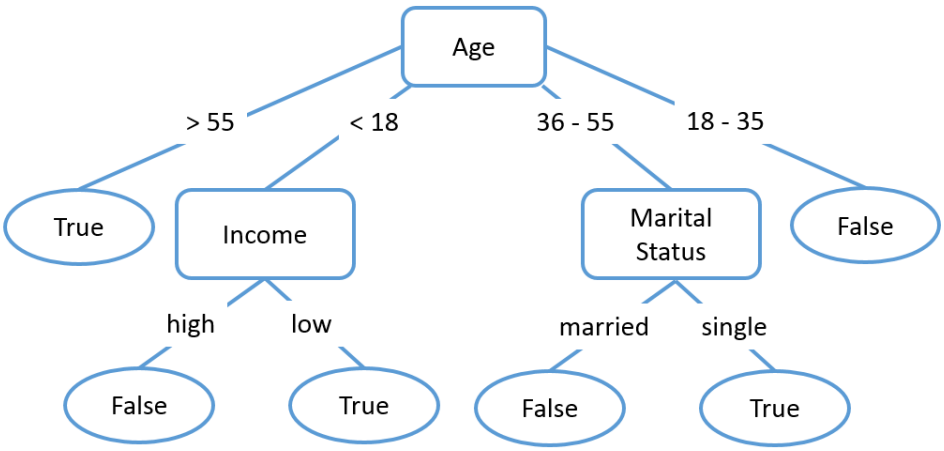


Figure 4.2. Decision tree, with two types of classes, i.e, "True" and "False".

The training set is denoted as $S = s_1, s_2, ...$ Each s_i represents a sample from the training set with a p-dimensional vector $(x_{1i}, x_{2i}, ..., x_{pi})$, where each x_p represents an attribute of the sample as well as its class.

The tree is constructed from the top to down, doing a greedy search. At each node of the tree, it chooses an attribute x that most effectively splits the samples into subsets enriched in one class or the other. To do this splitting, the C4.5 uses Information Gain (IG) as shown in the equation 4.1, where k is the number of partitions of the attribute selected, n_i is the number of records in the partition i.

IG measures the reduction in Entropy after the split, in other words, this let us know how much information we gained by doing the split using a x attribute. The attribute with highest information gain is chosen to make the decision. This process is applied recursively to the child nodes. A node stops the splitting when all the samples are of the same class or when the splitting does not improve the predictive power of the model (Quinlan, 1993).

$$GAIN_{split} = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)) \tag{4.1}$$

4.2.2 Random Forest

The Random Forest algorithm is an ensemble learning algorithm that uses multiple decision trees developed by Tin Kam Ho (1995). The idea is to build lots of Trees in such a way to make the correlation between the Trees smaller.

The Random Forest starts selecting random samples from the training set, this selection is known as *bagging*. The idea behind bagging is the combination of learning models increases the classification accuracy. Then is selecting randomly k attributes from a total m of attributes, this process usually called *feature bagging* or *subspace sampling* (Flach, 2012), typically for classification problem is used $m = \sqrt{k}$. With these sample of attributes and subset of samples is created a decision tree.

This process is repeated T times to generate T decision trees. When a sample is going to be evaluated, it is evaluated by all the T trees, and the final classification is a result of the majority class predicted by the ensemble. A graphic representation of this process is shown in Fig 4.3.

4.2.3 Multilayer Perceptron

Multilayer Perceptron (MLP) algorithm is an evolution of the perceptron algorithm that is capable of learning only linearly separable data. MLP has a more complex structure than the perceptron. The topology of an MLP is shown in Figure 4.4.

The network can contain many intermediary layers between the Input Layer and Output Layer, known as Hidden Layers.

In the feed-forward neural network only, there can be connections only between nodes with the next layer. Similarly to perceptron, MLP has activation functions (i.e., linear, sigmoid and hyperbolic tangent functions) that allow to nodes in the hidden and output layer to produce non-linear outputs.

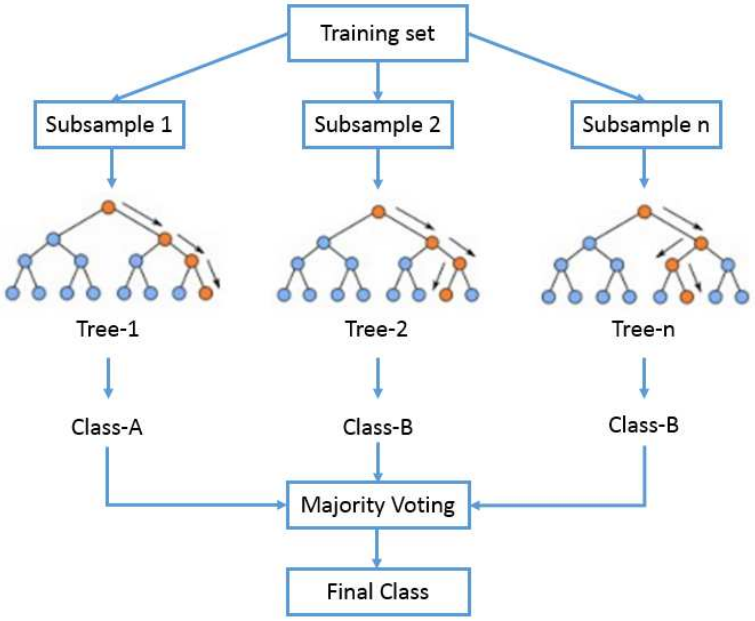


Figure 4.3. Random Forest.

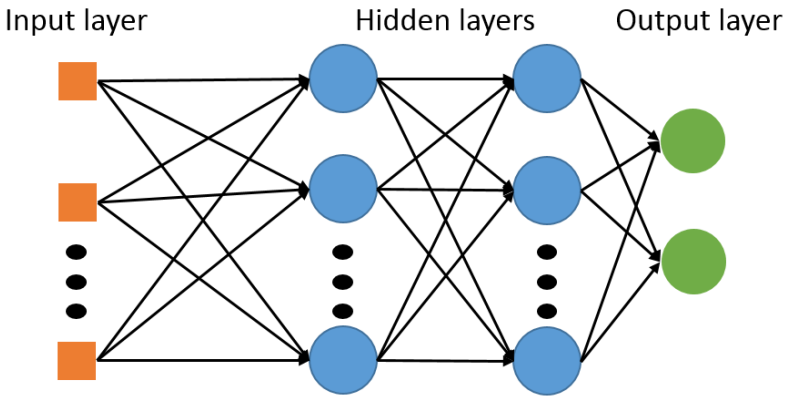


Figure 4.4. Multilayer Perceptron topology.

To adjust the weights, this model applies a technique called *back-propagation*. This technique uses the difference observed between the output result and the expected result from a layer, to generate an error value and uses this value to update the weights to the previous layer. It means the weights are updated in backward direction through the network. This process is repeated until the result is close to the expected output value (Tan et al., 2006).

4.2.4 Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm, for classification of linear and nonlinear data. This method aims to construct a hyperplane or a group of hyperplanes that differentiate two classes very well.

SVM does this process using other two hyperplanes on either side of the separating hyperplane, in a way to maximize the margin, i.e., the distance between the two extra hyperplanes. Defining the separating hyperplane this way tends to avoid overfitting. The points that lie in the margins are known as support vectors. The solution is a linear representation of these points. This can be observed in Figure 4.5.

When SVM deals with a non-linear classification problem, it uses a technique called the kernel trick. These functions map a low dimensional input space into a high dimensional feature space, so that the classes become linearly separable (Tan et al., 2006).

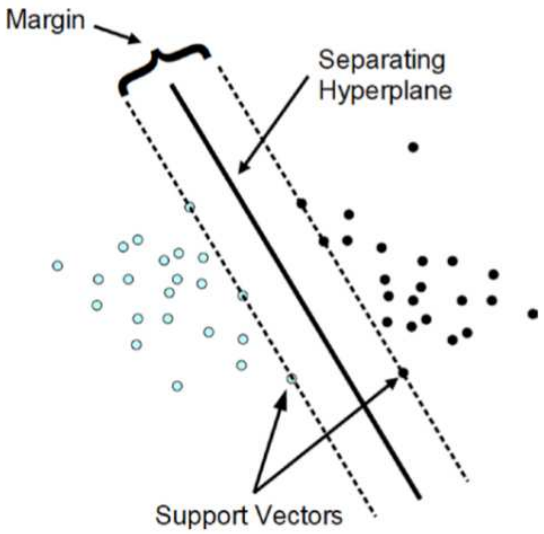


Figure 4.5. SVM classification example. Source (Oommen et al., 2008)

4.3 Performance Assessment

In this section, we describe the measures used in this work to evaluate the quality of the prediction models.

4.3.1 Cross Validation

Cross-validation is a method to estimate their performance of a prediction model, this kind of method divides into two segments the data, i.e., a training set and testing set to validate the model.

In this study, we used a K-Fold cross-validation, this method divides the samples into K equal parts, using $K-1$ parts to be the training set and 1 part to be the testing. This process is repeated until all the parts are used as a testing set. At the end Accuracy, and other statistical measures, of the model, is estimated by averaging the accuracies in all K cases (Refaeilzadeh et al., 2011).

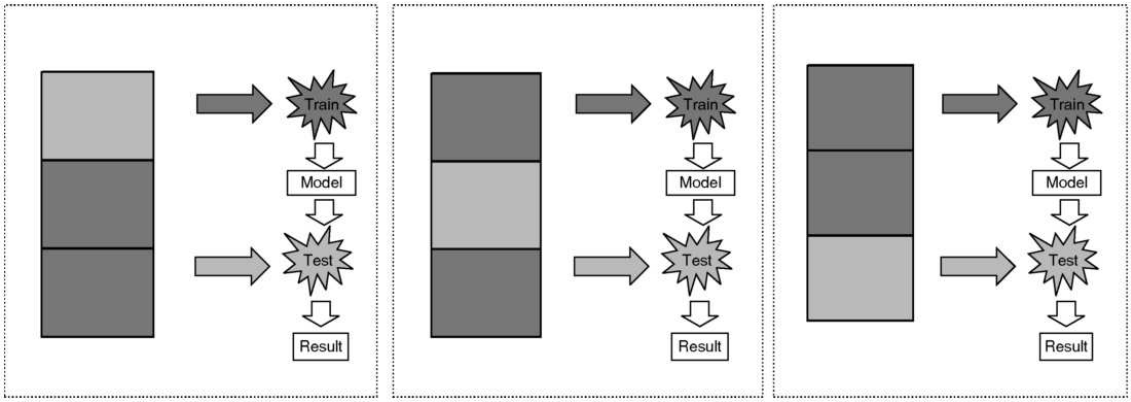


Figure 4.6. Procedure of three-fold cross-validation. Source (Refaeilzadeh et al., 2011)

4.3.2 Confusion Matrix and Evaluation Measures

The Confusion Matrix is a table commonly used in machine learning to evaluate a classifier's quality. It allows seeing the "Actual Class" versus "Predicted Class" from the predictive model as shown in Figure 4.7.

In the confusion matrix, the True Positives (TP) are all the samples from the positive class correctly classified as positive. Similarly, the True Negatives (TN) are all the samples from the negative class correctly classified as negative class. In contrast, the False Positives (FP) are all the samples from the negative class incorrectly classified as positives class, and the False Negatives (FN) are all the samples from the positive class incorrectly classified as negative class.

Following, we describe some statistic measures that can be calculated from the confusion matrix:

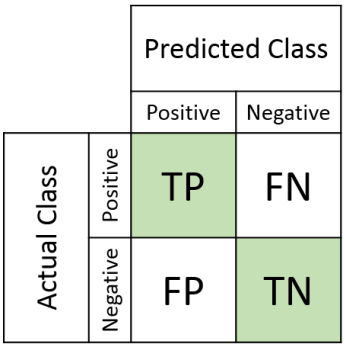


Figure 4.7. Confusion Matrix.

Accuracy (ACC): It measures the proportion of samples classified correctly with respect to the total of samples in the dataset. It can be stated as the equation 4.2.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

Precision (Pr): It measures the proportion of samples correctly classified as positive with respect to all the samples classified as positive. It can be stated as the equation 4.3.

$$Pr = \frac{TP}{TP + FP} \tag{4.3}$$

Sensitivity or Recall (Sn): Also called true positive rate or probability of detection, it measures the proportion of samples classified as positives. It can be stated as the equation 4.4.

$$Sn = \frac{TP}{TP + FN} \tag{4.4}$$

Specificity (Sp): Also called true negative rate, it measures the proportion of samples classified as negatives. It can be stated as the equation 4.5.

$$Sp = \frac{TN}{TN + FP} \tag{4.5}$$

F-measure or F1 score (F1): It is the harmonic mean of the Precision and Sen-

sitivity. It can be stated as the equation 4.6.

$$F1 = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{4.6}$$

Matthews correlation coefficient (MCC): It is a correlation coefficient between the actual and predicted classes in a binary classification; it returns a value between -1 and $+1$. A value of $+1$ represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and actual. It can be stated as the equation 4.7.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)))}} \tag{4.7}$$

4.3.3 Area under the ROC curve

Additionally, the Area under the ROC curve (AUROC) was used. The ROC curve is created by plotting the Sensitivity (Sn) also known as true positive rate versus the false positive rate (FPR) and can be calculated (1-specificity), to illustrate the diagnostic performance of a binary classifier as its discrimination threshold is varied. A good classification model shows a curve above and distant to the diagonal line, leading to an area under the curve close to 1.

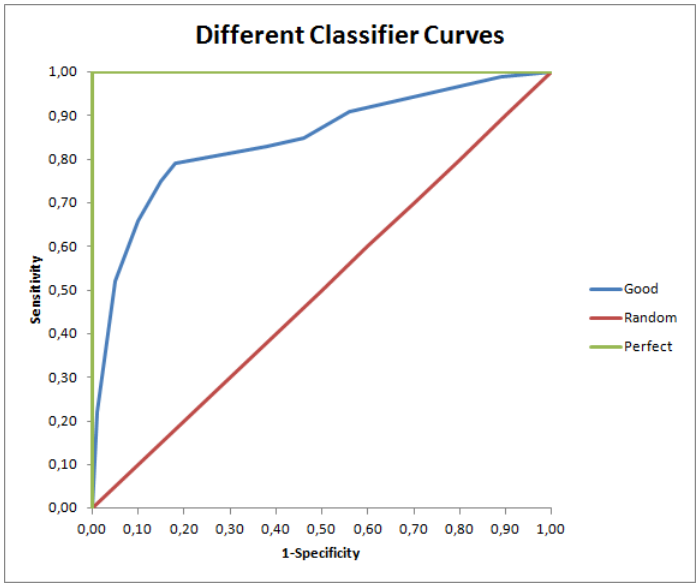


Figure 4.8. Area under the ROC curve.

As shown in the Figure 4.8, the classifier represented by the line in blue is

a good classifier. A perfect classification is represented by the green line, and a random classification is represented by the red line.

4.3.4 Feature Selection

Feature Selection methods make an automatic selection of the most relevant attributes in the training set to use in the model construction. These methods let us identify and remove unneeded attributes that do not contribute to the accuracy of the predictive model or even decrease the accuracy of it.

There exist two types of feature selection: (i) Filter Methods: These methods apply a statistical measure to assign a score to each attribute. These methods consider each feature independently. (ii) Wrapper Methods: These methods consider the feature selection as a search problem, where multiple combinations of sub-set of features are evaluated using a specific machine learning algorithm (Flach, 2012).

4.3.4.1 Information Gain

The information gain is part of the filter methods. This method is used in many decision trees algorithms (e.g., ID3, C4.5, etc.) for the splitting process. This method measures the attribute's information gain with respect to the class (J.Han, 2012).

The first step is to calculate the entropy of the class, defined as follows:

$$Entropy(D) = -\sum_{j=1} p(j|D) log_2 p(j|D) \tag{4.8}$$

Where D is the dataset, $p(j|D)$ is the relative frequency of class j in the dataset D. The entropy measures the homogeneity; when the samples are equally distributed among all the classes means less information, the worst scenario is the value of 1 (i.e., totally random), in contrast, if all the samples correspond to a single class imply most information, the value of the entropy is 0 (i.e., perfectly classified) (Tan et al., 2006).

The Information Gain is calculated as the difference between the prior entropy of classes and posterior entropy. As shown in the equation 4.9.

$$GAIN = Entropy(D) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)) \tag{4.9}$$

Where k is the number of partitions of the attribute selected, n_i is the number of records in the partition i.

4.3.4.2 Correlation-based Feature Selection

The Correlation-based Feature Selection (CFS) is another filter method. The CFS aims to select subsets of features highly correlated with the class and minimally correlated between them. The merit of a subset s containing k features is calculated with the Equation 4.10, where $\overline{r_{cf}}$ represents the average value of the features-class correlation and $\overline{r_{ff}}$ represents the average value of the feature-feature correlation (Tan et al., 2006).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{4.10}$$

4.3.4.3 Wrapper Selection method

The Wrapper Selection method finds the best subset of features to be used in a specific machine learning algorithm, evaluating this subset of features with a specific measure (i.e., accuracy, AUROC, F1, etc.) applying in the most of cases 5-fold cross-validation (Kohavi and John, 1997). This method is primarily a search problem. To do the search process, different search algorithms can be used, e.g., Forward selection. It starts with an empty set of features and adds a feature one at a time, as long as the model is improved. Backward elimination, in turn, starts with the complete set of features and enhance the model removing features one at a time (Flach, 2012). The cycle of this method is shown in figure 4.9.

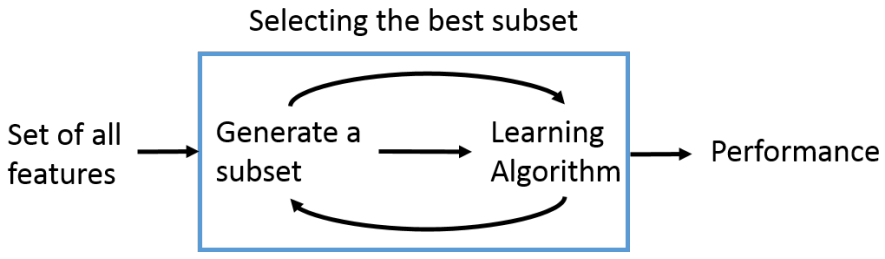


Figure 4.9. Wrapper Selection process.

Chapter 5

Methods

5.1 Dataset Generation

The main problem in this study is the information available. Since from the first discovery of a sRNA in *Escherichia coli*, only a limited number of sRNA were identified among bacteria. Even with the new sRNAs discovered in the last years, by new computational methods and high-throughput techniques such as genomic tiling microarrays and deep sequencing, the Databases with information about sRNA are far from desirable. The Databases specific for sRNA are: (i) sRNAMap (Huang et al., 2009), is an old database from 2009 with a lack of updates from this creation. (ii) sRNAdb (Pischimarov et al., 2012) a better database from 2012, also without updates, and (iii) BSRD (Li et al., 2013) the current referent database of sRNA. Despite others important sources of sRNA like Rfam (Nawrocki et al., 2015) that contains only few sRNAs families. Also, exist other good sources specific for *E. coli K12* like RegulonDB (Zhou and Rudd, 2013), Ecocyc (Keseler et al., 2013), and Ecogene (Zhou and Rudd, 2013).

5.1.1 Training Set

In this study, we used information from BSRD (Li et al., 2013) because this database contains over nine times the number of sRNA experimentally validated than any other database. BSRD (Li et al., 2013) obtained their data from many databases such as Rfam (Nawrocki et al., 2015), sRNAMap (Huang et al., 2009), RegulonDB (Zhou and Rudd, 2013), literature mining from PubMed and manually extraction from many relevant articles, that turns BSRD the sRNA database with the most updated information.

The training set is composed of two classes (i.e., positive and negative class) with the same number of sequences, so as to avoid an imbalanced distribution that can influence the performance of the predictive model.

The positive dataset: BSRD (Li et al., 2013) allows the download of all the experimentally validated sequences. The total number of these sequences are 897, among 64 bacteria sub-strains. We apply two filters to this information. (i) Filter by its size, selecting only sequences with a size between 50 - 500 nt following the information in the literature about sRNAs. (ii) Filter by their type, there are four types of sRNAs in BSRD (Li et al., 2013) (i.e., trans-encoded antisense, cis-encoded antisense, protein binding and regulatory element). We select only trans-encoded antisense, and cis-encoded antisense as those are the focus of this study and because they have more sequences validated as well as more information about it. Furthermore, there are only 32 non-redundant sequences of the other two types (i.e., 25 regulatory elements and 7 protein binding). After applying these two filters, we obtained 802 sequences of sRNA.

To remove redundant sequences, we used CD-HIT (Li and Godzik, 2006) to eliminate sequences with similarities above 90% and obtained the most representative sequences for our training set, leading to 720 sequences.

For evaluation purposes, we removed all the sequences similar or belonging to any of the sequences contained in our testing sets applying BLAST (Altschul et al., 1990). Thus, our final positive data set is constituted by 576 non-redundant sequences of sRNAs, labeled as "sRNA".

The negative dataset: In the literature, there are many approaches to build the negatives sets (e.g., using CDS sequences, a shuffle of CDS sequences, a shuffle of intergenic regions, etc.). We follow the same approach of many prediction tools of ncRNA, which consists of shuffling a real sequence while preserving the mono-nucleotide and di-nucleotide frequencies. We applied this process to each one of the sequences from the positive set using the tool uShuffle (Jiang et al., 2008) that does 1000 permutations to obtains the negative sequence. Due to the preserving, the di-nucleotide frequencies of the sequences comprise a robust method for the comparison of both sets (Clote et al., 2005).

This process led to the same quantity of sequences of the positive dataset, labeled as "nonsRNA".

5.1.2 Testing Sets

We used many testing set to evaluate our predictive model. The main comparison of this study is with the approach of Barman et al. (2017). Currently, it is the only ab initio tool specific for sRNA prediction in bacteria and follows a methodology similar to Tran et al. (2009) as well as ours.

5.1.2.1 Multi-Tool Testing Set

This testing set was built as a training set of sRNA prediction approach by Arnedo et al. (2014). This prediction approach applies a multi-objective method using results of many prediction tools of homology approach (i.e., RNAz2 (Gruber et al., 2010), vsFold (Dawson et al., 2007), Alifoldz (Washietl and Hofacker, 2004), dynalign (Mathews and Turner, 2002), QRNA (Rivas and Eddy, 2001), MSRAi (Coventry et al., 2004), zMFold (Zuker, 2003)) to obatin a final identification.

This set is composed of a positive set of 182 sequences of sRNAs from *Salmonella enterica* serovar Typhimurium LT2 (SLT2) and a negative set composed of non-sRNA sequences, generating two random sequences by each sequence from the positive set. To obtain these random sequences Barman et al. (2017) used shuffleseq (Rice et al., 2000) to shuffle ten times the complete genome of SLT2 preserving the mono-nucleotide composition and also used the real coordinate of the sRNAs to extract random sequences of the same size. Furthermore, this set was used by Barman et al. (2017) as a training set of their model, and they used it to make a comparison with the results obtained by Arnedo et al. (2014).

There are few differences between the sets of Barman et al. (2017) and Arnedo et al. (2014). Barman et al. (2017) remove 11 redundant sequences of sRNAs and duplicate the number of sequences in the negative set. In our case, we used in our positive set the same sequences published by (Barman et al., 2017), the only difference between is the removal of 11 sequences that exceed the size of 500 nt. Barman et al. (2017) describe the use of a negative set of twice the size of the positive set but do not specify which of the ten published negative sets were used to build it. In our case, we built five negative sets, as a result of joining Negative_set1 with Negative_set2, Negative_set3 with Negative_set4, Negative_set5 with Negative_set6, Negative_set7 with Negative_set8 and Negative_set9 with Negative_set10. As a result, we built five testing sets with same positive set and a negative set with each one of the negative sets created from Barman et al. (2017).

The importance of this testing set lies in the evaluation of multiple approaches (even with homology approaches), perhaps it cannot be accurate or fair at all, but

let us obtain an overview of the robustness of the methods in the identification of the same sequences of sRNAs.

5.1.2.2 Bacterial Testing Set

We select a group of four bacterial genomes by their medical importance and the quantity of sRNA available to make a group of testing sets. All of them are pathogenic bacteria that cause several infections in humans. The positive sets were made of sRNAs obtained from BSRD (Li et al., 2013) and the negative set was made creating two shuffled sequences preserving the di-nucleotide frequencies of each sequence in the positive dataset, following the approach of Barman et al. (2017) so as to obtain the double of negative sequences than the positive set. An important difference from Barman et al. (2017) is the inclusion of gram-positive bacteria in our tests.

Bacteria strain	**Type**	**Nº Positive Class**	**Nº Negative Class**
Listeria monocytogenes EGD-e AL591824.1	Gram-positive	120	240
Streptococcus agalactiae NEM316 AL732656.1	Gram-positive	36	72
Escherichia coli K12 MG1655 U00096.3	Gram-negative	93	186
Salmonella enterica senovar Typ. LT2 AE006468.1	Gram-negative	118	236

Table 5.1. Bacterial Testing Sets.

The purpose of these testing sets is to evaluate in different species of bacteria the single sequence sRNA classification.

5.1.2.3 Escherichia coli K12 sRNA collection

The main subject to evaluate the Genome-wide prediction was *Escherichia coli* K12; this organism is frequently used as a model organism in microbiology studies. It was also used by Tran et al. (2009) and Barman et al. (2017) to test their predictive models in Genome-wide prediction.

In this study, we focused on collecting as many sRNAs as possible from *Escherichia coli* K12 substrain MG1655. The advantage of using this strain is that there is a lot of documentation available as well as databases specific of their sRNAs.

To avoid non-validated sequences or without trustful sources, we made the collection with the following procedures. First, we selected all the sequences of sRNAs from Ecogene (Zhou and Rudd, 2013). This project is one of the best sources of genomic and proteomic information about *Escherichia coli* K12 MG1655. Additionally, it was the main source of the GenBank file from NCBI, which turns this

database into the most trustworthy source of sRNAs. A total of 63 sRNAs sequences were extracted from Ecogene (Zhou and Rudd, 2013). Second, we collected sRNAs from others databases including Rfam (Nawrocki et al., 2015), Ecocyc (Keseler et al., 2013), BSRD (Li et al., 2013) and RegulonDB (Gama-Castro et al., 2016). To preserve a good data quality between these databases, we selected only sequences experimentally validated and at least present in more than one database. As a result, we obtained 30 sequences different from Ecogene (Zhou and Rudd, 2013), leading to a total of 93 sRNAs, as shown in Table 5.2.

Database	**Nº Sequences**	**Nº Selected**
Ecogene	63	63
RegulonDB	125	91
Ecocyc	57	55
Rfam	156	43
BSRD	108	101
Final unified selection		93

Table 5.2. *Escherichia coli* K12 database sRNA selection.

5.2 Bacterial Genomes

In this study, we used two different genomes, one from *Escherichia coli* K12 MG1655 and the other from *Salmonella enterica* serovar Typhimurium LT2, for genome-wide prediction evaluation. Both of these bacteria were used in many different experiments of previously proposed methods, which turns them into perfect candidates for applying our analyses and tests.

5.2.1 Escherichia coli K12 MG1655

The ab initio approaches by Tran et al. (2009) and Barman et al. (2017) use the same substrain i.e., MG1655. Tran et al. (2009) used the version NC_000913.2, while Barman et al. (2017) used the latest version NC_000913.3, which is also the version used in this work.

RefSeq	**INSDC**	**Size (Mb)**	**GC%**	**Protein**	**rRNA**	**tRNA**	**Other RNA**	**Gene**	**Pseudogene**
NC_000913.3	U00096.3	4.64	50.8	4,140	22	89	67	4,498	184

Table 5.3. Genome *Escherichia coli* K12 substrain MG1655.

To extract more information about the selected sRNAs, we apply some analysis about the position of the sRNAs sequences in the genome.

Statistics	
Average size	136.94
Max. size	436
Min. size	30
Sense overlapping	12
Antisense overlapping	21
Without overlapping	60
N° Forward strand	53
N° Reverse strand	40

Table 5.4. *Escherichia coli* K12 sRNAs Genome Analysis.

5.2.2 Salmonella enterica serovar Typhimurium LT2

We used the substrain of SLT2, i.e., AE006468.1, due to preserving the same coordinates from BSRD (Li et al., 2013), furthermore it is the same version used by Barman et al. (2017).

RefSeq	**INSDC**	**Size (Mb)**	**GC%**	**Protein**	**rRNA**	**tRNA**	**Other RNA**	**Gene**	**Pseudogene**
NC_003197.1	AE006468.1	4.86	50.8	4,452	22	85	10	4,608	41

Table 5.5. Genome SLT2.

As performed for *Escherichia coli* K12, we extracted more information about the sRNAs by their position in the genome and their sizes.

Statistics	
Average size	163.25
Max. size	1236
Min. size	42
Sense overlapping	23
Antisense overlapping	11
Without overlapping	86
N° Forward strand	72
N° Reverse strand	48

Table 5.6. SLT2 sRNAs Genome Analysis.

5.3 Feature Collection

We selected an extensive collection of attributes from three ncRNA prediction tools, all of them with an ab initio approach. A brief description of each one is provided below:

Tran et al. (2009): This is the first ab initio tool for computational identification of regulatory ncRNA genes in prokaryotic genomes. Tran et al. (2009) use three types of features, i.e., *Folding statistics*, *Ensemble statistics* and *Structural statistics*. The features used in this approach were used for many other tools, the work of Tran et al. (2009) is the most valuable source due to the available documentation of each one of their features.

RNAcon: This method predicts and classifies ncRNAs. The approach of Panwar et al. (2014) is composed of two predictors. The first one filters all the possible CDS using *Primary sequence attributes*. The second one classifies families of ncRNA using *Graph properties from Structural Information*, this last type of attribute was included in our set of features.

HLRF: The prediction tool has the largest quantity of features. Internally it is composed of multiple prediction tools that extract many features prior to machine learning model building. The aim of HLRF (Lertampaiporn et al., 2014) is to predict lncRNAs, but their results include many other families of ncRNAs. This tool has 300 different features, but we only selected 137 that does not make it from any homology approach. The problem with this tool is the lack of documentation of each of their features. Despite of this fact, we used the description in their paper to decide about their importance.

In total, this work collects 264 different features, divided into eight categories as summarized in Table 5.7.

5.3.1 Primary Sequences attributes

This type of features correspond to all the features of the primary structure of the RNA, explained in the Chapter 3. These attributes are the frequency of K nucleotides in the sequence, in this study we included mono, di, tri and tetra-nucleotide frequencies. Most of the studies use at least mono-nucleotide and di-nucleotide frequencies, but HLRF (Lertampaiporn et al., 2014) select a group of 20 frequencies of tetra-nucleotides and Barman et al. (2017) use only tri-nucleotides

Feature Group	Nº of features
Primary sequence attribute	104
Folding Statistics	32
Structural Statistics	20
Graph Properties from Structural Information	20
Base-pair features	28
Triplet structure-sequence features	32
Structural Robustness features (SC – derived features)	11
Ensemble Statistics	17
Total	264

Table 5.7. Summary of the 264 ncRNA features.

frequencies as features in their predictive model, that is the reason because we included the 64 tri-nucleotides frequencies in this set.

5.3.2 Ensemble Statistics

As explained in the Section 3.3 all the transcripts of sRNA adopt a secondary structure, but this structure is not unique by sequence.

The set of all the possible secondary structures adopted by RNA is called Ensemble of secondary structures. This ensemble has two principal structures: (i) The centroid: which corresponds to the structure more similar to all the structures in the ensemble. (ii) The minimum free energy (MFE): which is the structure with the lowest value of free energy (i.e., the energy released by folding a completely unfolded RNA molecule), the purpose of calculating the mfe is due to lower the free energy more probably the secondary structure to be adopted.

5.3.3 Folding Statistics

These measures are calculated comparing the mfe between the real sequence and a group 1000 aleatory sequences of the same size and same di-nucleotide frequencies. The main two features in this set are:

Z-score: This value corresponds, to the standard deviation that the real sequence deviates from the mean of the shuffled sequences. The z-score is calculated by the formula below:

$$Zscore(x) = \frac{\text{mfe}(x) - \mu}{\sigma} \tag{5.1}$$

Where μ and σ represent the mean and the standard deviation respectively of the set of 1000 random sequences.

Partition value (pvalue): Correspond, to the number of random sequences n that have a mfe lower than the original sequence normalized by the number of the total of the random sequences N. The p-value is calculated with the formula below:

$$pvalue(x) = \frac{n}{N} \tag{5.2}$$

5.3.4 Structural Statistics

All these features were presented by Tran et al. (2009). These features are obtained from the mfe. This set includes the number of structural elements present in the structure, number of nucleotides, and the average length of each structural element, i.e., loop, internal-loop, multi-loop, steam, and bulge.

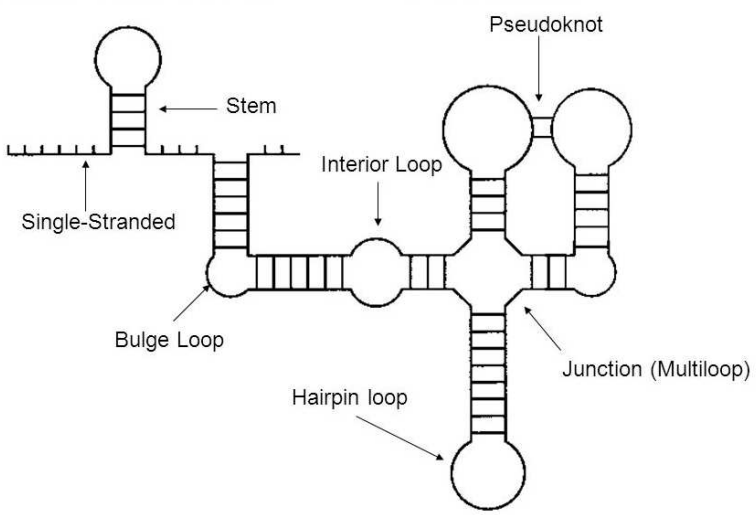


Figure 5.1. Secondary structure elements.

5.3.5 Graph Properties from Structural Information

All these types of features come from the tool RNAcon (Panwar et al., 2014); this tool first uses IPknot (Sato et al., 2011) software to predict some specific element from secondary structure then it used in the igraph (Csárdi and Nepusz, 2006) R package to calculate all the 20 different graph properties of all the predicted

structures. The nodes are represented by the nucleotides and edges are the bonds between the nucleotides as shown in Figure 5.2.

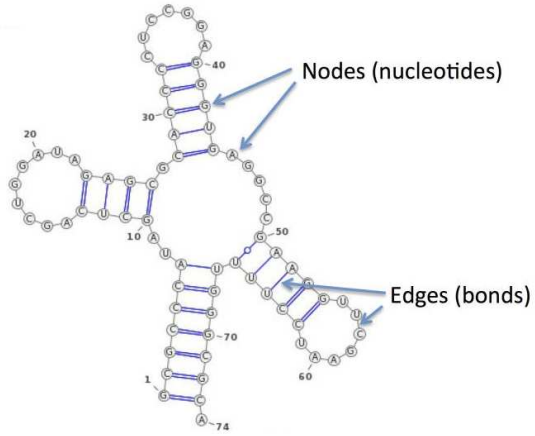


Figure 5.2. Secondary structure RNAcon. Source (Panwar et al., 2014)

5.3.6 Base-pair distance features

The base pair is the result unit of pairing two nucleotides. The possible base pairing are A with U and C and G, due to the number of hydrogen bonds between the nucleotides (i.e., two between A and U; three between C and G), this ability to form this bonds makes the bp more stable. These features are from HLRF (Lertampaiporn et al., 2014), all these features are calculated from the secondary structure predicted, e.g., the total of base pairs present in the structure, the number of base pairs AU and GC, the normalized possible bp by the length of the sequence, the structural base-pair distance. Also, HLRF (Lertampaiporn et al., 2014) makes some statistics of bp in a structural element such as steam and loops.

5.3.7 Triplet structure-sequence features

This type of feature was developed by Xue et al. (2005) to miRNA precursors detection. It focuses on information of every three adjacent nucleotides in the secondary structure predicted by RNAfold (Lorenz et al., 2011). There exist 8 possible structures "(((", "((.", "(..", "(.(", ".((", ".(.", "..(" and "...". Where brackets "(" represent a paired nucleotide and dot "." represented unpaired nucleotide. Considering the middle nucleotide among the adjacent nucleotide which can be A, C, G or U, we obtained 32 possible structure combinations. The final feature value is the number of appearances of this structure element, which we denote as "U(((", "A(((", etc.

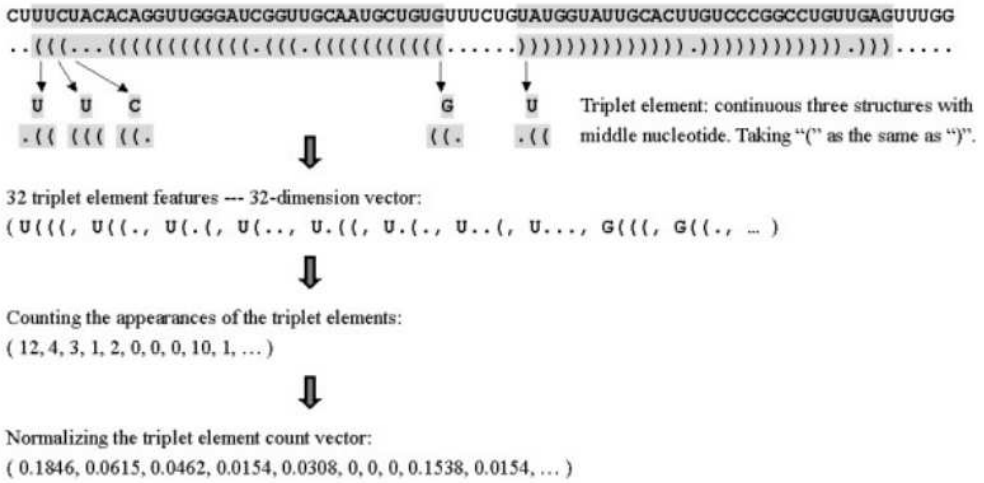


Figure 5.3. Triplet sequences feature extraction process. Source (Xue et al., 2005)

5.3.8 Structural Robustness features

This type of feature comes from the study by Lee and Kim (2008). This study proposed a new measure for miRNA prediction called Self-contained index score (SC). The SC is obtained following this steps: for each RNA sequence w of length L folding into a particular mfe secondary structure, the true sequence w is embedded in the middle of two random sequences x and y of equal length, obtaining a sequence xwy of length $3L$. This new sequence is folded and measures the proportion of the original structure of w preserved in the larger structure. This process is repeated 1000 times to generate a single statistic with a value between 0.0 to 1.0, with 1.00 represent a maximum degree of self-containment. HLRF (Lertampaiporn et al., 2014) use this feature and combine with other measures to create new features. An example of these process is shown in Figure 5.4.

5.4 Feature Selection

To select the best set of features from our large set of features collected, we applied different machine learning methods to do this task. These methods are useful to identify and remove the irrelevant and redundant features in our training set. This kind of features does not contribute to the classification process and can decrease the accuracy of our model.

Before applying the feature selection algorithms, we evaluate each set of types of features in the machine learning algorithms selected, to measure its performance

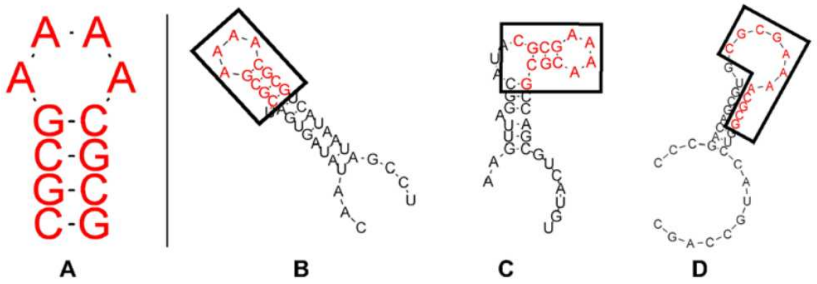


Figure 5.4. Example of varying degrees of structure conservation, (A) The original RNA folding into a hairpin. The secondary structures (B), (C) and (D) is the result of putting the original sequence in the middle of two other sequences; (B) Perfect conservation, (C) Lose one base pair, (D) Complete disruption of the structure. Source (Lee and Kim, 2008)

and determinate which type of feature contributes more to the classification. Furthermore, we test the tri-nucleotide features used by Barman et al. (2017) to evaluate their performance. The algorithms selected are: SVM without kernel (SVM_P), SVM with RBF kernel (SVM_RBF), Multilayer Perceptron (MLP), Random Forest (RF), C4.5 and Naive Bayes (NB).

Type	**Algorithm**	**ACC** (%)	**Pr** (%)	**Sn** (%)	**SPC** (%)	**F1** (%)	**MCC** (%)	**AUROC**	**Nº Att.**
Primary seq.	SVM_P	61.7	62.5	58.5	64.9	60.4	23.5	0.617	84
	SVM_RBF	55.9	59.9	35.8	76.0	44.8	12.9	0.559	
	MLP	58.5	57.8	62.8	54.2	60.2	17.1	0.608	
	RF	54.0	53.9	55.7	52.3	54.8	8.0	0.560	
	C4.5	54.0	54.1	52.3	55.7	53.2	8.0	0.553	
	NB	53.9	54.3	49.7	58.2	51.9	7.8	0.553	
Tri-nt	SVM_P	59.7	60.0	58.3	61.1	59.2	19.5	0.597	64
	SVM_RBF	56.5	58.4	45.1	67.9	50.9	13.4	0.565	
	MLP	57.6	58.1	54.7	60.6	56.4	15.3	0.599	
	RF	52.3	52.1	56.1	48.4	54.0	4.5	0.525	
	C4.5	53.0	53.6	43.4	62.5	48.0	6.0	0.534	
	NB	54.0	54.3	50.3	57.6	52.3	8.0	0.535	

Table 5.8. Summary of results by type sets.

Type	Algorithm	ACC (%)	Pr (%)	Sn (%)	SPC (%)	F1 (%)	MCC (%)	AUROC	Nº Att.
Tetra-nt	SVM_P	56.2	57.3	48.6	63.7	52.6	12.5	0.562	20
	SVM_RBF	52.1	56.5	18.2	85.9	27.6	5.7	0.521	
	MLP	55.0	55.5	50.9	59.2	53.1	10.1	0.579	
	RF	54.4	54.2	56.9	51.9	55.5	8.9	0.552	
	C4.5	50.7	50.7	50.3	51.0	50.5	1.4	0.510	
	NB	54.3	54.7	50.2	58.5	52.4	8.7	0.552	
Folding Stat.	SVM_P	76.2	77.8	73.4	79.0	75.5	52.5	0.762	31
	SVM_RBF	74.5	72.9	78.0	71.0	75.3	49.1	0.745	
	MLP	71.9	73.2	69.1	74.7	71.1	43.8	0.784	
	RF	76.0	81.1	67.7	84.2	73.8	52.6	0.820	
	C4.5	75.7	86.1	61.3	90.1	71.6	53.7	0.769	
	NB	73.7	77.1	67.4	80.0	71.9	47.8	0.815	
Structural Stat.	SVM_P	68.7	73.6	58.2	79.2	65.0	38.2	0.687	20
	SVM_RBF	58.0	56.4	70.8	45.1	62.8	16.5	0.580	
	MLP	65.6	69.7	55.4	75.9	61.7	31.9	0.701	
	RF	67.6	69.3	63.2	72.0	66.1	35.4	0.728	
	C4.5	65.7	67.2	61.5	70.0	64.2	31.5	0.661	
	NB	61.4	60.3	66.7	56.1	6.3	22.9	0.684	
Graph Prop.	SVM_P	64.2	64.6	62.7	65.6	63.6	28.3	0.641	20
	SVM_RBF	63.0	61.2	71.0	55.0	65.8	26.4	0.630	
	MLP	62.9	63.7	59.7	66.0	61.6	25.7	0.665	
	RF	60.9	60.8	60.9	60.8	60.9	21.7	0.658	
	C4.5	63.1	64.8	57.3	68.9	60.8	26.4	0.633	
	NB	58.7	56.5	75.0	42.4	64.5	18.4	0.645	
Bp distance	SVM_P	70.7	73.7	64.2	77.1	68.6	41.7	0.707	29
	SVM_RBF	65.7	70.1	54.9	76.6	61.5	32.2	0.657	
	MLP	68.2	69.7	64.6	71.9	67.0	36.6	0.735	
	RF	66.1	67.1	63.0	69.1	65.0	32.2	0.728	
	C4.5	60.9	61.8	57.5	64.4	59.5	21.9	0.618	
	NB	64.7	67.0	57.8	71.5	62.1	29.6	0.693	
Triplet seq.	SVM_P	68.1	69.4	64.6	71.5	66.9	36.2	0.681	32
	SVM_RBF	66.8	71.3	56.1	77.4	62.8	34.3	0.668	
	MLP	61.3	61.5	60.2	62.3	60.9	22.6	0.663	
	RF	67.1	67.8	65.1	69.1	66.4	34.2	0.735	
	C4.5	63.6	64.9	59.4	67.9	62.0	27.4	0.639	
	NB	67.3	69.7	61.1	73.4	65.1	34.8	0.733	

Table 5.8. Summary of results by type sets.

Type	Algorithm	ACC (%)	Pr (%)	Sn (%)	SPC (%)	F1 (%)	MCC (%)	AUROC	Nº Att.
SCI	SVM_P	71.9	89.1	49.8	93.9	63.9	48.7	0.719	11
	SVM_RBF	64.2	77.0	40.6	87.8	53.2	32.3	0.642	
	MLP	76.0	85.7	62.3	89.6	72.2	54.0	0.797	
	RF	72.7	76.8	65.1	80.4	70.5	46.0	0.786	
	C4.5	73.3	86.8	54.9	91.7	67.2	50.0	0.742	
	NB	68.8	78.0	52.3	85.2	62.6	39.7	0.736	
Ensemble Stat.	SVM_P	66.2	67.0	63.9	68.6	65.4	32.5	0.662	17
	SVM_RBF	65.0	65.0	64.9	65.1	65.0	30.0	0.650	
	MLP	62.7	63.1	60.9	64.4	62.0	25.4	0.666	
	RF	63.9	64.5	61.8	66.0	63.1	27.8	0.689	
	C4.5	62.4	64.0	56.6	68.2	60.1	25.0	0.648	
	NB	63.6	62.5	68.2	59.0	65.2	27.4	0.691	

Table 5.8. Summary of results by type sets.

As we can see in Table 5.8, the three worst set of attributes are the Tetra-nucleotide, Tri-nucleotide and the set composed of their two sets (i.e., Primary seq.). With these results, we want to prove that the use of primary sequences features alone is not enough to create a good prediction model of sRNA. This experiment uses our training set composed of many families of sRNAs and not only composed by families presented in a single bacteria substrain like other approaches, and conveniently tests their models in substrains phylogenetically close like Barman et al. (2017).

Other analysis from the Table 5.8 is the best set of features that are the Folding statistics and the SCI features. This can be evidence of the relevant information from the mfe, and the possible structure conservation (SCI) of the sRNA like the miRNA in eukaryotic cells.

To select the best attributes we used three different techniques (i.e., Information Gain, CSF, and Wrapper method).

5.4.1 Information Gain

We applied IG with 10-fold cross-validation, and then selected all the features with a score larger than 0.059 obtaining 34 features, this set of attributes was tested with the six algorithms selected.

As we can show in Table 5.9, the best model was from Random Forest algorithm.

Algorithm	ACC (%)	Pr (%)	Sn (%)	SPC (%)	F1 (%)	MCC (%)	AUROC	Nº Att.
SVM_P	76.48	79.0	72.0	80.9	75.4	53.2	0.765	34
SVM_RBF	74.65	73.3	77.6	71.7	75.4	49.4	0.747	
MLP	72.66	74.3	69.3	76.0	71.7	45.4	0.790	
RF	76.13	82.0	67.0	85.2	73.7	53.1	0.821	
C4.5	75.26	87.2	59.2	91.3	70.5	53.3	0.770	
NB	73.78	79.5	64.1	83.5	71.0	48.5	0.808	

Table 5.9. Information Gain results.

5.4.2 Correlation-based Feature Selection

We applied the Correlation-based Feature Selection (CFS) in our training set obtaining 19 relevant features. Then, we used the features in the six algorithms selected leading to the results shown in Table 5.10.

Algorithm	ACC (%)	Pr (%)	Sn (%)	SPC (%)	F1 (%)	MCC (%)	AUROC	Nº Att.
SVM_P	76.39	79.2	71.5	81.3	75.2	53.0	0.764	19
SVM_RBF	74.65	72.9	78.5	70.8	75.6	49.5	0.747	
MLP	72.48	74.3	68.8	76.2	71.4	45.1	0.789	
RF	75.43	80.3	67.4	83.5	73.3	51.5	0.828	
C4.5	76.30	86.3	62.5	90.1	72.5	54.7	0.774	
NB	76.39	84.2	64.9	87.8	73.3	54.2	0.841	

Table 5.10. CFS results.

5.4.3 Wrapper Selection Method

The final method applied was Wrapper. This procedure chooses the best sub-set of attributes tailored to a specific classification algorithm to obtain a significant improvement in one particular evaluation measure. In our case, we chose two relevant measures, i.e., F1 score and ROC curve using two different search methods (i.e., best first and greedy stepwise). Almost all these models overpass the results from IG and CFS methods. The criteria to choose the model was taking into account Precision, Sensitivity, SPC, F1, AUROC with Precision not lower than Sensitivity, due to the high FP rates in ab initio approaches. The selected model of this study was the Random Forest algorithm with the search method best first. This model has the best overall results with the low quantity of False Positives reflected in all our testing sets.

The 22 attributes from the selected model are described in the Table 5.12.

Algorithm	Evaluation measure	Search Method	ACC (%)	Pr (%)	Sn (%)	SPC (%)	F1 (%)	MCC (%)	AUROC	Nº Att.
SVM_P	AUROC	Best First	78.82	87.6	67.2	90.5	76.0	59.3	0.788	6
		GreedyStepwise	78.82	87.6	67.2	90.5	76.0	59.3	0.788	6
	F1	Best First	78.90	78.1	80.4	77.4	79.2	57.8	0.789	14
		GreedyStepwise	78.90	78.1	80.4	77.4	79.2	57.8	0.789	14
SVM_RBF	AUROC	Best First	74.21	71.0	81.9	66.5	76.1	49.0	0.742	6
		GreedyStepwise	74.21	71.0	81.9	66.5	76.1	49.0	0.742	6
	F1	Best First	73.61	69.3	84.7	62.5	76.3	48.4	0.736	8
		GreedyStepwise	73.61	69.3	84.7	62.5	76.3	48.4	0.736	8
MLP	AUROC	Best First	80.38	85.9	72.7	88.0	78.8	61.5	0.866	9
		GreedyStepwise	78.30	83.4	70.7	85.9	76.5	57.3	0.854	6
	F1	Best First	78.03	80.1	73.8	82.3	77.1	56.3	0.832	10
		GreedyStepwise	78.65	83.7	71.2	86.1	76.9	57.9	0.829	4
RF	AUROC	Best First	78.47	84.2	70.1	86.8	76.5	57.8	0.868	22
		GreedyStepwise	78.56	83.5	71.2	85.9	76.9	57.8	0.865	15
	F1	Best First	78.21	82.3	75.0	84.5	76.7	56.9	0.825	12
		GreedyStepwise	77.17	80.5	71.7	82.6	75.8	54.7	0.819	7
C4.5	AUROC	Best First	77.08	84.4	66.5	87.7	74.4	55.4	0.817	14
		GreedyStepwise	77.08	84.4	66.5	87.7	74.4	55.4	0.817	13
	F1	Best First	78.82	85.5	69.4	88.2	76.6	58.7	0.799	22
		GreedyStepwise	78.73	86.1	68.6	88.9	76.3	58.7	0.799	16
NB	AUROC	Best First	78.99	81.9	74.5	83.5	78.0	58.2	0.872	30
		GreedyStepwise	78.99	81.9	74.5	83.5	78.0	58.2	0.872	30
	F1	Best First	77.69	75.1	82.8	72.6	78.8	55.7	0.816	11
		GreedyStepwise	77.69	75.1	82.8	72.6	78.8	55.7	0.816	11

Table 5.11. Results from Wrapper method.

Attribute	Description
Prob	Probability that the MFE of the given RNA sequence is different from a distribution of MFE computed with random sequences
zG	The absolute value of the z-score (zG)
au/L	au is number of base A paired with base U in the secondary structure, normalized by length of the sequence
P_gc/bp	P_gc is the percentage of base G paired with base C in the secondary structure divided by the total number of base pairs in the secondary structure
SC/len	Self-contained index score (SCI) divided by the length of sequence
c3	Third structure combination with base C
c5	Fifth structure combination with base C
pairprob7	Pairing probability of C-U
pairprob10	Pairing probability of U-U
sumbits	# Count out-of-frame tripletts
AAAA	Frequency of tetranucleotide “AAA”
AGGA	Frequency of tetranucleotide “AAA”
stem_ave	Average length (nt) per stem
totalinternal_ave	Average length (nt) per all internal loops
wss	Measures the base-pair distance between the cluster centroid with all the structures within that cluster.
Variance_of_coreness	Variance of the k-core, that is the maximal sub-graph in which each vertex has at least degree k.
AAA	Frequency of trinucleotide “AAA”
ATT	Frequency of trinucleotide “ATT”
GCA	Frequency of trinucleotide “GCA”
GAT	Frequency of trinucleotide “GAT”
CCA	Frequency of trinucleotide “CCA”
CCC	Frequency of trinucleotide “CCC”

Table 5.12. Final selection of Attributes.

5.5 Genome-wide prediction

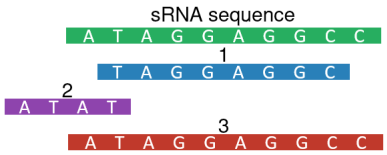
One of the primary focus of this work is to present an approach to the Genome-wide identification of sRNA in prokaryotes. We follow the same technique of window-slides used by Tran et al. (2009) and Barman et al. (2017). This technique consists, in dividing the genome into windows or slides of l size and advancing through the strand with s steps obtaining an overlapping o between the windows. A brief description of both methodologies is presented below:

Tran et al. (2009) use, three different window sizes, $l = 100, 120, 10$ because there are the three peaks of the ncRNA-length distribution in *E. coli*, the steps used was the half of the size of the window selected $s = l/2$; this process is applied in both strands of the genome. Tran et al. (2009) used a lot of filters to reduce their high quantity of putative sequences, such as sequence conservation, promoter and terminator information; and use a neural network to do the prediction process.

Barman et al. (2017) uses only intergenic regions from the forward strand of the genome. The size of the windows is 145 nt $l = 145$ because this is the median of the length of the sRNAs in BSRD (Li et al., 2013) with a step of 50 nt $s = 50$. Barman et al. (2017) do not take into account the sense of the sequences. They always use the coordinates in the forward strand. In other words, if the sRNA sequence is located in the reverse strand, the reverse complement is used.

A problem with previously proposed methods is the lack of clarity in describing what is considered a true positive. Tran et al. (2009) describes if there exists an overlapping with the real sRNA is considered a TP, but they do not specify how much of overlapping must exist. Barman et al. (2017) uses the window closer to the center of the sRNA without considering factors like the size of the sRNA.

This is a very important measure in Genome-wide prediction. In this study, we follow the method of Pena-Castillo et al. (2016) applied in a homology approach. This methodology uses two measures: (i) Minimum percentage overlap. It quantifies what percentage of the putative sequence is covered by the real sRNA. (ii) Minimum percentage reciprocal overlap. It quantifies what percentage of the real sRNA is covered by the putative sequence. Following this approach, the acceptable values to be considered as a True Positive (TP) sequence are 80% of minimum sequence overlap or 50% of reciprocal overlap. An example of these measures is shown in Figure 5.5.



Sequence	Min % sequence overlap	Min % reciprocal sequence overlap
1	100%	80%
2	50%	20%
3	100%	100%

Figure 5.5. Examples of True Positives in Genome-wide prediction. In the first case, the sequence overpass, the condition of the Minimum percentage overlap and has the minimum value to pass the Minimum percentage overlap. The second case, pass with the Minimum percentage overlap of 50% and does not pass the second measure. The last example is the perfect match with the real sRNA.

5.5.1 Proposed Framework

The main difference between our method and other approaches is the non-use of any experimental information about the genome, the only known information used are the coordinates of coding regions. The pipeline process of the framework is described in the next steps as well as in the Figure 5.6:

1. Input File: It reads a GenBank file of the prokaryotic genome to be analyzed.

2. Windows slide module: The both strands of the genome are divided into windows of the same size l, in our case of 145 nt with steps s of 50 nt or half of l.

3. CDS extraction: The coordinates of the coding sequences (CDS) are extracted from Genbank file.

4. Promoter extraction: Additionally to the Genbank file, it is extracted a complete sequence of the genome, and it is used with the BPROM (Solovyev, 2011) tool to predict all the σ^{70} promoters in both strands of the genome. Due to the limitations of this tool (i.e., the max size of sequence), each strand of the genome is divided into four parts of the same size. Next, it is applied the promoter prediction, and then the resultant coordinates are remapped in the original coordinates in the genome.

5. CDS filter: It is filtered all the windows with an overlapping over 110 nt with a CDS region in any strand, because the major part of the sRNA sequences with a CDS overlap only occurs in the start or end of the CDS with a size of 110 nt.

6. Promoter Filter: Then, it is extracted as a final test group all the windows that have one of these conditions: (a) If a predicted promoter is located 200 nt upstream the putative window. (b) If there exists an overlap with a predicted promoter region.

7. Feature Extraction: Once obtaining the final group of sequences, it is extracted all the 22 features used in our prediction model.

8. Prediction: In the last step, it is applied the prediction model to each sequence record to determinate if an sRNA is present or not in the sequence.

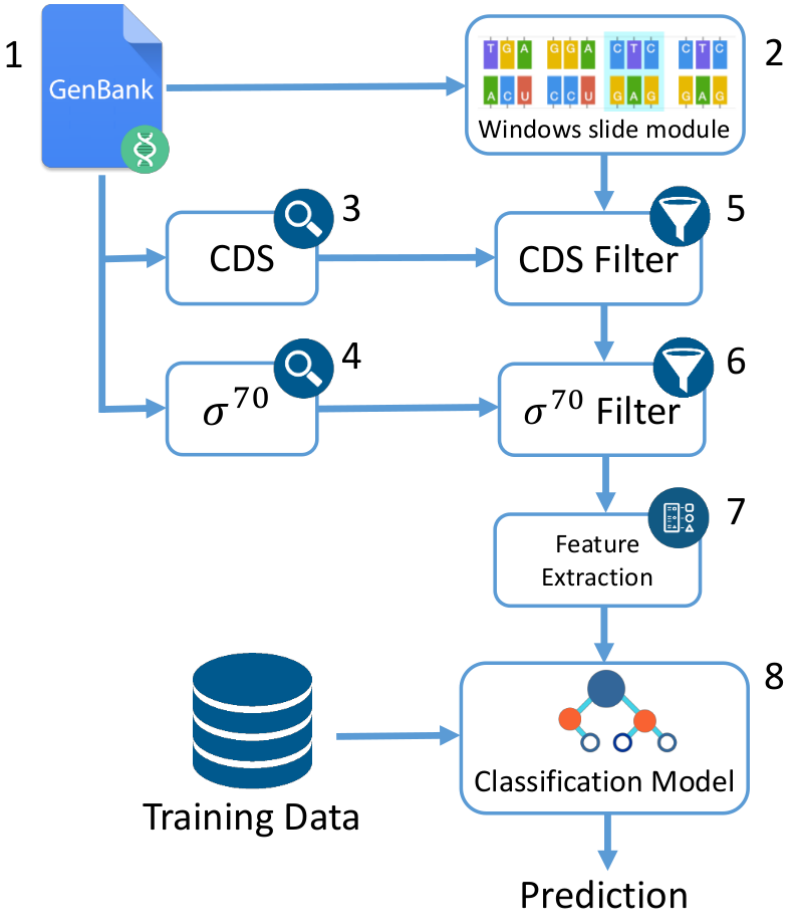


Figure 5.6. Genome-wide Prediction Framework.

Chapter 6

Results and Discussion

The prediction model was tested with two different methodologies. The first one is the single sequence prediction; in these tests, we use the testing-sets described in Section 5.1.2. The second one is the Genome-wide prediction of two bacteria substrain. In both methodologies, we put emphasis in comparison with the only ab initio whose the computational tool is available, i.e., Barman et al. (2017).

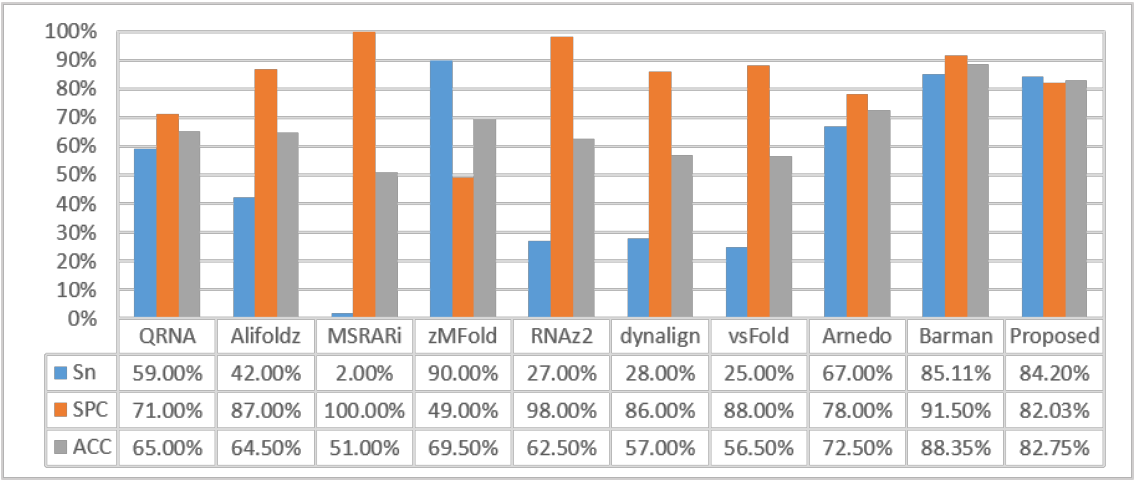
6.1 Single sequence Prediction Results

6.1.1 Multi-tool testing set

The first set to be evaluated is the Multi-tool testing set. This testing set made by (Arnedo et al., 2014) is possibly the most relevant among the testing sets, due to the comparisson of many homology tools i.e., QRNA (Rivas and Eddy, 2001), Alifoldz (Washietl and Hofacker, 2004), MSARi (Coventry et al., 2004), zMFold (Zuker, 2003), RNAZ 2.0 (Coventry et al., 2004), Dynalign (Mathews and Turner, 2002), vsFold (Dawson et al., 2007), Arnedo et al. (2014), and it was used by (Barman et al., 2017) to compare and build its prediction model.

The result of our prediction measures corresponds, to the average of 10-fold cross-validation of the five sets described in Subsection 5.1.2.1.

As shown in Figure 6.1, the proposed method has the best set of values only surpassed by Barman et al. (2017), and after us the approach from Arnedo et al. (2014). The comparison between different type of approaches can be unfair, but this test lets us obtain an overview of the robustness of our proposed method among others. Two important factors to be considered: (i) Our training set only contains two sequences from this testing set, unlike Barman et al. (2017) and Arnedo et al.



	QRNA	Alifoldz	MSRARi	zMFold	RNAz2	dynalign	vsFold	Arnedo	Barman	Proposed
Sn	59.00%	42.00%	2.00%	90.00%	27.00%	28.00%	25.00%	67.00%	85.11%	84.20%
SPC	71.00%	87.00%	100.00%	49.00%	98.00%	86.00%	88.00%	78.00%	91.50%	82.03%
ACC	65.00%	64.50%	51.00%	69.50%	62.50%	57.00%	56.50%	72.50%	88.35%	82.75%

Figure 6.1. Multi-tool testing set result chart.

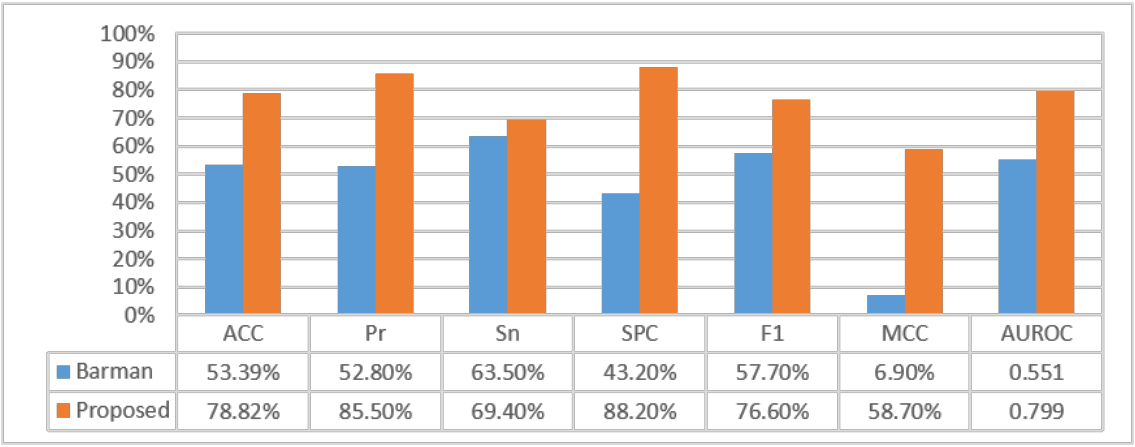
(2014) whose training sets are composed of all these sequences. (ii) The negative set used has a different approach compared with our case, because it obtains a random sequence from an entirely shuffle genome only preserving the mono-nucleotides frequencies. Nonetheless, our model demonstrated good results detecting this type of negative sequences.

To make a fair comparison with Barman et al. (2017), we apply its model to our training set. As shown in Figure 6.2, our method has acceptable performance, clarifying that these results are from 10-fold cross-validation. To apply the prediction with the Barman et al. (2017) approach, we put all the sRNAs sequences of the reverse-strand without the reverse complement. The results from Barman et al. (2017) show lower performance with considerable difference in almost all the measures.

6.1.2 Bacterial testing set

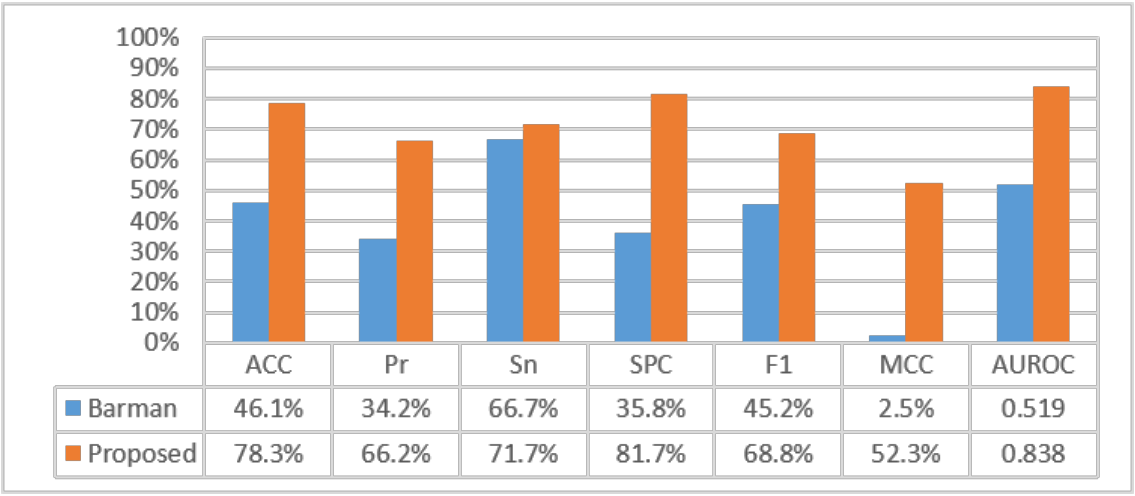
One of the differential factors in this study is not to restrict the tests to only one bacterial family. We also applied our prediction model to gram-positive as well as gram-negative bacteria. As explained in Section 5.1.2.2, the negative sets are twice as big than the size of the positive set.

As shown in the results of the four testing sets, our proposed method is overpassing Barman et al. (2017) approach. All measured AUROC reach values over 0.838, and the significant difference between the two approaches is the ability of the proposed method to detect the non-sRNA sequences, where the average FPR of



	ACC	Pr	Sn	SPC	F1	MCC	AUROC
Barman	53.39%	52.80%	63.50%	43.20%	57.70%	6.90%	0.551
Proposed	78.82%	85.50%	69.40%	88.20%	76.60%	58.70%	0.799

Figure 6.2. Performance measures with Proposed Method training set.



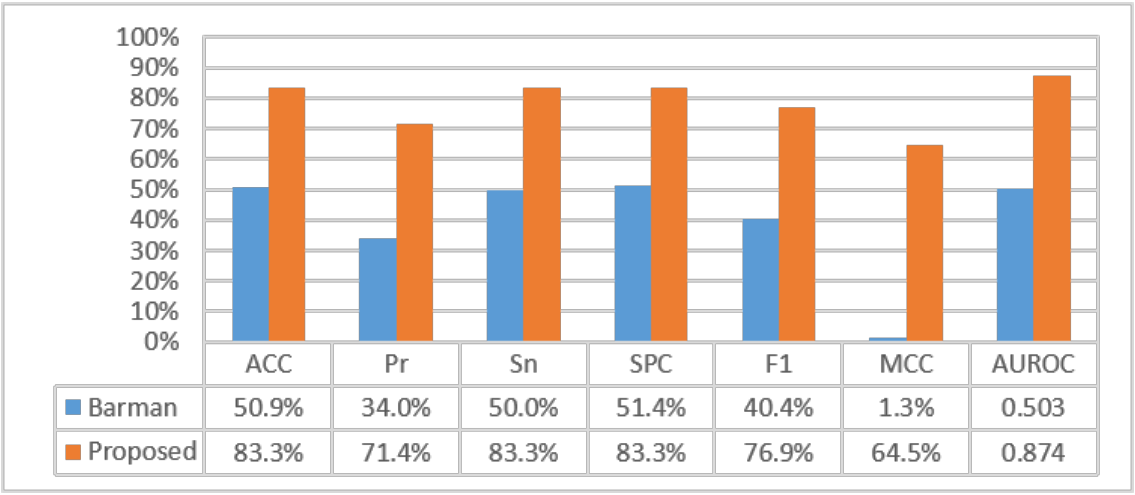
	ACC	Pr	Sn	SPC	F1	MCC	AUROC
Barman	46.1%	34.2%	66.7%	35.8%	45.2%	2.5%	0.519
Proposed	78.3%	66.2%	71.7%	81.7%	68.8%	52.3%	0.838

Figure 6.3. *Listeria monocytogenes* Testing set results.

Barman et al. (2017) approach is 50.52%, which is a high value in comparison with 18.83% of the proposed method.

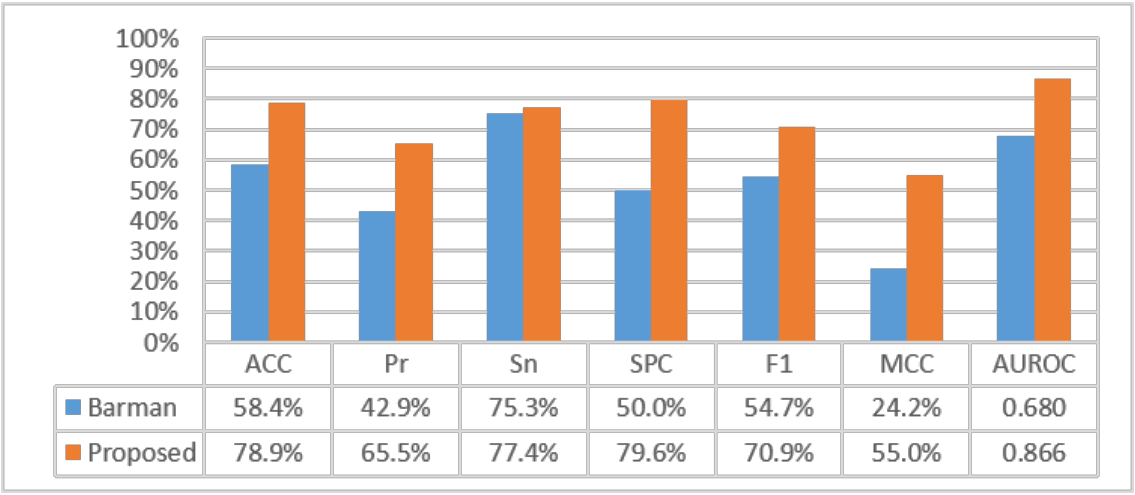
6.1.3 Average single sequence Testing sets results

It was calculated the average of the six testing sets to obtain a better overview of the prediction results in single sequence testing sets. As shown in Table 6.1 and Figure 6.7, the presented method obtained a significant difference in almost all the measures. The capacity of the proposed method to identify the non-sRNA sequences is highly superior to Barman et al. (2017); also, we reach a good value of AUROC with 0.226 of improvement. The smallest difference is in the Sensitivity,



	ACC	Pr	Sn	SPC	F1	MCC	AUROC
Barman	50.9%	34.0%	50.0%	51.4%	40.4%	1.3%	0.503
Proposed	83.3%	71.4%	83.3%	83.3%	76.9%	64.5%	0.874

Figure 6.4. *Streptococcus agalactiae* Testing set results.



	ACC	Pr	Sn	SPC	F1	MCC	AUROC
Barman	58.4%	42.9%	75.3%	50.0%	54.7%	24.2%	0.680
Proposed	78.9%	65.5%	77.4%	79.6%	70.9%	55.0%	0.866

Figure 6.5. *Escherichia coli* K12 Testing set results.

probably because the model of Barman et al. (2017) has the tendency to classify more sequences as sRNA, therefore, leading to a high quantity of TP.

Method	**ACC** (%)	**Pr** (%)	**Sn** (%)	**SPC** (%)	**F1** (%)	**MCC** (%)	**AUROC**
Barman et al.	58.61	48.17	72.30	51.18	56.65	23.61	0.653
Proposed method	80.89	71.28	79.70	82.11	75.18	60.23	0.879

Table 6.1. Average results of single sequence Testing sets.

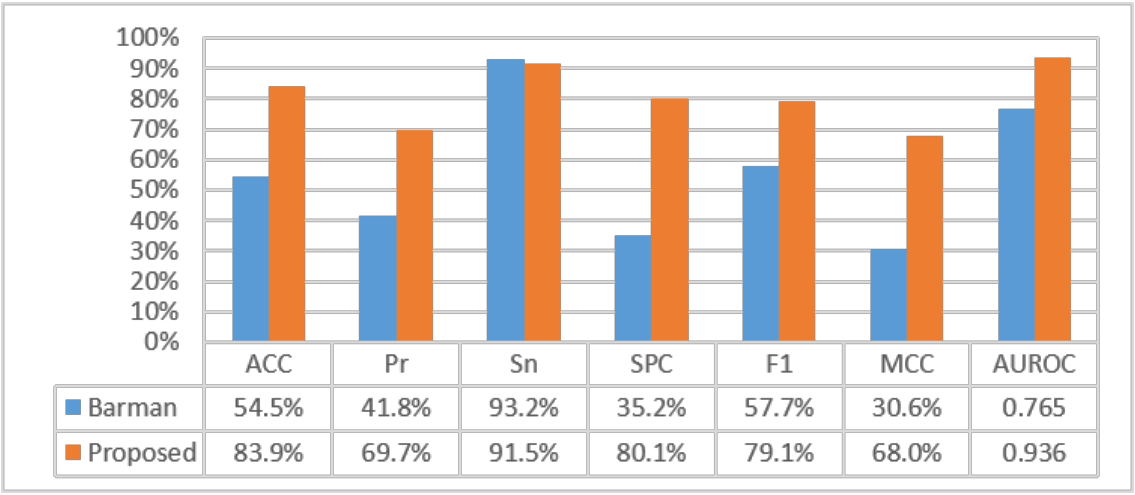


Figure 6.6. SLT2 Testing set results.

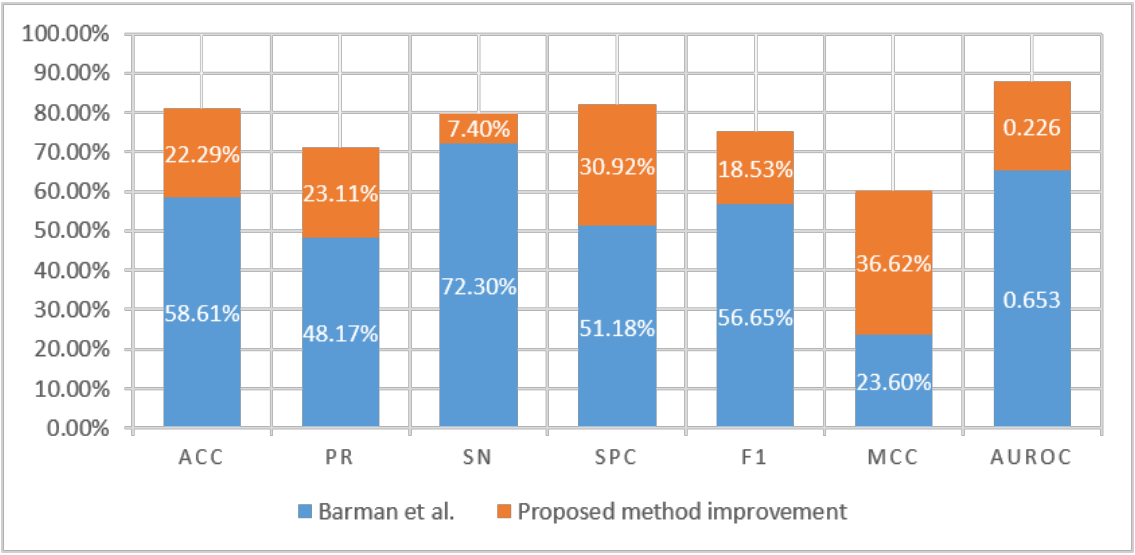


Figure 6.7. Improvement over Barman et al. results.

6.2 Genome-wide Prediction Results

In Genome-wide prediction, we use two bacterial species, i.e., *Salmonella enterica* serovar Typhimurium LT2 and *Escherichia coli* (E. coli) K12. The window size selected for both tests was 145 because it is the median size of the sequences collected from BSRD (Li et al., 2013).

Escherichia coli K12 MG1655 prediction

Both sets use the last version of *Escherichia coli* K12 MG1655, i.e., NC_000913.3. Characteristics of both methods are described in Table 6.2. As shown, only less than 1% are sRNAs windows, even filtering some parts of the whole genome. It thus represents a huge challenge for any prediction tool. The Barman et al. (2017) set includes only sequences from the positive-strand of the genome. The sense of the sequence matter in our method. The Barman et al. (2017) method and the proposed method use a window of 145 nt because this is the median of the size of the experimentally validated sequences of sRNAs from BSRD (Li et al., 2013).

The proposed Framework, in this case, uses a window $l = 145$ with a step $s = 73$, the half of size of the window follows the methodology used by Tran et al. (2009).

The Barman et al. (2017) approach uses a window of 145 nt with a step of 45 nt following the methodology described in their article and manual of their tool.

	Proposed Method			**Barman et al.**		
	sRNA	**non-sRNA**	**Total**	**sRNA**	**non-sRNA**	**Total**
Windows Positive-strand	73	7992	8065	280	58238	58518
Windows Negative-strand	30	7954	7984	-	-	-
Total	103	15946	16049	280	58238	58518
Percentage of set	0.64%	99.36%		0.48%	99.52%	

Table 6.2. *Escherichia coli* K12 testing sets statistics.

In this test, we include the putative sequences published by Tran et al. (2009). Furthermore, we cannot reproduce their methodology without the information of all the filters applied in their work and the absense of the sequence data of the training set.

Method	**Putative sequences**	**ACC (%)**	**FP Rate**	**Pr (%)**	**Sn (%)**	**SPC (%)**	**F1 (%)**	**MCC (%)**	**AUROC**	**sRNA cont.**	**sRNA recov.**
Barman et al.	43223	26.4	73.8	0.5	77.9	26.2	1.0	0.6	0.555	87	81
Tran et al.	16571	-	-	0.4	-	-	-	-	-	-	74
Proposed method	6347	60.63	39.4	1.0	64.1	60.6	2.0	4.0	0.674	63	47

Table 6.3. Performance measures genome-wide prediction *Escherichia coli* K12.

As shown in Table 6.3, our method could deliver a good performance. Our number of putative sequences is seven times smaller than Barman et al. (2017). The lower value is the Sensitivity in comparison with Barman et al. (2017) due to the fact that it is easier to present high TP if it is prioritized over high FP Rate, i.e.,

73.80%. In other measures, we have the best values, and we recovered 74.60% of the sRNAs contained in the set.

Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 prediction

The genome prediction was made using the genome version AE006468.1. The sRNAs to be founded in the genome are not the same of Barman et al. (2017), due to the fact that the information provided in their published data is incomplete, i.e., the absent information of the strand of each sRNA and the inaccurate references of them made it impossible to do a test. The sRNAs used in this experiment are extracted from BSRD (Li et al., 2013), totaling 120 sRNAs. In this case, we used the same step from Barman et al. (2017) of $s = 45$.

As shown in Table 6.4, in comparison with the *Escherichia coli* K12 MG1655, Barman et al. (2017) obtain almost the same quantity of windows, in the opposite, we obtain a fewer number of windows. The difference between our number of windows is because BPROM (Solovyev, 2011) predicted fewer promoters σ^{70} than the first genome.

Even with the differences between the number of windows in each set, as shown in Table 6.5 both models have similar values when compared with *Escherichia coli* K12 MG1655.

	Proposed Method			**Barman et al.**		
	sRNA	**non-sRNA**	**Total**	**sRNA**	**non-sRNA**	**Total**
Windows Positive-strand	75	5042	5117	377	60238	60615
Windows Negative-strand	15	3789	3804	-	-	-
Total	90	8831	8921	377	60238	60615
Percentage of set	1.01%	98.99%		0.62%	99.38%	

Table 6.4. *Salmonella enterica* serovar Typhimurium testing sets statistics.

Method	**Putative sequences**	**ACC (%)**	**FP Rate**	**Pr (%)**	**Sn (%)**	**SPC (%)**	**F1 (%)**	**MCC (%)**	**AUROC**	**sRNA cont.**	**sRNA recov.**
Barman et al.	47802	21.56	78.8	0.7	84.1	21.2	1.3	1.0	0.553	109	105
Proposed method	3166	65.23	34.7	1.7	57.8	65.3	3.2	4.8	0.659	35	27

Table 6.5. Performance measures genome-wide prediction *Salmonella enterica* serovar Typhimurium.

Discussion

The proposed method is the next step in prediction models of sRNAs started by Tran et al. (2009). This approach had good results in the identification of sRNAs, demonstrating good performance in single sequence prediction even in comparison with homology tools, and comparing with four different bacterial strains (i.e., two gram-positive and two gram-negative). Excluding the Multi-tool testing set, the proposed method overpasses with a significant difference in all the testing-sets as well as the Genome-wide prediction in the comparison Barman et al. (2017) turning our proposed method in the best ab initio tool available for sRNA prediction. This work selected 22 new attributes in the search to make the best prediction model focusing on obtaining a good balance in Precision and Specificity, which was demonstrated in the six single sequence testing sets where the negative class is twice as big than the positive class, as well as in both sets of Genome-wide prediction where only less than 1% of the samples are sRNAs. We made a robust validation of our model among the six single sequence testing sets as well as in the experiments with both genomes, where our model demonstrated the same behaviours in their results, specially in the detection of non-sRNA, which is probably the most valuable characteristic of our model considering that a genome has, obviously, a huge amount of non-sRNA sequences.

We demonstrated that the use of primary sequence features alone is not enough to make a prediction model of sRNA. Also, the methodology used to make our negative sets building random sequences from the real ones, and preserving the di-nucleotides frequencies, presented successful results. That was demonstrated with the random sequences used by Barman et al. (2017) in the Multi-tool testing set, and it was also observed in the four bacterial testing sets and with both genome predictions.

The most significant problem in this field of study is the information about sRNAs. In the collection of sRNAs of *Escherichia coli* K12 MG1655, which is one of the most studied organisms, we detected much inaccurate information in the databases. Furthermore, in the construction of our training set of sRNAs from BSRD (Li et al., 2013) we found errors in coordinates, strand, etc. If the data are inaccurate, we cannot obtain good prediction models. The absence of proper quality data of sRNA will be the primary problem to future works in this field.

The proposed framework of Genome-wide prediction does not use any known information about the genome (e.g., promoters, terminators, etc.). That can help and reduce the search-space. The only information used was the coordinates of

CDS extracted from the GenBank file, and the promoters predicted by BPROM (Solovyev, 2011). The results obtained, even being better than the other available tools, are not good enough to use directly to find new families of sRNAs. This framework can be used as an excellent complement to other tools and new methods to search sRNAs. A good strategy is filtering all the unannotated putative windows of the results. In our future studies, we will prioritize the techniques in the window extraction, such as using more than one tool to predict promoters and terminators. Obtaining better windows will allow us to reach better results. The sub-structures of sRNAs sequences must be as close as possible to the real position and size of the real sRNA to result in accurate prediction.

In the current literature, we found many works related to ncRNA prediction, some of them cited and used in this work, others not even mentioned, that are very obscure in the way they present their results. Testing some of them, we could not reach the same results reported in their paper, which weakens the confidence in those methods. In this study, we applied many tests to validate our method and the approach of Barman et al. (2017). Even not reaching high results, our approach overcomes the other procedures. The proposed methodology to validate our prediction model, as well as the testing sets constructed in this work, will be available online, with the objective to be a real benchmark to future prediction tools and avoid confusing result reports in this area.

Chapter 7

Conclusions and Future Work

The first conclusion of this work is that the sRNA prediction with ab initio approach is still an open problem in bioinformatics. It is still needed to find new strong characteristics of sRNAs to construct better prediction models. Also, we do not know if this can be reached due to the limited current knowledge of this type of ncRNA and the high diversity of their characteristics.

The proposed method described in this work demonstrated to be the best ab initio prediction tool of sRNA in the literature, overpassing in performance in the single sequence prediction with an improvement of 22% in ACC, 23% in Precision and specially 30.92% in Specificity, as well as in Genome-wide prediction comparing with the method of Barman et al. (2017), i.e., the last and the only ab initio computational tool available.

Additionally, our prediction tool has a good performance in Precision, Sensitivity, and Specificity in single sequence prediction, and overpasses the tools based on homology approaches. The presented Genome-wide prediction framework has a high quantity of putative sequences; hence, it can be used with different filters like transcription signals or utilized with others prediction tools.

We concluded that it is not enough the use of only primary sequence features to predict sRNA. Another contribution of this work is the 22 new set of features presented to sRNA prediction. Furthermore, the 264 features generated by our tool can be a useful resource for data-mining in the bioinformatics field, especially in the detection of other types of ncRNA. Additionally, we validated in sRNAs the usefulness of the z-score and p-value of mfe to distinguish ncRNA from random sequences.

The central focus of this work was not only making a robust prediction model of sRNA, but also to create a real computational tool for researchers so as to allow

the prediction of new families of sRNAs and obtain advance in knowledge about them and the regulatory networks in prokaryotes.

7.1 Future work

The main future work of this proposal is to develop an online version of our prediction tool because it is currently limited to Linux operative systems and due to the high complexity to install all the software requirements for installing the tool and to provide a more user-friendly application.

Although the provided methodologies and results obtained are quite good in comparison with the other tools, there are some improvements that can still be made regarding the Genome-Wide prediction framework using other transcription signals such as terminators as a filter that can reduce the search space used in conjunction with promoters. Furthermore, some improvements can be made in our prediction model using the types of sRNAs that use chaperone proteins to do the regulation process.

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403--10.

Altuvia, S. (2007). Identification of bacterial small non-coding RNAs: experimental approaches. *Current Opinion in Microbiology*, 10(3):257--261.

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

Arnedo, J., Romero-Zaliz, R., Zwir, I., and Del Val, C. (2014). A multiobjective method for robust identification of bacterial small non-coding RNAs. *Bioinformatics*, 30(20):2875--2882.

Ayodele, T. O. and Zhang, Y. (2010). Machine Learning Overview. In *New Advances in Machine Learning*, chapter 2. InTech.

Azhikina, T. L., Ignatov, D. V., Salina, E. G., Fursov, M. V., and Kaprelyants, A. S. (2016). Role of Small Noncoding RNAs in Bacterial Metabolism. *Biochemistry (Moscow)*, 80(13):1633--1646.

Barman, R. K., Mukhopadhyay, A., and Das, S. (2017). An improved method for identification of small non-coding RNAs in bacteria using support vector machine. *Scientific Reports*, 7(September 2016):1--8.

Beisel, C. L. and Storz, G. (2011). Networks. *FEMS Microbiology Review*, 34(5):866--882.

Clote, P. and Bayegan, A. (2015). Network Properties of the Ensemble of RNA Structures. *PloS one*, 10(10):e0139476.

Clote, P., Ferré, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA (New York, N.Y.)*, 11(5):578--591.

Coventry, A., Kleitman, D. J., and Berger, B. (2004). MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(33):12102--12107.

Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695:1--9.

Dawson, W. K., Fujiwara, K., and Kawai, G. (2007). Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. *PLoS ONE*, 2(9).

Feng, L., Rutherford, S. T., Papenfort, K., Bagert, J. D., van Kessel, J. C., Tirrell, D. A., Wingreen, N. S., and Bassler, B. L. (2015). A qrr noncoding RNA deploys four different regulatory mechanisms to optimize quorum-sensing dynamics. *Cell*, 160(1):228--240.

Flach, P. (2012). *Data, Machine Learning: The Art and Science of Algorithms that Make Sense of*. Cambridge University Press.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Del Moral-Chavez, V., Rinaldi, F., and Collado-Vides, J. (2016). RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1):D133–D143.

Gottesman, S. and Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor perspectives in biology*, 3(12):a003798.

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. (2010). RNAZ 2.0: Improved noncoding RNA detection. *Pacific Symposium on Biocomputing 2010, PSB 2010*, 79:69--79.

Guillier, M. and Gottesman, S. (2006). Remodelling of the Escherichia coli outer membrane by two small regulatory RNAs. *Molecular Microbiology*, 59(1):231--247.

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. (2009). sRNAMap: Genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Research*, 37(SUPPL. 1):D150--D154.

Hüttenhofer, A. and Vogel, J. (2006). Experimental approaches to identify non-coding RNAs. *Nucleic Acids Research*, 34(2):635--646.

Jack0m Getty Images (2017). Bacteria Cell Anatomy and Internal Structure.

Javayel, S., Papenfort, K., and Vogel, J. (2011). *The small RNAs of Salmonella*. Horizon Scientific Press, Norwich, UK.

J.Han, J.Pei, M. (2012). *Data Mining: Concepts and Techniques*, volume 3. Elsevier.

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9(1):192.

Junge, K., Imhoff, F., Staley, T., and Deming, J. W. (2002). Phylogenetic diversity of numerically important Arctic sea-ice bacteria cultured at subzero temperature. *Microbial Ecology*, 43(3):315--328.

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A. M., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Schröder, I., Shearer, A. G., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R. P., Paulsen, I., and Karp, P. D. (2013). EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(D1):605--612.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273--324.

Langley, P. and Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):54--64.

Lee, M. T. and Kim, J. (2008). Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Computational Biology*, 4(8).

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S., and Bassler, B. L. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell*, 118(1):69--82.

Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B., and Ruengjitchatchawalya, M. (2014). Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. *Nucleic Acids Research*, 42(11).

Li, L., Huang, D., Cheung, M. K., Nong, W., Huang, Q., and Kwan, H. S. (2013). BSRD: a repository for bacterial small regulatory RNA. *Nucleic acids research*, 41(D1):D233--D238.

Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658--1659.

Li, W., Ying, X., Lu, Q., and Chen, L. (2012). Predicting sRNAs and their targets in bacteria. *Genomics, proteomics & bioinformatics*, 10(5):276--284.

Livny, J. (2005). sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Research*, 33(13):4096--4105.

Livny, J., Teonadi, H., Livny, M., and Waldor, M. K. (2008). High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PLoS ONE*, 3(9).

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*.

Massé, E., Salvail, H., Desnoyers, G., and Arguin, M. (2007). Small RNAs controlling iron metabolism. *Current Opinion in Microbiology*, 10(2):140--145.

Massé, E., Vanderpool, C. K., and Gottesman, S. (2005). Effect of RyhB small RNA on global iron use in Escherichia coli. *Journal of Bacteriology*, 187(20):6962--6971.

Mathews, D. H. and Turner, D. H. (2002). Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(2):191--203.

Michaux, C., Verneuil, N., Hartke, A., and Giard, J.-C. (2014). Physiological roles of small RNA molecules. *Microbiology*, 160(Pt_6):1007--1019.

Mitchell, T. M. (2006). *The Discipline of Machine Learning*, volume 3. Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Mitchell, T. M. et al. (1997). *Machine learning*. McGraw-Hill Boston, MA.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015). Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Research*, 43(D1):D130–D137.

Oommen, T., Misra, D., Twarakavi, N. K. C., Prakash, A., Sahoo, B., and Bandopadhyay, S. (2008). An objective analysis of support vector machine based classification for remote sensing. *Mathematical Geosciences*, 40(4):409--424.

Panwar, B., Arora, A., and Raghava, G. P. (2014). Prediction and classification of ncRNAs using structural information. *BMC Genomics*, 15(1):127.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology*, 2(4):251--262.

Pena-Castillo, L., Gruell, M., Mulligan, M. E., and Lang, A. S. (2016). Detection of Bacterial Small Transcripts From Rna-Seq Data: a Comparative Assessment. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 21, pages 456--467.

Pischimarov, J., Kuenne, C., Billion, A., Hemberger, J., Cemič, F., Chakraborty, T., and Hain, T. (2012). SRNAdb: A small non-coding RNA database for gram-positive bacteria. *BMC Genomics*, 13(1).

Pitman, S. and Cho, K. H. (2015). The mechanisms of virulence regulation by small noncoding RNAs in low GC gram-positive pathogens. *International Journal of Molecular Sciences*, 16(12):29797--29814.

Quinlan, J. (1993). C4. 5: programs for machine learning. *Machine Learning*, 240:302.

Refaeilzadeh, P., Tang, L., and Liu, H. (2011). Cross-validation. *Advances in Oto-Rhino-Laryngology*, 71:1--9.

Rice, P., Longden, L., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276--277.

Rivas, E. and Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8.

Rose, D., Hertel, J., Reiche, K., Stadler, P. F., and Hackermüller, J. (2008). NcDNAlign: Plausible multiple alignments of non-protein-coding genomic sequences. *Genomics*, 92(1):65--74.

Sato, K., Kato, Y., Hamada, M., Akutsu, T., and Asai, K. (2011). IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):85--93.

Sears, C. L. (2005). A dynamic partnership: Celebrating our gut flora. *Anaerobe*, 11(5):247--251.

Solovyev, V. (2011). Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies*, number January, chapter 4, pages 61--78. Nova Science Publishers.

Sponk Wikimedia Commons (2010). Difference DNA RNA-EN.

Sridhar, J. and Gunasekaran, P. (2013). Computational small RNA prediction in bacteria. *Bioinformatics and biology insights*, 7:83.

Stetter, K. O. (2006). History of discovery of the first hyperthermophiles. *Extremophiles*, 10(5):357--362.

Storz, G., Vogel, J., and Wassarman, K. M. (2011). Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Molecular Cell*, 43(6):880--891.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*, volume 67. Pearson Education.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673--4680.

Tin Kam Ho (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278--282.

Tomizawa, J., Itoh, T., Selzer, G., and Som, T. (1981). Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA. *Proceedings of the National Academy of Sciences*, 78(3):1421--1425.

Tran, T. T., Zhou, F., Marshburn, S., Stead, M., Kushner, S. R., and Xu, Y. (2009). De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics*, 25(22):2897--2905.

Valentin-Hansen, P., Johansen, J., and Rasmussen, A. A. (2007). Small RNAs controlling outer membrane porins. *Current Opinion in Microbiology*, 10(2):152--155.

Washietl, S. and Hofacker, I. L. (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of Molecular Biology*, 342(1):19--30.

Waters, L. S. and Storz, G. (2009). Regulatory RNAs in Bacteria. *Cell*, 136(4):615--628.

Xue, C., Li, F., He, T., Liu, G.-P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, 6:310.

Zhou, J. and Rudd, K. E. (2013). EcoGene 3.0. *Nucleic Acids Research*, 41(D1):613--624.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406--3415.