

Activity 1- Part 2

by

© Ruslan Masinjila

COMP 690A RESEARCH METHODS IN COMPUTER SCIENCE

Professor: Dr. Oscar Meruvia-Pastor

Memorial University of Newfoundland

St. John's

Newfoundland

0.1 Introduction

sRNAs are bacterial small regulatory RNAs, usually less than 200 nucleotides in length[?]. sRNAs are also called non-coding RNAs as they are not translated into a protein. sRNAs play an essential role in gene expression regulation in bacteria and have become a rising class of regulatory RNAs [?]. They are involved in several biological functions such as virulence, metabolism, and environmental stress response [?]. The sRNAs exert their functions when they interact with mRNAs (messenger RNAs) or proteins. These mRNAs or proteins are called the targets of the sRNAs. There have been many sRNAs discovered in recent years; however, their corresponding targets are yet to be found. To understand the roles and functions of sRNAs, it is important to find out their targets and thus, identifying targets of sRNAs has become an essential piece of bacterial RNA science.

There are several programs developed in previous studies for finding sRNA targets [?]. We will discuss more on these in Chapter ???. These programs generate many false positives which reduce the accuracy of the program. The goal of this thesis is to develop a machine learning approach for sRNA Target Prediction to reduce the number of false positives. We compared the performance of this program (sRNARFTarget) with two state-of-the-art programs, CopraRNA [?] and IntaRNA [?]. Our results show that sRNARFTarget substantially outperforms IntaRNA in terms of precision, recall, and running time. However, using comparative genomics as it is done in CopraRNA still achieves the most accurate sRNA target predictions.

Additionally, we have implemented two python scripts to run interpretability models on top of sRNARFTarget to facilitate understanding sRNARFTarget predictions.

We will discuss related work about sRNA target prediction and interpretability of machine learning models in Chapter 0.1. Chapter 0.1 describes the methodology: data collection and processing, feature extraction, machine learning models' training, model selection, and benchmarking. Lastly, we will discuss the programs created for sRNARFTargets' interpretation. Chapter 0.1 presents results and discussion. Chapter ?? is the conclusion. The program code and supplementary files are available at [?, ?]. More citation resources are available at Monash University[?].