Title

by

© Student Firstname & Lastname

A Project Report submitted to the

School of Graduate Studies

in partial fulfillment of the

requirements for the degree of

Master of Science

Supervisor: Dr. Supervisors Name

Department of Computer Science

Memorial University of Newfoundland

July 2021

St. John's Newfoundland

Abstract

Bacterial small regulatory RNAs (sRNAs) play a vital role in the regulation of gene expression in bacteria. sRNAs regulate gene expression by interacting with mRNAs or proteins. Bacterial sRNAs are involved in various processes, such as environmental stress response, metabolism, and virulence. We need to identify the mRNAs and/or proteins that these sRNAs interact with, to understand the functional roles of sRNAs. These mRNAs or proteins are called targets of the sRNAs. There are several computational tools available for sRNA target prediction; however, these tools have a high number of false positives, and the most accurate tool requires sRNA sequence conservation across bacteria. As a result of this research project, a machine-learning-based method (sRNARFTarget) for sRNA target prediction applicable to any bacterium or sRNA has been developed. This method substantially outperforms current non-comparative-genomic methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

sRNAs are bacterial small regulatory RNAs, usually less than 200 nucleotides in length[1]. 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 [2]. They are involved in several biological functions such as virulence, metabolism, and environmental stress response [2]. 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 tar-

gets [3]. 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 [4] and IntaRNA [5]. 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 1. Chapter 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 1 presents results and discussion. Chapter **??** is the conclusion. The program code and supplementary files are available at [6, 7]. More citation resources are available at Monash University[8].

Bibliography

[1] Wikipedia. Small RNA. Wikimedia Foundation. https://en.wikipedia.org/wiki/Small_RNA. (Accessed on 04/10/2021).

[2] E. Gerhart. Chapter Three - Small RNAs in Bacteria and Archaea: Who They Are, What They Do, and How They Do It. volume 90 of *Advances in Genetics*, pages 133 – 208. Academic Press, 2015. http://www.sciencedirect.com/science/article/pii/S0065266015000036.

[3] Adrien Pain, Alban Ott, Hamza Amine, Tatiana Rochat, Philippe Bouloc, and Daniel Gautheret. An assessment of bacterial small RNA target prediction programs. *RNA Biology*, 12(5):509–513, 2015. https://doi.org/10.1080/15476286.2015.1020269.

[4] Patrick R. Wright, Andreas S. Richter, Kai Papenfort, Martin Mann, Jörg Vogel, Wolfgang R. Hess, Rolf Backofen, and Jens Georg. Comparative genomics boosts target prediction for bacterial small RNA. *Proceedings of the National*

Academy of Sciences, 110(37):E3487–E3496, 2013. https://doi.org/10.1073/pnas.1303248110.

[5] Andreas S. Richter, Anke Busch, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 10 2008. https://doi.org/10.1093/bioinformatics/btn544.

[6] Winston Haynes. *Wilcoxon Rank Sum Test*, pages 2354–2355. Springer New York, New York, NY, 2013. https://doi.org/10.1007/978-1-4419-9863-7_1185.

[7] R Documentation. wilcox.test: Wilcoxon rank sum and signed rank tests. Data Camp. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test. (Accessed on 07/10/2020).

[8] Monash University. Citing and referencing: Websites and social media. https://guides.lib.monash.edu/citing-referencing/apa-websites-social-media. (Accessed on 07/10/2020).