Computational prediction, experiment design and statistical validations of non-coding regulatory RNA

Hyejin Cho
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons, and the Computational Biology Commons

Recommended Citation
Computational prediction, experiment design and statistical validations of non-coding regulatory RNA

by

Hyejin Cho

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:

Hui-Hsien Chou, Co-Major Professor
Peng Liu, Co-Major Professor
Gwyn Beattie
Jin Tien
Kan Wang

Iowa State University
Ames, Iowa
2015

Copyright © Hyejin Cho, 2015. All rights reserved.
DEDICATION

This dissertation is dedicated to my husband, family, and friends. A special feeling of gratitude to my wonderful husband, Jaehyun Kim, who has been a constant source of support and encouragement all the time. I give my deepest expression of love and appreciation to my loving mother and two sisters for the encouragement that they gave and the sacrifices they made during my training at the graduate program.

I also dedicate this dissertation to many friends who have supported me throughout the study. I will always appreciate all they have done, especially Deukwoo Kwon for helping me develop my statistical analysis skills, and my best friend Sooyoung Jung for being there for me throughout the entire doctorate program.

Without their unconditional support, this I would not have reached the end of the journey.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................................................... v

LIST OF TABLES ................................................................................................................................................ vii

ACKNOWLEDGMENTS ........................................................................................................................................ viii

ABSTRACT ............................................................................................................................................................ ix

CHAPTER 1. GENERAL INTRODUCTION ...........................................................................................................1

  Introduction ...................................................................................................................................................... 1

  Thesis Organization ......................................................................................................................................... 5

  Literature Review ........................................................................................................................................... 6

  References ...................................................................................................................................................... 9

CHAPTER 2. THERMODYNAMICALLY OPTIMAL WHOLE-GENOME TILING MICROARRAY DESIGN AND VALIDATION

  Abstract ......................................................................................................................................................... 14

  Introduction .................................................................................................................................................... 15

  Microarray Design ......................................................................................................................................... 18

    Sample procurement and genome confirmation ......................................................................................... 18

    Tiling microarray design .............................................................................................................................. 20

    Microarray manufacturing ......................................................................................................................... 21

  Microarray Validations .................................................................................................................................. 23

    Experiment protocol ................................................................................................................................... 23

    Probe level consistency validations ........................................................................................................... 27

    Gene level consistency validations ........................................................................................................... 30

    Novel transcript discoveries ...................................................................................................................... 36

  Conclusions .................................................................................................................................................... 36

  Acknowledgments ......................................................................................................................................... 38

  References .................................................................................................................................................... 38

CHAPTER 3. LIMITATION OF THERMODYNAMIC ANALYSIS THAT PREDICT NON-CODING RNAs IN PROKARYOTES .................................................................................................................. 44

  Abstract ......................................................................................................................................................... 44

  Introduction .................................................................................................................................................... 44

  Assembly ....................................................................................................................................................... 46

  Prediction and Validation ............................................................................................................................. 49

  Results .......................................................................................................................................................... 52

  Conclusion .................................................................................................................................................... 58

  Material and Methods ................................................................................................................................... 59

    Total DNA extraction ................................................................................................................................. 59

    Prediction using PICKY ............................................................................................................................... 59

    Microarray ................................................................................................................................................... 60
Overexpression (Recombination) ......................................................... 60
Quantitative PCR .............................................................................. 61
References ....................................................................................... 61

CHAPTER 4. PREDICTION OF HFQ-BINDING REGULATORY RNAS IN PROKARYOTES
BASED ON THERMODYNAMIC AND STRUCTURAL ANALYSES ......................... 63
Abstract ........................................................................................... 63
Introduction ....................................................................................... 64
Results .............................................................................................. 66
Hfq-binding ncRNA prediction ........................................................... 66
Microarray validations ....................................................................... 72
Comparison to the other ncRNA identification tools ......................... 76
ncRNA target gene predictions .......................................................... 77
Discussion ......................................................................................... 81
Materials and Methods ....................................................................... 83
Growth conditions for bacteria and microarray experiments ............... 83
Thermodynamic analysis using PICKY ............................................. 83
RNA Secondary structure prediction ................................................... 83
Perl programs used in ncRNA prediction .......................................... 84
References ....................................................................................... 84

CHAPTER 5. GENERAL CONCLUSIONS ................................................................. 88
General Discussion ............................................................................. 88
Recommendations for Future Research .............................................. 90
References ....................................................................................... 92

APPENDIX A. REPOSITORY OF SUPPLEMENTARY FILES ............................... 94
LIST OF FIGURES

Figure 2.1. Processing of biological replicates. In each day, two culture tubes under each treatment condition were combined and the RNA sample was extracted from the combined tubes. The extracted RNA samples from 2 different days (days 1 and 3, or days 2 and 4) were later combined to become the final replicate. ..................................24

Figure 2.2. Scatterplots with correlation values for 8 treatment replicate pairs of MG1655. On the graphs, ‘r’ denotes Spearman’s correlation and ‘ccc’ denotes concordance correlation coefficient. .................................................................29

Figure 2.3. Scatterplots with correlation values for the 10 treatment replicate pairs of C58. On the graphs, ‘r’ denotes Spearman’s correlation and ‘ccc’ denotes concordance correlation coefficient. .................................................................30

Figure 3.1. Overview of prediction procedure based on thermodynamic approach......... 50

Figure 3.2. Pearson’s correlation between predicted ncRNAs and their targets in E. coli MG1655. Correlations above FDR-cutoff value are shown as green and correlation below FDR-cutoff value are shown as red orange. ..................................................54

Figure 3.3. Pearson’s correlation between predicted ncRNAs and their targets in A. Tumefaciens C58.................................................................55

Figure 3.4. qPCR expression levels of 13 selected over-expressed ncRNAs and their target genes. ‘insert’ denotes cloned constructs and ‘pGem’ denotes empty vectors...............................................................................................................57

Figure 4.1. The common structures of Hfq-binding ncRNA exemplified by rybB. The Seed Region recognizes and binds target mRNA. The Hfq Site and Terminator are required for ncRNA functions...........................................................................................65

Figure 4.2. Example secondary structure of a predicted Hfq-binding ncRNA with poly-U tail. The RNA sequence and its associated ConxtexFold output are given as well. In ContextFold output, ‘(‘ and ‘)’ denote stem and ‘.’ denotes unbound RNA nucleotide. The ContextFold predicted secondary structure is visualized using PseudoViewer [13]. .................................................................................................69

Figure 4.3. Correlations between predicted Hfq-binding 7-bp or 8-bp ncRNAs and their predicted target regions, target genes, and target operons in E. coli. ........................................73

Figure 4.4. Correlations between predicted Hfq-binding 7-bp or 8-bp ncRNAs and their predicted targets in Agrobacterium tumefaciens.................................................................75

Figure 4.5. Network analyses of predicted ncRNA and their target genes. Each cluster represents an ncRNA candidate (larger center node) and its associated predicted
target genes (smaller surrounding nodes). The few thicker lines that connect clusters are overlapped KEGG pathway links. The many thinner lines connecting targets within a cluster are overlapped gene ontology network links, suggesting targets predicted for some ncRNAs have similar biological functions.
LIST OF TABLES

Table 2.1. Sequence-level comparisons cannot faithfully predict thermodynamic properties. ............................................................................................................................................................................................... 17

Table 2.2. Polymorphisms between lab bacteria genomes and official GenBank reference genomes. ............................................................................................................................................................................................... 20

Table 2.3. Results of tiling microarray design .............................................................................................................................................................................................................. 22

Table 2.4. Growth conditions of MG1655 and C58.................................................................................................................................................................................................. 25

Table 2.5. Detected stress response genes in MG1655 and C58.................................................................................................................................................................................................. 31

Table 2.6. Differentially expressed stress genes in each pair of conditions for MG1655.................................................................................................................................................................. 33

Table 2.7. Differentially expressed stress genes in each pair of conditions for C58. ................................................................................................................................................................................................ 34

Table 3.1. Comparison results with k-mer 55 .................................................................................................................................................................................................................. 47

Table 3.2. Results of NGS Assembly ................................................................................................................................................................................................................................. 48

Table 3.3. Predicted ncRNAs of E. Coli and A. Tumefaciens ................................................................................................................................................................................................. 52

Table 4.1. Results of Hfq-binding ncRNA prediction ............................................................................................................................................................................................................. 71

Table 4.2. Results of predicted Hfq-binding ncRNAs after FDR control at 5% .................................................................................................................................................................................................. 74

Table 4.3. Confirming ncRNA predictions by 3 other ncRNA-identification tools ........................................................................................................................................................................ 77
ACKNOWLEDGMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from lab members.

I would like to express my deepest gratitude to my advisor, Dr. Hui-Hsien Chou, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. He also patiently corrected my writing and financially supported my research. I would like to thank Dr. Peng Liu for helping and guiding me to develop my background in statistics. I also would like to thank Dr. Kan Wang and Dr. Gwyn Beattie for their valuable advices on bacterial studies. I also thank Drs. Kan Wang and Keunsub Lee for providing the *Agrobacterium tumefaciens* C58 strain and teaching us how to maintain this bacteria. Special thanks goes to Dr. Jin Tien for his advice on machine learning and participating in my committee.

I would like to thank Jennifer Chang, who as a good friend is always willing to help and give her best suggestions. Many thanks to Taryn Anderson, Matthew Hunt, and Bhagyashree Birla for their help in bacteria maintenance, RNA extraction, and other molecular biology experiments.

I would also like to thank many friends in US and in Korea. They have cheered me up, showed me outside of a graduate life and given me energy to continue this study.

Finally, I would like to thank my family — my husband, mother, two sisters, niece, and family-in-law. They are always supporting me and encouraging me with their best wishes.
Non-coding regulatory RNAs (ncRNAs) regulate a host of gene functions in prokaryotes, e.g., transcriptional regulation, RNA processing and modification, translation regulation and mRNA stability. Some ncRNAs have been identified experimentally but many are yet to be found. ncRNAs can be classified as either cis- or trans-acting. cis-ncRNAs perfectly complement their target genes and are usually encoded on the anti-sense strands of the targets. On the contrary, trans-ncRNAs regulate their target genes through short and often imperfect base-pairings with the targets, and are usually encoded elsewhere on the genome. A whole-genome thermodynamic analysis can be performed to identify all imperfect but stable base-pairings between all annotated genes and some genomic regions encoding ncRNAs from the same species. However, the sizes of these base-pairing regions are short and variable, and their melting temperatures vary greatly between perfectly and imperfectly matched targets. It is difficult to predict trans-acting ncRNAs solely based on the thermodynamic analysis. Therefore, we also have to consider known ncRNA structures to improve our predictions. We matched our prediction results to known ncRNAs, measured the correlation between ncRNA candidates and their predicted targets using whole-genome tiling microarrays, and we also performed gene network studies to find supporting evidences to our predictions.

In Chapter 2, we describe how we have designed, created, and validated whole genome tiling microarrays for *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 that are used for ncRNA validations. In Chapter 3, we explain ncRNA predictions employing only the
thermodynamic approach and why they did not produce conclusive results. In Chapter 4, we illustrate how we have redeveloped our prediction method according to lessons learned from the previous thermodynamic only analysis, and we present a systematic, novel 5-step approach to predicting Hfq-binding ncRNAs in *E. coli* and *Agrobacterium tumefacies* based on thermodynamic analyses as well as known structural properties of this class of ncRNAs. Our new prediction results match to known ncRNAs, have high correlation between ncRNA candidates and their predicted targets as measured by the whole-genome tiling microarrays, and are confirmed by 3 other ncRNA identification software tools. We also compared the predicted ncRNA targets to known ncRNA targets and performed a gene ontology network analysis to infer the ncRNA functions. The 5-step prediction method is generally applicable to other prokaryote species and may help advance ncRNA research in prokaryotes.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Prokaryote species express small non-coding RNAs (ncRNAs) to regulate their gene expressions [1]. ncRNAs can be classified as either cis- or trans-acting. cis-acting ncRNAs act in a perfectly matched manner to their target genes and are usually encoded on the anti-sense strands of the targets. On the contrary, trans-acting ncRNAs regulate their target genes usually through imperfect base-pairing with the targets, and they are usually encoded elsewhere on the genome. ncRNAs play many important roles in gene expression, including transcriptional regulation, RNA processing and modification, translation regulation and mRNA stability [2]. Many efforts have been spent to find ncRNAs in prokaryotes due to their various important functional roles and identification of ncRNA candidates in different species has been increased dramatically owing to high-throughput methods and novel computational algorithms [3]. However, the variable sizes of ncRNAs (up to 500 base pair), their multitude potential encoding regions in the 5’ and 3’ untranslational regions of genes or the intergenic region, and their imperfect and short base-parings with target genes make prediction of trans-acting ncRNAs challenge. To face the challenge, we have developed a systematic method to predict trans-acting ncRNAs based on thermodynamic analysis and some unique features of ncRNAs to sensibly predict trans-acting ncRNAs. *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 have been chosen to test our method, because *E. coli* MG1655 has been well studied and many known ncRNAs exist in *E. coli* and *Agrobacterium tumefaciens* C58 is an important model species but not
many known ncRNAs exist in C58. Therefore choosing these two species allows us to both confirm our prediction methods with *E. coli* and to apply it on the discovery of novel ncRNAs in C58.

At the beginning of the study, the genomes of the two strains MG1655 and C58 have been *de novo* sequenced and assembled using next-generation sequencing (NGS) techniques. This helps improve our microarray design and ncRNA prediction quality. Single nucleotide polymorphisms (SNPs) between our lab strains and the GenBank reference sequences were discovered from the newly assembled genomes. The SNP locations in MG1655 were compared with our predicted ncRNA locations, and we found the majority of *E. coli* SNPs were located in known ncRNAs in GenBank report and predicted ncRNAs. This supports the notion that our efforts in sequencing and assembling the genomes do influence the quality of the ncRNA predictions and the microarray design. We describe the newly assembled sequences of *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 in Appendix A. We decided to use whole-genome tiling microarrays to validate our ncRNA predictions for several reasons: less RNA material is required for microarray than NGS [4]; the lack of polyA tail in prokaryotes mRNA makes it difficult to build full-length cDNA sequencing libraries; NGS did not perform well for GC-rich CDSs with light normalization; and rRNA removal in strong normalization can cause non-coding regulatory RNA losses [5]. Microarray is able to detect predicted ncRNA sites and their association with predicted targets; if the predicted ncRNA encoding regions are to express RNA transcripts, the expressed transcripts can bind strongly to the target genes, resulting in high correlation values that can be inferred from microarray data. If found, these high correlation regions
are potential encoding sites of ncRNAs. Even if high correlation values are not identified, some predicted ncRNA sites may still be function if they indeed express RNA transcripts, and that can be readily detected by the tiling microarrays. For these purposes, two high density whole-genome tiling microarrays have been designed to cover all predicted ncRNAs and all annotated target genes and their anti-sense strands on the *E. coli* MG1655 and the *Agrobacterium tumefaciens* C58 genomes.

In the first attempt of prediction, we predicted *trans*-acting ncRNAs solely based on the whole-genome thermodynamic analysis software PICKY [6]. PICKY identifies all imperfect but strong base-pairings between all annotated genes and some untranslated or intergenic regions from the same species. These potential regions are selected if they are within a certain thermodynamic threshold showing that they can form imperfect but strong base-parings with annotated genes. However, after predicting ncRNAs this way but failing to validate them in experiments, we later learned that *trans*-acting ncRNAs may not always bind strongly to their targets; some ncRNAs form just short 6-12 base-pairs (bp) with their targets [7]. The melting temperatures of predicted ncRNAs to targets can be highly variable, thus our first prediction method based solely on PICKY thermodynamic analysis did not provide conclusive results.

We subsequently revised the prediction method of *trans*-acting ncRNAs. Some known *trans*-acting ncRNAs require the Hfq binding protein to bind their target genes, and these so called Hfq-binding ncRNAs have unique structural properties that our new prediction method takes into consideration. In addition to the stem loop, which is common to many ncRNAs, Hfq-binding ncRNAs share three special regions on their single-strand stretches: the seed region, Hfq-binding site, and poly-U tail. The new method starts by
finding poly-U regions in both directions of the genome. After finding these poly-U regions, a whole-genome thermodynamic analysis was performed to identify all short base-pairings 6~12 bp between all annotated genes and all potential seed regions upstream from the poly-U regions; if any seed region expresses ncRNA transcripts, the transcripts can potentially bind to target genes. Next, secondary RNA structures were predicted for all identified candidate ncRNAs and the predicted structure information was used as a filter with conditions that three required structural regions should be located in single stranded stretches on the predicted structure. The unique structural properties of Hfq-binding ncRNA helped to narrow down potential ncRNA sites detected by PICKY and improve our prediction method.

Several validation steps were then performed for the predicted Hfq-binding ncRNAs in E. coli and Agrobacterium tumefaciens. First, predicted ncRNAs in E. coli were compared to known ncRNAs in GenBank Report #U00096.3 and predicted ncRNAs in Agrobacterium tumefaciens were compared to confirmed ncRNAs in Lee, et al [8]. Next, our custom-designed whole-genome tiling microarrays were used to measure the correlation between all our predicted ncRNAs and their targets. High correlations can suggest associations between predicted ncRNAs and their targets, and we found many predicted ncRNAs and their predicted targets were highly correlated. To further validate our predictions, we also confirmed our ncRNA candidates with 3 other non-coding RNA identification software and found that they were largely in agreement with our predictions. The best advantage of our prediction method is that the thermodynamic analysis with PICKY allowed us to predict ncRNAs as well as their potential targets. In addition, our novel prediction method is ab initio — Our method starts from a whole genome as input and directly finds potential
ncRNA encoding regions and their targets. Our method does not require any *a priori* study while the other 3 RNA identification software tools require some prerequisite studies to prepare input sequences from potential encoding areas of ncRNAs. As mentioned earlier, our method also predicts targets of ncRNA candidates, too. So we finally performed a gene network analysis with gene ontology data to understand the predicted targets and found many targeted genes of each predicted ncRNA serve similar molecular functions.

The details of each step outlined above will be discussed in the following chapters. We will also describe some other useful purposes of our tiling microarrays, expansion of the prediction method with other structural properties, and some future perspectives of this project such as artificial ncRNA design and applications to other species in the general conclusion chapter.

**Thesis Organization**

This dissertation starts with an introduction chapter. Basic principles of bacterial small noncoding RNAs and their important roles in life cycle are outlined in order to describe why this prediction of ncRNAs in prokaryotes was chosen for the PhD study. The next chapter summarizes how to predict ncRNAs in two attempts. The final chapter describes how to validate predicted ncRNAs. Chapter 2, which has been submitted for publication, explains the design of whole genome tiling microarray, validation of the design and further application of our custom-designed tiling microarray. Chapter 3 describes limitations of thermodynamic approach for predicting *trans*-acting ncRNAs in prokaryotes. Chapter 4, which has also been submitted for publication and modified for this dissertation,
describes a novel prediction method for Hfq-binding ncRNAs in prokaryotes. Chapter 5 present the general conclusions of this PhD study and possible future research based on our novel prediction method for Hfq-binding ncRNAs.

**Literature Review**

1) Properties of noncoding ncRNA

Bacterial small non-coding RNAs (ncRNA)s regulate gene expression positively and negatively [1]. ncRNAs have important roles in regulation such as transcriptional regulation, RNA processing and modification, translation regulation and mRNA stability [2]. ncRNAs are very structured and hold several structural properties such as stem-loops. Some ncRNAs binds to target genes incorporated with Hfq protein [9]–[11]. These ncRNAs, called Hfq-binding ncRNAs, have common structural properties [12]. Poly-U tail is one of the properties and becomes the start regions to search for our prediction method [13].

2) De novo assembly

The genomes of the two strains *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 have been *de novo* sequenced and assembled using next-generation sequencing (NGS) and the Velvet software [14]. Velvet was selected among other de novo assembly tools such as AbySS [15], MIRA [16] and Zorro [17] after performance tests. Beside Velvet, other bioinformatics tools and our own many Perl programming were also used to finish the assembly. Minimus2, which is the part of the AMOS software package [18], is used to merge Velvet output contigs to form longer scaffolds. BLAT is then used to align merged contigs to the reference genomes [19]. The alignment information is important to correctly orient
some contigs, find repeated contigs and fill in the gaps among aligned contigs. We have also 
created our own Perl scripts to close sequencing gaps and finish the assembly.

3) Thermodynamic approach

Thermodynamic approach using PICKY is used for prediction and tiling microarray design 
[6], [20], [21]. For prediction, special ‘Examine’ function in PICKY is used to examine DNA 
fragments upstream from poly-U regions to determine if they match to any regions in the 
whole genome. This is calculated according to the Nearest-Neighbor Model of nucleotide 
hybridizations, the most precise model to date [22], [23]. For tiling microarray, PICKY 
allow us to achieve maximum whole-genome coverage under the thermodynamic 
constraints of each target genome, because PICKY compares all query probes to all input 
genes and reports only the highest found above a user-defined threshold.

4) RNA secondary structure prediction

Many software tools, such as IPKnot [24], CentroidFold [25], and ContextFold [26], are 
considered to predict ncRNA secondary structure. ContextFold was chosen to predict 
secondary structures of our ncRNA candidates among many RNA structure prediction 
programs because it is ranked top on secondary structure predictions of short RNAs (20-
200 nt) and medium-sized RNAs (201-800 nt) [27].

5) Tiling microarray design

Gene expression microarrays are gradually being replaced by next-generation sequencing 
(NGS) techniques because NGS techniques can detect novel RNA transcripts and provide a 
better dynamic range of measured gene expression values [28]. Nevertheless, microarrays 
are still commonly used in some other applications. In this project, we focus on another 
application where microarrays are still viable for identifying novel transcripts and genome
variations and our design is optimized and efficient one. The melting temperatures of each probe candidate with its intended target and with its closest nontargets anywhere in the genome are calculated by PICKY according to the thermodynamic nearest-neighbor models of prefect matches [23], mismatches [22], [29]–[31], bulges [32], and dangling-ends [33]. The equations used by PICKY are deterministic according to thermodynamic principles and PICKY exhaustive applies these equations to all potential probe candidate and nontarget hybridizations. Therefore, PICKY designs are thermodynamically optimal and provide high probe specificity [20].

6) Confirmation with other 3 ncRNA identification software

Our predicted Hfq-binding ncRNAs were confirmed with other 3 different programs to identify ncRNAs for validation. The first method was GraPPLE which identifies noncoding RNA using graph properties and Support Vector Machine (SVM) [34], [35]. Secondary structure of ncRNA are converted to graph properties in this method [34]. The second method was RNAcon and this method predicts noncoding RNAs based on sequence composition [36]. This method is most recent one and produces higher sensitivity and accuracy than other two methods according to the author’s claim. The third one was Coding Potential Calculator (CPC) and CPC distinguishes coding RNAs from noncoding RNAs based on sequence features [37].

7) Biological network for association between predicted ncRNAs and their predicted target genes

Using the network analysis and visualization software Mango (http://www.complex.iastate.edu/download/Mango/), we first represented predicted ncRNAs and their target genes. KEGG database and Gene Ontology database (version 2,
http://geneontology.org/gene-associations/gene_association.ecocyc.gz) are imposed to the generated network by MANGO to detect relationships between predicted ncRNAs and their targets.

8) Artificial design of ncRNA

Our new predicted ncRNAs may allow us to design artificial ncRNA to regulate specific target gene. To pursue this effort, machine learning techniques are needed to find common structural properties shared by the confirmed ncRNAs. First, we will use software, called ContextFold to predict secondary structures of confirmed ncRNAs. From predicted structures, we will consider using pattern recognition by machine learning methods to find the common structural features [38]. The kernel-based approach based on dual graph representation [42] or/and comparative RNA structure analysis such as LocARNa [43] are machine learning methods for pattern recognition. In addition, known RNA structural data can be downloaded from RNA related databases such as Rfam [39], fRNAdb [40], or BSRD [41] to build training data for machine learning method. If we successfully find common structural properties among confirmed ncRNAs, we may be able to design artificial small ncRNAs using common structural properties to regulate specific target genes we choose.

References


CHAPTER 2. THERMODYNAMICALLY OPTIMAL WHOLE-GENOME TILING MICROARRAY DESIGN AND VALIDATION

A paper submitted to Microarrays

Hyejin Cho¹,² and Hui-Hsien Chou¹,³,⁴

Abstract

Microarray is an efficient apparatus to interrogate the whole transcriptome of species. Microarray can be designed according to annotated gene sets, but this design method is not applicable to unannotated species and the resulting microarrays cannot be used to identify novel transcripts. Alternatively, a whole-genome tiling microarray can be designed using only genomic sequences without gene annotations, and it can be used to detect novel RNA transcripts as well as known genes. The difficulty with tiling microarray design often lies in the tradeoff between probe-specificity and coverage of the genome. Sequence comparison methods based on BLAST or similar software are commonly employed in microarray design, but they cannot precisely determine the subtle differences in thermodynamic properties between probe targets and partially matched nontargets during hybridizations. Using the whole-genome thermodynamic analysis software PICKY to

¹ Department of Genetics, Development and Cell Biology, Iowa State University
² Primary researcher and author
³ Department of Computer Science, Iowa State University
⁴ Author for correspondence
design tiling microarrays, we can achieve maximum whole-genome coverage under the thermodynamic constraints of each target genome. The resulted tiling microarrays are thermodynamically optimal in the sense that all selected probes share the same target and closest nontarget melting temperature separation range and no additional probes can be added without violating the specificity of the microarray to the target genome. Here we describe the design of two whole-genome tiling microarrays for \textit{Escherichia coli} MG1655 and \textit{Agrobacterium tumefaciens} C58 using this new method and their experimental validation results.

Keywords: prokaryote transcriptome; tiling microarray design; thermodynamics; hybridization; PICKY software; microarray validation.

\textbf{Introduction}

Different types of microarray exist, and they all have DNA probes on the microarray surface to hybridize, or capture, targeted sequences in the samples that are poured over them. Microarrays differ in their probes design and their intended targets. The most common microarrays are designed to detect gene expressions; their probes are designed according to annotated gene sets and are used to detect individual gene expressions. Although gene expression microarrays have been in use for more than a decade and have produced a large volume of biological data, they are gradually being replaced by next-generation sequencing (NGS) techniques because NGS techniques can detect novel RNA transcripts and provide a better dynamic range of measured gene expression values [1].

Nevertheless, microarrays are still commonly used in some applications. Sequence fragment capturing microarrays work by grabbing specific genome fragments or RNA
transcripts of interest to researchers, hence enriching the targeted samples. The captured fragments can then be sequenced and analyzed using NGS techniques [2,3]. In this work, we focus on another application where microarrays are still viable for identifying novel transcripts and genome variations — the whole-genome tiling microarrays. A tiling microarray is designed against a genome, not a gene set, and can be used to detect all transcripts expressed from the genome regardless if they are annotated genes or novel transcripts. The latter may include short regulatory RNAs that are recently the interest of study [4–6].

We have designed two whole-genome tiling microarrays for *Escherichia coli* (*E. coli*) MG1655 and *Agrobacterium tumefaciens* C58 using the whole-genome thermodynamic analysis software PICKY that has been used to design traditional microarrays [7–9]. PICKY analyzes the whole genome to identify thermodynamically unique probes. The melting temperatures of each probe candidate with its intended target and with its closest nontargets anywhere in the genome are calculated by PICKY according to the thermodynamic nearest-neighbor models of perfect matches [10], mismatches [11–14], bulges [15], and dangling-ends [16]. The equations used by PICKY are deterministic according to thermodynamic principles and PICKY exhaustively applies these equations to all potential probe candidate and nontarget hybridizations. Therefore, PICKY designs are close to optimal and have been validated [17].

Sequence-level comparison software such as BLAST [18] are often used to estimate probe specificity by calculating the matched bases (identity) or the longest match stretch to nontargets [19,20], but these estimates are less precise than thermodynamics. For example, a previous tiling microarray probe set containing 409,807 probes for a bacterial
species were screened using PICKY, and 21,773 (5.3%) of the probes were found to have the potential to hybridize to identified nontargets with melting temperatures less than 5°C from or even higher than their target melting temperatures [21] (W83.picky is described in Appendix A). In Table 2.1, example data from this probe set shows that probes of the same length (50 bp), the same identity to nontarget (24 bp; <50%) and the same match stretch (14 bp; <30%) can still have PICKY estimated melting temperatures ranging from 28°C to 68°C. Furthermore, the lowest probe-to-target melting temperature at 54.33°C is much lower than the highest probe-to-nontarget melting temperature at 68.56°C, thus it is impossible to set a microarray hybridization temperature that allows all probes to function effectively. The reason for these discrepancies is because thermodynamic equations are inherently nonlinear. Depending on the hybridization structure, a wide range of melting temperatures can be calculated. This probe set was designed using a sophisticated pipeline involving BLAST to screen for cross-hybridizations and was considered optimal by that standard [22], but evaluations using thermodynamics revealed the weakness of sequence-level specificity estimates.

Table 2.1. Sequence-level comparisons cannot faithfully predict thermodynamic properties.

<table>
<thead>
<tr>
<th>Probe (top strand) and nontarget (lower strand) match (complementary bases in uppercase; mismatched bases in lowercase)</th>
<th>Identity (bp)</th>
<th>Stretch (bp)</th>
<th>Nontarget Melting Temp.(°C)</th>
<th>Target Melting Temp.(°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tagagtAaaAAaaCAataAaAGAcattaAaAGAAAAATGATTTTgtttTtt</td>
<td>24</td>
<td>14</td>
<td>28.39</td>
<td>55.14</td>
</tr>
<tr>
<td>tgttgtAaGTacctGTTacgtTCTccgccCTCTTTTACTAAAAAaagtAt</td>
<td>24</td>
<td>14</td>
<td>29.92</td>
<td>54.33</td>
</tr>
<tr>
<td>CTTgAaaItgaaTcAaatcctctaTaaaTCAATGATATGAtacaataACA</td>
<td>24</td>
<td>14</td>
<td>38.04</td>
<td>61.27</td>
</tr>
</tbody>
</table>

Table 2.1. Sequence-level comparisons cannot faithfully predict thermodynamic properties.
Our goal in this work is to test whether PICKY, which was originally developed to take annotated gene sets as input and design gene-specific probes, can be repurposed to also design tiling microarrays with maximum probe coverage but maintain the same thermodynamic specificity of PICKY designed probes. After the microarrays were designed and manufactured, the two bacteria species were grown under 10 different treatment conditions to trigger gene expression changes. Subsequently, samples extracted from them were applied to the tiling microarrays to validate their design quality and also to uncover novel transcripts.

Microarray Design

Sample procurement and genome confirmation

The E. coli MG1655 strain was obtained from CGSC E. coli Genetics Resources at Yale University (http://cgsc.biology.yale.edu, CGSC# 6300). The Agrobacterium tumefaciens C58 strain was obtained from Dr. Kan Wang’s lab (http://www.agron.iastate.edu/ptf/employee/director.aspx). Bacteria were recovered from the delivery medium and grown under standard conditions (37°C in Luria-Bertani
medium for MG1655 and 28°C in YEP medium for C58). The QIAGEN DNeasy blood & Tissue kit (#69504) was used to extract total DNA from both bacteria. The Qubit 2.0 Fluorometer was used to precisely quantify DNA concentration in the samples and the Experion DNA 12K Analysis Kit was used to check the DNA quality. The total DNA was eluted in 100 uL buffer and 50 uL of that was sent for sequencing confirmation.

The genomes of the two bacteria MG1655 and C58 were resequenced using the Illumina HiSeq 2000 instrument and de novo assembled using the Velvet software [23]. Minimus2, which is part of the AMOS software package, was used to merge Velvet contigs to form longer scaffolds [24]. BLAT was then used to align merged contigs to the reference genomes [25]. The alignment is important to correctly orient some contigs, find repeated contigs and fill in the gaps among aligned contigs. The reference genomes were used to guide the assembly of the contigs, but not the individual reads. The At plasmid of C58 was not successfully assembled due to lack of matched contigs, thus the reference sequence was used in subsequent design.

We have found hundreds of single nucleotide polymorphisms between the assembled genomes and the database reference genomes, which support our initial concern that the bacteria we obtained might not match the database sequences exactly. These polymorphisms, which are summarized in Table 2.2, might cause slightly less precise tiling microarray design if left unidentified. The resequencing confirmation step is not expensive and it helps improve the design quality, but it is entirely optional. The NGS data can be obtained from NCBI Short Read Archive database (http://www.ncbi.nlm.nih.gov, SRX806374 and SRX806654), and the assembled new genomic sequences are described in Appendix A.
Table 2.2. Polymorphisms between lab bacteria genomes and official GenBank reference genomes.

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Genome size</th>
<th>Single nucleotide polymorphisms</th>
<th>Identity to GenBank reference genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli MG1655</td>
<td>4,639,675 bp</td>
<td>265</td>
<td>99.9945 %</td>
</tr>
<tr>
<td>A. tumefaciens C58</td>
<td>5,746,078 bp</td>
<td>203</td>
<td>99.9965%</td>
</tr>
</tbody>
</table>

**Tiling microarray design**

Based on the assembled genome sequences, we designed the two whole-genome tiling microarrays for *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 using PICKY [9,26]. The design process is summarized as follows. The genome sequences were broken up into 100 base-pair (bp) fragments without overlaps — these were treated as targets for probe design to ensure even distribution of the tiling probes. Separately, 50 bp fragments centered on the boundaries between the target fragments (25 bp on either side of a boundary) were extracted and treated as unintended targets for probe design (i.e., fragments to avoid) to ensure that tiling microarray probes will not inadvertently target the boundaries between fragments. PICKY was run using both the target and unintended fragments as input. The benefit of this approach is that we can take full advantage of the probe specificity calculation of PICKY while making it design tiling microarrays with evenly distributed probes. The following parameters specific to tiling microarray design were given to PICKY: maximum match length 18, minimum match length 8, minimum sequence similarity 66%, and minimum melting temperature difference 5°C. All other PICKY
parameters were taken at their default values, including the screening of both strands of each input sequence to ensure probe specificity in either direction. The minimum and maximum match length parameters and the minimum sequence similarity parameter ensure that a wide range of random partial nontarget matches will be screened thermodynamically by PICKY. The minimum melting temperature difference ensures that only probes unique to the target fragments will be selected.

After running PICKY the first time, it turned out that some target fragments did not have matching probes under the stringent design parameters. To increase the number of useful probes, we ran PICKY again with the following new input. The target fragments were separated into two different sets: one contained fragments without probes as a new target set, and the other contained fragments that had probes found during the first PICKY run. The second set was combined with a modified boundary fragment set to form the new unintended fragment set. The modified boundary fragments were shortened to 40 bp centered on the boundaries between the target fragments. The second PICKY run used the same parameters as in the first run, but it produced additional probes because the shortened boundary fragments allow more borderline probes to be selected.

**Microarray manufacturing**

Microarray probes obtained from both PICKY runs were merged to obtain the final design output. When designing tiling microarray for the C58 bacterium, the pTi plasmid of C58 was also added to the design data set to increase the versatility of the C58 tiling microarray. We have chosen the NimbleGen Custom Microarray Service to manufacture the tiling microarrays. The NimbleGen microarray platform has a synthesis cycle limitation of
148 on custom designed microarray probes [27]. Therefore, 19 MG1655 and 20 C58 probes were removed because they exceeded the limit.

Probes for the exogenous gene encoding hygromycin resistance to *E. coli* and *Agrobacterium tumefaciens* were added: 1000 hygromycin probes were added to the C58 probe set and 2000 hygromycin probes were added to the MG1655 probe set. These probes can be used as quality controls if the hygromycin gene is added to each bacteria sample during the microarray hybridization protocol to help detect any technical bias. The manufacturer also added other control probes to the final probe set for proprietary quality control and microarray image alignment. All probes were synthesized in situ on the NimbleGen microarray surface using the 4x72K microarray layout, meaning that there were 4 independent microarrays per each NimbleGen glass chip and each microarray contained up to 72K probes. Final results of the microarray design are summarized in Table 2.3. Complete microarray design information and experiment data were deposited into the NCBI Gene Expression Omnibus (GEO) database with the Series Access Number GSE61738, which is a Super Series combining both microarray series for MG1655 and C58.

**Table 2.3. Results of tiling microarray design**

<table>
<thead>
<tr>
<th></th>
<th>MG1655</th>
<th>C58</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Microarray unique probe count</strong></td>
<td>67,435</td>
<td>71,498</td>
</tr>
<tr>
<td><strong>Avg. probe length</strong></td>
<td>41 bp</td>
<td>40 bp</td>
</tr>
<tr>
<td>% of 100-bp fragments without useful probes</td>
<td>6.13 %</td>
<td>4.48 %</td>
</tr>
<tr>
<td>% of genome covered by fragments with probes</td>
<td>93.87%</td>
<td>95.52%</td>
</tr>
<tr>
<td>Hygromycin control probes</td>
<td>2000</td>
<td>1000</td>
</tr>
</tbody>
</table>
Although NimbleGen has exited the custom microarray manufacturing business, the tiling microarrays can still be manufactured by other manufacturers given the original microarray design information. Naturally, some array-specific protocols such as labeling and image quantification may need to be modified accordingly if different microarray platforms are used.

**Microarray Validations**

**Experiment protocol**

*E. coli* MG1655 and *Agrobacterium tumefaciens* C58 cells were grown under 10 different treatments (1 standard and 9 stressed conditions) listed in Table 2.4. The significantly varied growth conditions can help induce large-scale gene expression changes that ideally should cover most of the transcriptome landscape of the two species. Cells were harvested after treatment at the harvest point given in Table 2.4. To control noise and bias, samples were pooled, randomized, blocked and replicated. We treated each microarray as a statistical ‘block’ and randomly placed samples onto microarrays to balance variances from batch processes and positional effects [28]. Two biological replicates were produced for each bacterium under each treatment condition. The biological replicates were prepared by pooling samples according Figure 2.1; each biological replicate eventually was made from 4 different cell cultures grown at 2 different days [29].
Figure 2.1. Processing of biological replicates. In each day, two culture tubes under each treatment condition were combined and the RNA sample was extracted from the combined tubes. The extracted RNA samples from 2 different days (days 1 and 3, or days 2 and 4) were later combined to become the final replicate.

Two biological replicates were performed for each species under each treatment condition. The Qiagen RNeasy Mini Kit (#74104) was used to purify total RNA after on-column DNase digestion to remove DNA contamination (#79254). The total RNA were reverse-transcribed to cDNA using Life Technologies Superscript double stranded cDNA synthesis kit (#11917-020) with a random primer set (#48190-011). Residue RNA were then removed using RNase H (NEB #M0297). The cDNA samples were labeled using the NimbleGen One-Color Labeling Kit (#06370411001) and quantified using a Nanodrop ND-1000 spectrophotometer.
Table 2.4. Growth conditions of MG1655 and C58

<table>
<thead>
<tr>
<th>Name</th>
<th>Conditions</th>
<th>Harvest point</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>E. coli MG1655</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reached mid-log phase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>O.D. 600nm 0.6~0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cold shock</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reached half of O.D. 600nm of Standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heat shock</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low pH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UV treat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low carbon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low nitrogen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low C &amp; N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reached quarter of O.D. 600nm of Standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oxidative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reached half of O.D. 600nm of Standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Osmotic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Agrobacterium tumefaciens C58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reached mid-log phase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>O.D. 600nm 0.6~0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Conditions</th>
<th>Harvest point</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reached mid-log phase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>O.D. 600nm 0.6~0.8</td>
</tr>
</tbody>
</table>
Table 2.4 continued

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Reached half of O.D. 600nm of Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold Shock</td>
<td>Grown at 17°C for 13 hour then grown at 28°C in YEM medium</td>
<td></td>
</tr>
<tr>
<td>Heat shock</td>
<td>Grown at 40°C for 11 hour then grown at 28°C in YEM medium</td>
<td>Same as above</td>
</tr>
<tr>
<td>Low pH</td>
<td>Grown at 28°C in AB5.5 medium [31]</td>
<td>Reached mid-log phase</td>
</tr>
<tr>
<td>Low Iron</td>
<td>Grown at 28°C in AB7 without Fe medium (AB7 Fe-) [32]</td>
<td>Same as above</td>
</tr>
<tr>
<td>Oxidative</td>
<td>Grown at 28°C in 40 mL YEM medium with 130 μL 1% Hydrogen peroxide</td>
<td>Reached half of O.D. 600nm of Standard</td>
</tr>
<tr>
<td>Cold shock &amp; Oxidative</td>
<td>Grown at 17°C for 13 hour then grown at 28°C in YEM medium</td>
<td>Reached quarter of O.D. 600nm of Standard</td>
</tr>
<tr>
<td>Low pH &amp; Cold shock</td>
<td>Grown at 17°C for 13 hour then grown at 28°C in AB5.5 medium</td>
<td>Same as above</td>
</tr>
<tr>
<td>Low pH &amp; Heat shock</td>
<td>Grown at 40°C for 11 hour then grown at 28°C in AB5.5 medium</td>
<td>Same as above</td>
</tr>
<tr>
<td>Low pH &amp; Low Iron</td>
<td>Grown at 28°C in AB5.5 without Fe medium (AB5.5 Fe-)</td>
<td>Same as above</td>
</tr>
</tbody>
</table>

Microarray hybridizations were carried out on a NimbleGen Hybridization Workstation 4 (#05223652001) after dissolving labeled cDNA samples in the Hybridization Kit (#05583683001) with appropriate Sample Tracking Controls added (#05223512001). After the manufacturer recommended overnight hybridization (about 16
hours), microarrays were washed with the NimbleGen Wash Buffer Kit (#05584507001) and scanned using a GenePix 4100A Microarray Scanner at the maximum resolution of 5 µm for the Cy3 channel.

Scanned microarray images, which contained 4 microarrays on each chip, were processed using the NimbleGen DEVA 1.2.1 software [33]. The DEVA software aligned and anchored the microarray images using special alignment probes on the microarray surface and then split the images into 4 subarrays for the 4x72K NimbleGen layout before quantifying them into individual probe values. Although manual alignments can be performed, we found it unnecessary for all the microarray images processed. The DEVA software also provided automatic RMA normalizations (Robust Multi-Array Analysis) across each set of microarray data for MG1655 and C58 to reduce outliers and make data comparisons more meaningful. Two microarray samples were removed from the final data sets due to large bubbles on the microarray surface. In all, 38 microarray samples (18 for MG1655 and 20 for C58) were used for the following validation analyses. The data can be obtained from NCBI GEO database using the Super Series Access Number GSE61738

**Probe level consistency validations**

For each strain and treatment condition that has two successful biological replicates, we calculated Spearman’s correlation and concordance correlation coefficient (CCC) between them to validate the biological replicate consistency of the microarrays. Spearman’s correlation was used to measure the reproducibility of two replicates [34]. CCC was used to provide a better indicator of accuracy and precision of agreement between two biological replicates [35]. Biological replicates are more variable than technical replicates because the two biological samples included cultures that were independently grown,
harvested and subjected to microarray protocols, thus correlation between biological replicates can be as low as 30% [36].

Scatterplots for each pair of replicates with regression line and correlation coefficients are shown in Figure 2.2 for *E. coli* MG1655. Both statistical tests were performed with 95% confidence level. The oxidative condition showed the highest correlation values. The cold shock condition had the lowest correlation values, but its Spearman’s correlation value 0.7847 and CCC value 0.7676 still indicate high correlations between the replicates. Therefore, we conclude that the biological replicates of *E. coli* MG1655 under all treatment conditions are consistent enough to indicate that the MG1655 tiling microarray is reliable at the probe level.
Figure 2.2. Scatterplots with correlation values for 8 treatment replicate pairs of MG1655. On the graphs, ‘r’ denotes Spearman’s correlation and ‘ccc’ denotes concordance correlation coefficient.

Scatterplots for each pair of biological replicates with regression line and correlation coefficients for Agrobacterium tumefaciens C58 are similarly shown in Figure 2.3. They show that the two biological replicates under each treatment condition are consistent with each other with a high correlation value at the 95% confident level. The low pH & heat shock condition produced the highest correlation value of 0.9398. The cold shock and oxidative condition produced a lower correlation value of 0.7393, which is still high
enough to conclude the replicates are highly related. Therefore, we also conclude that the C58 tiling microarray is reliable at the probe level.

**Figure 2.3. Scatterplots with correlation values for the 10 treatment replicate pairs of C58.** On the graphs, ‘r’ denotes Spearman’s correlation and ‘ccc’ denotes concordance correlation coefficient.

**Gene level consistency validations**

Significant analysis of microarray (SAM) [37], [38] and one-way ANOVA [39] were conducted to detect differentially expressed genes in *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 under the 10 treatment conditions. As stated earlier, most of the conditions are stress conditions that can promote stress gene responses. The two statistical
analyses were performed to validate that the tiling microarrays can detect biologically significant gene expression changes. SAM detects differentially expressed genes across all 10 conditions for each bacterium. One-way ANOVA tests were performed for a few sets of probes targeting some known stress related genes to confirm that their means differ across the 10 conditions for each bacterium, which also signals stress responses of cells.

SAM was performed with unpaired two class (control and treatment), delta value of 0.06 and fold change of 2. It found 34 differentially expressed genes in *E. coli* MG1655, including 22 known stress response genes such as *dnaX* [40], *entF* [41], *groL* [42], and *copA* [43], [44]. For *Agrobacterium tumefaciens* C58, SAM was ran with unpaired two class (control and treatment), delta of 0.065 and fold change of 3 to limit the number of differentially expressed genes reported. It detected 46 differentially expressed genes, including 16 known stress genes such as *livJ* [45], *dadA* [46], and *rpoH* [47]. The one-way ANOVA tests were performed with small subsets of known stress response genes at the 95% confidential level. Stress response genes detected by SAM and/or confirmed by one-way ANOVA test are summarized in Table 2.5.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>dnaX</em></td>
<td>Temperature sensitive for replication and growth [40]</td>
<td>SAM, ANOVA</td>
</tr>
<tr>
<td><em>groL</em></td>
<td>Acid tolerance response [41]</td>
<td>SAM, ANOVA</td>
</tr>
<tr>
<td><em>hslO</em></td>
<td>Heat shock response [42]</td>
<td>SAM</td>
</tr>
<tr>
<td><em>entF</em></td>
<td>Transcriptional regulation in C and N limited cultures [43]</td>
<td>SAM, ANOVA</td>
</tr>
</tbody>
</table>
Table 2.5 continued

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function and Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>copA</strong></td>
<td>Essential element in copper homeostasis and copper proteins; involved in oxidative stress protection [44], [45]</td>
</tr>
<tr>
<td><strong>cusS</strong></td>
<td>Copper tolerance in anaerobic [45]</td>
</tr>
<tr>
<td><strong>nusA</strong></td>
<td>Cold shock response [46]</td>
</tr>
<tr>
<td><strong>uvrA</strong></td>
<td>DNA repair and SOS response [47]</td>
</tr>
<tr>
<td><strong>aceE</strong></td>
<td>Induce the oxidative and acid resistance gene yfjD [43]</td>
</tr>
<tr>
<td><strong>katF</strong></td>
<td>Control of catalase-hydroperoxidase [48]</td>
</tr>
<tr>
<td><strong>cydA</strong></td>
<td>Control cytochrome bd oxidase on LB but not on minimal medium [49]</td>
</tr>
<tr>
<td><strong>mreBCD</strong></td>
<td>Involved in cell shaping and osmotolerant [50]</td>
</tr>
</tbody>
</table>

**Agrobacterium tumefaciens C58**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function and Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cspA</strong></td>
<td>Cold shock protein [51]</td>
</tr>
<tr>
<td><strong>rpoH</strong></td>
<td>Temperature sensitive, control heat shock protein [52]</td>
</tr>
<tr>
<td><strong>groEL</strong></td>
<td>Heat shock protein (stress protein) [53]</td>
</tr>
<tr>
<td><strong>dnaK</strong></td>
<td>Heat shock protein [54] but also induced by other stresses [53]</td>
</tr>
<tr>
<td><strong>livJ</strong></td>
<td>ABC transporter associated with the uptake of metal ions and involved in antioxidative stress defense [55]</td>
</tr>
<tr>
<td><strong>dadA</strong></td>
<td>Catalyzes the oxidative deamination of D-amino acids [56]</td>
</tr>
<tr>
<td><strong>fepC</strong></td>
<td>Outermembrane receptor [57]</td>
</tr>
<tr>
<td><strong>virF</strong></td>
<td>Virulence genes induced under several stresses, such as acidic condition or mitomycin C attack [53]</td>
</tr>
<tr>
<td><strong>chvF</strong></td>
<td>Induced by acidic pH [38]</td>
</tr>
<tr>
<td><strong>sitF</strong></td>
<td>Related to iron uptake [58]</td>
</tr>
</tbody>
</table>

* Superscript † means gene families.
After ANOVA tests, multiple pairwise comparison tests (Tukey HSD [39] and Dunnett’s test [38]) were conducted at 95% confident level as post-hoc tests to find out which pairs of treatment conditions have distinctive stress gene expression differences. Differentially expressed stress genes tested and detected for each pair of conditions are listed in Table 2.6 for *E. coli* MG1655 and Table 2.7 for *Agrobacterium tumefaciens* C58. All differentially expressed stress genes confirmed by ANOVA were also found by the post hoc tests except the ‘fepC’ gene in *Agrobacterium tumefaciens* C58. The p-value of *fepC* from ANOVA is 0.00207 but Tukey HSD or Dunnett’s test cannot identify it as differentially expressed in all pairs of conditions. It may be inferred that the means for this gene in all conditions are different from each other and there is no pair of conditions that is significantly different to allow detection by the post-hoc tests.

### Table 2.6. Differentially expressed stress genes in each pair of conditions for MG1655.

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>CS</th>
<th>HS</th>
<th>pH</th>
<th>UV</th>
<th>C-</th>
<th>N-</th>
<th>C-N-</th>
<th>Oxi</th>
<th>Osmo</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SD**: uvrA
- **CS**: groL, nusA, aceE, gadC, uvrA, entF
- **HS**: uvrA, entF, livJ
- **pH**: cydA, cydA, cydA, mreBCD
**Table 2.6 continued**

<table>
<thead>
<tr>
<th>UV</th>
<th>C-</th>
<th>N-</th>
<th>C-N-</th>
<th>Oxi</th>
<th>Osmo</th>
</tr>
</thead>
</table>

**Table 2.7. Differentially expressed stress genes in each pair of conditions for C58.**

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>CS</th>
<th>HS</th>
<th>pH</th>
<th>Fe-</th>
<th>Oxi</th>
<th>CS &amp; Oxi</th>
<th>pH &amp; CS</th>
<th>pH &amp; HS</th>
<th>pH &amp; Fe-</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>groEL vir</td>
<td></td>
<td>sit</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>groEL</td>
<td>dnaK</td>
<td>groEL</td>
</tr>
<tr>
<td>Fe-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sit</td>
<td></td>
<td>sit</td>
<td></td>
</tr>
</tbody>
</table>
It is worth noting that SAM generally detected more differentially expressed stress response genes than ANOVA can confirm. For example, the heat shock response gene hslO was detected by SAM even though ANOVA was not able to confirm that its means are significantly different among the treatment conditions for E. coli MG1655. More interestingly, many stressful conditions triggered uvrA gene expressions, which is the SOS response gene in E. coli MG1655. For Agrobacterium tumefaciens C58, stress response genes are induced more by combined treatment conditions. For example, many stress response genes are differentially expressed not just by heat shock but by heat shock and low pH combined. We can conclude that the tiling microarrays detected sensible gene expression changes that conform to our expectation with regard to known cell stress response gene behaviors.
Novel transcript discoveries

One of the stated benefits of a tiling microarray is that it can detect unexpected expressions as well as annotated gene expressions. Indeed, the two tiling microarrays for *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 detected significant numbers of RNA expressions from non-gene-coding regions on the two genomes. For example, our tiling microarray detected all 65 non-coding regulatory RNAs (ncRNAs) annotated in GenBank report U00096.3 for *E. coli* MG1655. Among the 65 ncRNAs, *dicF* is computationally predicted to target the *hslV* gene according to the Bacterial Small Regulatory RNA database (BSRD) [58]. Since our tiling microarray covers both the *dicF* ncRNA and its target gene, we can calculate the correlation value between them is -0.4819401, which agrees with the predictions. Because the whole genome is monitored, when there are new predictions of such regulatory activities, the correlation values can be extracted from our tiling microarray data without having to design new experiments to validate the predictions. We have evidence of several ncRNA and target gene correlations, but it will take another publication to analyze these data and report their biological significance. It suffices to say here that the two tiling microarrays do allow novel transcript discoveries as we have anticipated.

Conclusions

In this work we have described the design strategies and validation experiments of two whole-genome tiling microarrays for *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 bacteria. The tiling microarrays are *thermodynamically optimal* for the two genomes based on the rigorous calculations conducted by the *PICKY* software [17]. This means that
all probes selected have maximum specificity toward their target genome regions and no additional probes can be added to the microarrays without jeopardizing its specificity under the given design constraints. On average, there is a unique microarray probe every 100 bp along the genomes to detect any transcripts coming from that region. Therefore, transcripts longer than 100 bp are likely to be detected by at least one tiling microarray probe. The 100 bp selection window can be adjusted upward or downward depending on the microarray probe count and user preferences, but we do not expect the selected probes to increase significantly when this window is reduced because most, if not all, thermodynamically optimal probes should have been found by PICKY at the current 100 bp window size given the ~40 bp non-overlapping probe length.

The tiling microarray probes can detect transcripts expressed from both strands of the genomes because most of the common cDNA conversion and labeling protocols automatically produced double-stranded DNAs from original RNA transcripts. Given gene annotation information and bioinformatic techniques such as gene predictions, determining the actual strand of expression is not difficult for most genes [59]. One can also use some other methods such as RT-PCR to confirm the expressing strand for a few difficult transcripts.

We believe tiling microarrays are useful for many gene expression studies, especially for novel and non-model species that have not been annotated yet. In fact, tiling microarrays can help identify novel gene expressions and facilitate the annotation effort of these species. Microarrays tend to produce data much faster (in just 2 days), can tolerate a few mismatched bases due to polymorphisms, tend not to be overwhelmed by excessive bacterial rRNAs as NGS does, and do not usually require sophisticated computing capacity
to interpret the data. The software used in this study and the data produced by the experiments are freely available to other researchers who may wish to design their tiling microarrays.

**Acknowledgments**

The authors would like to thank Dr. Kan Wang and Dr. Gwyn A Beattie for their valuable advices to this work. We also thank Drs. Kan Wang and Keunsub Lee for providing the *Agrobacterium tumefaciens* C58 strain and teaching us how to maintain these bacteria. We thank our colleagues Taryn Anderson and Matthew Hunt for their help in bacteria maintenance and RNA extraction, and Jennifer Chang for discussion about statistical methods. This work is supported by the National Science Foundation grant DBI-0850195.

**References**


33. “Roche NimbleGen | DNA Microarray Support.”


CHAPTER 3. LIMITATION OF THERMODYNAMIC ANALYSIS THAT PREDICT NON-CODING RNAs IN PROKARYOTES

Abstract

Bacterial small non-coding RNAs perform an important role in regulating their target genes. One class of ncRNA is trans-acting ncRNA, which regulate their target genes usually through imperfect base pairing with the targets and is usually encoded elsewhere on the genome. This property of trans-acting ncRNAs makes prediction and identification difficult. Our prediction method with thermodynamic analysis using PICKY is able to identify not only perfectly matched targets but also the closest non-targets with their melting temperature. Escherichia. coli MG1655 and Agrobacterium tumefaciens C58 were chosen to test the method, were de novo sequenced and assembled using next-generation sequencing (NGS) to improve prediction quality, and were grown under 10 different conditions to induce more transcriptomes. Whole-genome tiling microarray are designed with the newly assembled genome and generated to validate predicted ncRNAs using two bacterial cells grown under 10 different conditions. In addition, over-expression method is used to validate associations between predicted ncRNAs and their predicted target genes.

Introduction

Prokaryotic species express small non-coding RNAs (ncRNAs) to regulate their gene expression. ncRNAs can be classified as either cis- or trans-acting. cis-acting ncRNAs act in a perfectly matched manner to their target genes and are usually encoded on the anti-sense strands of the targets. On the contrary, trans-acting ncRNAs regulate their target genes
usually through imperfect base pairing with the targets and are usually encoded elsewhere on the genome. This property of trans-acting ncRNAs, i.e. forming imperfect base pairing with their targets, makes it difficult to predict and identify them solely based on sequence comparisons. We have developed a systematic method to identify trans-acting ncRNAs in prokaryotes. E. coli and A. tumefaciens have been chosen to test this novel method that is based on thermodynamic analysis and tilling microarrays. E. coli is a model organism for prokaryotes and ncRNAs in E. coli have been well studied. In contrast, few ncRNAs of A. tumefaciens are yet discovered. Therefore, choosing these two species allows us to both confirm our prediction methods and to discover novel ncRNAs.

The method starts with a whole-genome thermodynamic analysis to identify all imperfect but strong base-pairings between all annotated genes and some untranslated and intergenic regions from the same species; if these regions were to express RNA transcripts, the transcripts can bind strongly to the target genes. Therefore, these regions are potential encoding sites of trans-ncRNAs. If some predicted ncRNA sites indeed express RNA transcripts, their chance of being functional is high. To validate the predicted ncRNAs, two high density whole-genome tiling microarrays have been designed to cover all predicted ncRNAs and all annotated target genes and their anti-sense strands for the E. coli MG1655 strain and the A. tumefaciens C58 strain. To design the highest quality tiling microarrays, the MG1655 and C58 genomes have been independently sequenced and assembled using next-generation sequencing techniques. Several hundred single nucleotide polymorphisms (SNPs) have been found between the reference genomes of the two species and the newly assembled genomes. These SNPs are found to influence some target genes, some predicted ncRNAs and some microarray probes. Therefore, obtaining updated
genomic sequences matching the bacteria strains stored in our lab helps improve the sensitivity and quality of the subsequent validation experiments.

To validate and confirm the predicted ncRNAs, *E. coli* MG1655 and *A. tumefaciens* C58 strains were grown under 10 different laboratory conditions to trigger the expression of various indigenous ncRNAs in cells. The various RNA transcripts contained within the transcriptome were identified using the tiling microarrays. To validate association between the predicted ncRNAs and their predicted target genes, over-expression method was used. Plasmids containing inserted interesting ncRNAs and targets were built to be overly expressed using a designed RNA polymerase promoter and they are used to subsequently validate their associations. This procedure was conducted using precise quantitative PCR assays.

**Assembly**

It is essential that we have the correct genomic sequences of the MG1655 and C58 strains for ncRNA predictions, genome tiling microarray designs and all downstream data analyses. Although their reference genomic sequences can be downloaded from GenBank, we believe that there is an opportunity to improve the outcomes of our project by doing an independent *de novo* genome sequencing and assembly. This way, we can be certain that the bacteria strains in our lab will match the genomic sequences that will be subsequently using in our project. We have performed Next-Generation Sequencing (NGS) with 5 strains of *E. coli*, MG1655, W3110, JA221, M182, HT115(de3), and 1 strain of *A. tumefaciens* C58.
After extracting total DNA of 5 strains and obtaining the NGS reads from total DNAs, we attempted to assemble the three relevant genomes to the project: MG1655, W3110 and C58. We begin by comparing several de novo assembly tools for NGS data, including Velvet [1], ABysS [2], MIRA [3], and Zorro [4]. We are seeking better and more efficient assembly results, thus we initially tested their performance efficiency under the same conditions, and subsequently compared their assembly informatics such as the number of assembled contigs, the longest assembled contig length, average contig length, and the N50 length. We also considered practical issues such as the running time of each software program and their memory usage. We concluded that Velvet provides the best assembly result among all tested tools. Table 3.1 shows the results of comparison among the assembled MG1655 reads using the three major NGS assembly software tools we have tested.

**Table 3.1. Comparison results with k-mer 55**

<table>
<thead>
<tr>
<th></th>
<th>Velvet</th>
<th>ABysS</th>
<th>MIRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of contigs</td>
<td>867</td>
<td>902</td>
<td>634</td>
</tr>
<tr>
<td>Longest contig size</td>
<td>269,653</td>
<td>239,048</td>
<td>66</td>
</tr>
<tr>
<td>Avg. contig size</td>
<td>638.9</td>
<td>512.7</td>
<td>59</td>
</tr>
<tr>
<td>N50</td>
<td>57,717</td>
<td>48,967</td>
<td>17,382</td>
</tr>
</tbody>
</table>

Finding appropriate hash length is important to do de novo assembly using Velvet. After figuring out several optimal hash length value for each NGS read set, Velvet was conducted several times with several selected hash lengths to assemble NGS reads of 3 strains. Contigs in the assemblies obtained at unselected hash lengths were also used to close gaps. In addition to using Velvet, several other bioinformatics tools were needed to
finish the assembly. Minimus2, which is the part of the AMOS software package, was used to merge Velvet output contigs to form longer scaffolds [5]. Before merging Velvet contigs with Minimus2, we first used our own Perl scripts to remove redundant contigs. BLAT was then used to align merged contigs to the reference genomes [6]. The alignment information is important to correctly orient some contigs, find repeated contigs and fill in the gaps among aligned contigs. We also created our own Perl scripts to close sequencing gaps and finish the assembly.

The MG1655 and W3110 strains of *E. coli* were completely assembled. However, *E. coli* W3110 did not used for the study at all. The circular and linear chromosomes and the Ti plasmid of C58 were also completely assembled. The following table summarizes the results of the assembled NGS data sets.

**Table 3.2. Results of NGS Assembly**

<table>
<thead>
<tr>
<th></th>
<th>MG1655</th>
<th>W3110</th>
<th>C58</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of SNPs</td>
<td>265</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>Genome size</td>
<td>4,639,675</td>
<td>4,646,332</td>
<td>5,746,078</td>
</tr>
<tr>
<td>Overall identity</td>
<td>99.9943%</td>
<td>99.9960</td>
<td></td>
</tr>
</tbody>
</table>

### C58

- **Circular**: 56
- **Linear**: 142
- **Ti**: 5
All strains show very similar to GenBank reference sequences. The circular chromosome of the C58 strain produces the highest identity percentage to the reference sequence: 99.9980%. We have found a large empty area in the center of the At plasmid in C58. However, we cannot figure out what has caused the empty area in the At plasmid. There seems to be no NGS reads generated from this missing area. Therefore, the downloaded reference At plasmid sequence was used to replace the missing At plasmid in our assembly for C58.

We compared the previously predicted ncRNAs in MG1655 with the SNP positions discovered from the newly assembled genomes to estimate how many SNPs exist in predicted ncRNAs. In total, 181 out of 265 SNPs can be found in predicted ncRNAs and known ncRNAs (some ncRNAs overlap). Moreover, 78 of the 265 SNPs are found in mRNA genes and 34 among the 78 genes are targeted by the predicted ncRNAs. This proves that our effort in sequencing and assembling the genomes does improve the quality of the ncRNA predictions and subsequent tiling microarray designs.

**Prediction and Validation**

Thermodynamic analysis using PICKY [7] [8] was applied to predict ncRNAs in *E. coli* MG1655 and *A. tumefaciens* C58. Procedure of prediction in both strains is described as follows and Figure 3.1;
First, operon locations were identified on a newly assembled whole genome. We created a probe set by cutting each operon into 100 bp fragments with 20 bp overlaps and the created probes will be screened by PICKY. We screened these fragments against the whole genome and identified potential closest nontargets sites that can bind to them within a certain thermodynamic threshold. The following parameters were given to PICKY to discover potential closest nontargets: maximum match length 20, minimum match length 8, minimum sequence similarity 66%, and minimum melting temperature difference 5°C. All other PICKY parameters were taken at their default values, including the screening of both strands of each input whole genome to ensure the fragment specificity in either direction. The minimum melting temperature difference ensures that only closest nontarget sites
unique to the target fragments are selected. Selected nontargets were further filtered by their matched length and melting temperatures to the target fragments. Minimum value of matched length was given as input parameter and melting temperature of nontargets to the fragment compared with melting temperature of perfect targets to the same fragment then nontargets were discarded if the differences of melting temperatures were higher than 10°C. We only considered nontargets which have melting temperatures to fragments that were very close to melting temperature of perfect targets to same fragments. PICKY calculated melting temperatures of each fragment with its intended target and with its closest nontargets anywhere in the genome according to the thermodynamic nearest-neighbor models of prefect matches [9] and mismatches [10]–[13]. This feature of PICKY allowed us to predict trans-acting ncRNAs. The closest nontarget sites of 100bp fragment detected by PICKY with imposed parameter became candidates and many of these candidates were removed based on their lengths and melting temperature differences compared with perfected matched target of the same fragment.

For validation, we first calculated correlation coefficient values of pairs of the predicted ncRNA and False Discovery Rate (FDR) controlling was computed to provide more precise correlation coefficient values. Next, plasmids containing inserted validated ncRNAs and targets are constructed, overly expressed using a designed RNA polymerase promoter and used to subsequently validate their associations between predicted ncRNAs and their predicted targets. This procedure is accomplished using quantitative PCR assays. In total, 13 pairs of predicted ncRNAs and their predicted targets of E. coli MG1655 were selected for overexpression plasmid validation.
Results

Summarized results of predicted ncRNAs in both strains, MG1655 and C58 are shown in Table 3.3. 59,466 fragments with 100bp length in MG1655 and 63,941 fragments with 100bp length in C58 were created based on operon positions. After running with PICKY and filtering with threshold of melting temperatures and length, 541 potential ncRNAs in MG1655 and 220 potential ncRNAs in C58 were predicted. PICKY calculated sequence specificity between fragments and the entire genome based on thermodynamics. This means that PICKY predicted ncRNAs as well as its targets.

Table 3.3. Predicted ncRNAs of E. Coli and A. Tumefaciens

<table>
<thead>
<tr>
<th></th>
<th>E. Coli MG 1655</th>
<th>A. Tumefaciens C58</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Circular</td>
<td>Linear</td>
</tr>
<tr>
<td># of fragments</td>
<td>59,466</td>
<td>31,909</td>
</tr>
<tr>
<td>After Filtering</td>
<td>541</td>
<td>91</td>
</tr>
<tr>
<td>&lt; FDR 5%</td>
<td>310</td>
<td>5</td>
</tr>
</tbody>
</table>

Pearson’s correlations between predicted ncRNAs and their targets are calculated using the R environment (http://www.r-project.org) and then, several statistical steps were conducted to control false positives, i.e., to control the false discovery rate (FDR). Fisher’s Z transformation was performed to transform correlation coefficient values to a normal distribution, and then a test for whether the correlation coefficient is equal to zero for each predicted pair of ncRNA and target region was conducted with the p-value calculated. Finally, the resulting p-values were used to compute q-value with the ‘qvalue’ package in Bioconductor[14]. When the FDR was controlled at 5%, potential ncRNAs in both strains were ruled out if their q-values were greater than 0.05. Correlation coefficient
values of *E. coli* are shown in Figure 3.2 and correlation coefficient values of *A. tumefaciens* are shown in Figure 3.3. Cutoff values of correlations are lower than 40% in *E. coli* and a cutoff value of correlation is higher than 60% in *A. tumefaciens* while FDR controlling at 5% Figure 3.2 showed that correlations between some predicted ncRNAs and their targets with or without FDR controlling were located together near at ±0.4 points in *E. coli*. 
Figure 3.2. Pearson’s correlation between predicted ncRNAs and their targets in *E. coli* MG1655. Correlations above FDR-cutoff value are shown as green and Correlation below FDR-cutoff value are shown as red orange.
Figure 3.3. Pearson’s correlation between predicted ncRNAs and their targets in *A. Tumefaciens* C58.

310 predicted ncRNAs in MG1655 and only 5 predicted ncRNAs of circular and 2 predicted ncRNAs of linear chromosome in C58 remained with FDR controlling at 5% and there is no negative correlation value in C58, even though the cutoff value of correlation in *A. tumefaciens* is high. Small numbers of predicted ncRNAs in C58 and low correlation value
of predicted ncRNAs and their targets in MG1655 might infer that predictions were not conducted correctly.

Even though the correlation values are not strong, we continued validation procedures. Predicted ncRNAs with FDR controlling at 5% in both strains were compared to already known ncRNAs databases entries, such as GeneBank U00093.3 or Bacterial Small Regulatory RNA Database (BSRD) for E. coli [15]. Only 2 entries from BSRD matched with our predicted ncRNAs in E. coli. ncRNAs in A. tumefaciens were compared with other predicted and validated ncRNA entries in Lee et al. 2013 [16]. Our predicted ncRNAs in A. tumefaciens did not match with any ncRNAs in the paper.

We decided to validate our prediction results further using over-expression method even though our predicted ncRNAs matched only a few already known ones, because, we believe that there is still a possibility that our predicted ncRNAs are totally novel ncRNAs. To evaluate interactions between our predicted ncRNAs and their predicted targets, over-expression plasmids with induced promoters were constructed for selected predicted pairs in E. coli. In total, 13 predicted ncRNAs and their targets were selected for E. coli. Five of these predicted ncRNAs were selected based on their target genes, being related to stress response. Other 8 predicted ncRNAs were chosen based on their strong correlation values, either positively or negatively. Fragments of selected 13 ncRNAs and 13 target regions on their target genes were prepared by gBlock Gene Fragments at Integrated DNA Technologies (https://www.idtdna.com/site) and they were inserted into pGEM-3Zf(-) vector for over-expression. Promoters and terminators for selected gBlock pairs were designed to promote overexpression. PCR assays were conducted to evaluate expression. Constructs were transformed into E. coli DH5α. A strain with on empty pGEM-3Zf(-) vector
was grown as a control with 13 inserted and transformed ones. All samples were grown with and without IPTG. IPTG was used to induce expression. The expression of 13 selected ncRNAs and their targets are shown in Figure 3.4. To measure expression levels, normalized relative ratio of gene expression was calculated based on the $2^{(-\Delta \Delta C(T))}$ method [17] using experimental calibration data and qPCR results.

![Figure 3.4](image)

**Figure 3.4.** qPCR expression levels of 13 selected over-expressed ncRNAs and their target genes. 'insert' denotes cloned constructs and 'pGem' denotes empty vectors.

Results in Figure 3.4 did not show consistency regarding IPTG treatment. Samples with IPTG were not expressed higher than samples without IPTG. ncRNA samples with or without IPTG did not alter their regulatory targets negatively or positively with
consistency. Some ncRNA samples like ncRNA5 regulated their targets negatively but IPTG treatment did not working well at replicate 1 of ncRNA5. In most samples, replicates of them were not consistent with each other. Results of predicted ncRNAs in E. coli were not conclusive; therefore we did not continue this validation step for A. Tumefaciens.

**Conclusion**

Our prediction method with thermodynamic approach using PICKY is mainly based on melting temperatures of base pairing regions between ncRNAs and their targets. PICKY is able to identify not only perfectly matched target but also closest non-target of imperfect matches and reports their melting temperatures. We selected closest non-targets which have melting temperatures to their target fragments differing from the melting temperature of matched target to the same fragments of less than 10°C. This threshold value is coming from an assumption that base paring between trans-acting ncRNAs and their targets are strong and stable even though they form imperfect base paring. However, results of our predictions with thermodynamic approach did not provide strong evidence through validations; comparison with known databases provided few matches and results of over-expression using qPCR were not conclusive.

After failure to validate our predicted ncRNAs and their targets, we analyzed melting temperatures of known ncRNAs in GenBank data set to their targets which are predicted by PICKY. This analysis disclosed that melting temperatures of known ncRNAs to their target are variable, thus we could not find a specific range of melting temperature for the prediction threshold. We concluded that the condition of our thermodynamic approach,
that closest non-targets that have melting temperatures that are very close to the melting temperature of matched target, was not a useful selection criterion.

Trans-acting ncRNAs regulate target genes through partial complementary to one or more mRNA molecules and are encoded elsewhere on the genome. These ncRNAs form base pairs with their target mRNAs, perfectly or imperfectly, at short regions such as 6 to 12 bp [18]. Our thermodynamic analysis software PICKY allows identification of such base paring regions. However, the small region of base paring between ncRNAs and their target mRNA makes melting temperatures of the regions variable. Therefore, it is difficult to predict trans-acting ncRNAs solely based on the thermodynamic analysis and the lessons from these results promoted us to consider other elements to predict bacterial small non-coding RNA.

**Material and Methods**

**Total DNA extraction**

All 5 strains of *E. coli* were grown at 37 °C in Luria-Bertani (LB) medium and the C58 strain of *A. tumefaciens* was grown at 28 °C in YEP medium. The QIAGEN DNeasy extract kit was used to extract total DNA from all strains. Qubit 2.0 Flurometer was used to precisely quantify DNA concentration in our samples and the Experion DNA 12K Analysis Kit was used to check the DNA quality. The total DNA was eluted in 100 μL buffer and 50 μL of that was sent for Next-Generation Sequencing.

**Prediction using PICKY**

The software PICKY was used to predict ncRNAs as well as potential gene targets of the predicted ncRNA candidates in prokaryote. The whole *E. coli* and *A. tumefaciens*
genome sequences were loaded twice into PICKY as targets; one of the loaded instances for both genomes was reverse-complemented to represent the reverse strand. The Examination function in PICKY was used to examine the 100bp fragments as input to determine if they matched to any regions in the genomes. The following parameters of the Examine function was used: maximum match length 20, minimum match length 8, minimum sequence similarity 66%, and minimum melting temperature difference 5°C; we screened forward strands only, and used the better salt effect equation. All other PICKY parameters were taken at their default values. PICKY analysis results were saved with maximum reported nontargets set to 20 and maximum temperature difference set to 25°C.

**Microarray**

We use the whole-genome tiling microarray described in chapter 2-experiment protocol.

**Overexpression (Recombination)**

13 selected ncRNAs and 13 their targets are converted into up to 250bp gBlocks Gene Fragments by IDT (https://www.idtdna.com/pages/products/genes/gblocks-gene-fragments). Each fragment had restriction sites for BamHI and EcoRI enzyme and 5bp extra nucleotides beyond the restriction sites at both 3’- and 5’- ends. pGEM-3Zf(-) plasmid was chosen as a host vector. Each synthesized fragments were annealed. pGEM-3Zf(-) plasmids were purified using Qiagen QIAprep Miniprep Kit (Catalog #27104) and digested using BamHI and EcoRI restriction enzymes. Next, the prepared synthesized fragments and digested pGEM-3Zf(-) plasmids were ligated. The ligated plasmids were transformed into competent *E. coli* DH5α cells (NEB Catalog #C2988J). Finally ampicillin, IPTG and XGAL
were added into LB agar plates for white-blue screening and transformed cells were grown on the plates. After selecting an appropriate number of transformed *E. coli* colonies using white-blue screening, we grew them in LB + ampicillin medium until the cell culture concentration reached to OD 0.4. DNAs of the grown cells were extracted using the Qiagen QIAquick Gel Extraction Kit (Catalog #28704). Purified DNAs were sent for DNA sequencing to confirm they had the expected inserts.

**Quantitative PCR**

The Qiagen RNeasy Mini Kit (#74104) was used to purify total RNA after on-column DNase digestion to remove DNA contaminations (#79254). The total RNA was reverse-transcribed to cDNA using Life Technologies SuperScript® III First-Strand Synthesis SuperMix for qRT-PCR kit (#11917-020) with a random primer set (#48190-011). Next, the generated cDNA was amplified by Roche LightCycler Real-Time PCR system.

**References**


CHAPTER 4. PREDICTION OF HFQ-BINDING REGULATORY RNAS IN PROKARYOTES BASED ON THERMODYNAMIC AND STRUCTURAL ANALYSES

Modified from a paper to be submitted to RNA

Hyejin Cho1, #, Jennifer Chang1, Peng Liu2 and Hui-Hsien Chou1, 3, *

Abstract

Non-coding regulatory RNAs (ncRNAs) regulate a host of gene functions in prokaryotes. Some ncRNAs have been identified experimentally but many are yet to be found. Computational predictions facilitate the discovery of ncRNAs, allowing more focused experiments to be designed to validate the predictions. In this work, we present a systematic, novel 5-step approach to predicting Hfq-binding ncRNAs in *Escherichia coli* and *Agrobacterium tumefaciens* based on thermodynamic analyses and known structural properties of this class of ncRNAs. We matched our predicted results to known ncRNAs, measured the correlation coefficient values between ncRNA candidates and their predicted targets using a whole-genome tiling microarray, and confirmed our results using 3 other ncRNA identification software tools. We also compared the ncRNA targets predicted by our method to known ncRNA targets and performed a gene ontology network analysis; both analyses show that many predicted target genes that are controlled by the same ncRNA

1 Department of Genetics, Development and Cell Biology, Iowa State University

# Primary researcher and author
2 Department of Statistics, Iowa State University
3 Department of Computer Science, Iowa State University
* Author for correspondence
candidate may serve similar biological functions. Our prediction method is generally applicable to other prokaryotic species and may help advance ncRNA research in prokaryotes.

**Introduction**

Prokaryotic species express small non-coding RNAs (ncRNAs) to regulate their gene expressions. ncRNAs can be classified as either *cis-* or *trans-*acting. *cis-*ncRNAs perfectly complement their target genes and are usually encoded on the anti-sense strands of the targets. On the contrary, *trans-*ncRNAs regulate their target genes through short and often imperfect base-pairings with the targets, and are usually encoded elsewhere on the genome. Some *trans-*acting ncRNAs are Hfq-binding ncRNAs that depend on the chaperon protein Hfq for activation [1]. In addition to the stem loop common to many ncRNAs, Hfq-binding ncRNAs share 3 special regions on their single-strand stretches (Figure 4.1) [2]. The Terminator region contains the poly-U tail which is fundamental for Hfq function; shortening or removing the poly-U tail eliminates Hfq-binding ability [3]. The Hfq site is for Hfq protein binding and is generally AU-rich. Finally, the Seed region recognizes targets and directly base-pairs with mRNAs. Hfq protein facilitates ncRNA and mRNA target base-pairing [4].
Figure 4.1. The common structures of Hfq-binding ncRNA exemplified by *rybB*. The Seed Region recognizes and binds target mRNA. The Hfq Site and Terminator are required for ncRNA functions.

We have developed a systematic method to identify Hfq-binding ncRNAs based on the poly-U regions found on the genome, thermodynamic ncRNA-target binding analyses, and predicted RNA secondary structures. *E. coli* MG1655 and *A. tumefaciens* were chosen to test this novel method. *E. coli* is a model organism for prokaryotes and ncRNAs in *E. coli* have been well studied, in contrast, ncRNAs of *A. tumefaciens* are not yet discovered many. Therefore, choosing these two species may allow us to both confirm our prediction methods and to discover novel ncRNAs. Our method started by finding poly-U regions encoded by either strand of the genome. After finding these poly-U regions, a whole-
genome thermodynamic analysis is performed to identify all short base-pairings (6~12 base-pairs, or bp) between all annotated genes in both strains and all potential seed regions upstream from the poly-U regions; if any seed region expressed ncRNA transcripts, the transcripts could potentially bind to target genes. Subsequently, secondary RNA structures were predicted for all identified candidate ncRNAs and used as a filter based on the 3 required structural regions given in Figure 4.1.

Our method correctly predicted all 14 known Hfq-binding ncRNAs in *E. coli* (http://www.ncbi.nlm.nih.gov/nuccore/U00096.3). To confirm the predicted novel ncRNAs, a high density whole-genome tiling microarrays which covers all predicted ncRNAs and all annotated target genes in *E. coli* MG1655 and *A. tumefaciens* C58 were used to measure correlation coefficient values between predicted ncRNAs and their predicted targets. For further validation, our prediction results were confirmed by 3 existing ncRNA identification software tools and network analyses were performed to highlight interesting ncRNA-target groups in *E. coli* and *A. tumefaciens*.

**Results**

**Hfq-binding ncRNA prediction**

For ncRNA predictions in prokaryotes, we followed 5 steps that included thermodynamic and structural analyses. **PICKY** is an efficient software tool initially developed for designing optimal oligonucleotide probes for microarray purposes, but it can also be used in ncRNA thermodynamic analysis [5]–[7]. **PICKY** compares all ncRNA candidates with all input genes and reports the highest matches found above a user-defined threshold. **ContextFold** is a RNA secondary structure prediction tool [8] and was
chosen among many RNA structure prediction programs because it produces the best secondary structure predictions of short RNAs (20-200 bp) and medium-sized RNAs (201-800 bp) [9]. ncRNAs in prokaryotes are 70 to 500 bp long [10], which is within the optimal length for ContextFold prediction. The steps to predict Hfq-binding ncRNAs with poly-U tails is described below:

1) **Identifying poly-U tail regions.** Poly-U tails that are 8 bp or longer effectively terminate ncRNA transcription, facilitate Hfq-binding, and regulate target genes [3]. When the poly-U tail is shortened to 6 bp or less, the Hfq-binding ncRNA regulation of target genes is abolished. [3]. Therefore, our prediction started with finding all poly-U regions at least 7 bp long as encoded by both strands of the genome. To determine the best range upstream from each poly-U region to search, we tested sizes of 50 bp, 75 bp, and 100 bp since the average size of the 14 known *E. coli* ncRNAs is 110 bp. We found that the 75 bp length identified all 14 known *E. coli* Hfq-binding ncRNAs. Therefore, we chose the 75 bp region size that was upstream of the identified poly-U sites and subjected them to the following thermodynamic analysis step.

2) **Initial thermodynamic analysis.** Regions anchored by poly-U tails were potential encoding sites of Hfq-binding ncRNAs; if these regions expressed ncRNA transcripts, the transcripts should bind to their target genes. It is known that trans-ncRNAs form short and sometimes imperfect but stable base-pairings in the 7–12 bp range with their target genes [11]. Therefore, a whole-genome thermodynamic analysis using the Picky software was performed to identify all such short but stable base-pairings between all annotated *E. coli* genes and all identified poly-U anchored regions. After the thermodynamic analysis, candidate poly-U anchored regions that did not match any
annotated genes were removed, as well as regions that resided inside annotated gene-encoding regions because they were less likely to be trans-ncRNAs [12].

3) **Merging of potential ncRNAs.** The previous step identified short ncRNA candidates in the 100 bp range, but some ncRNAs in both strains are known to be longer than 300 bp. Therefore, predicted candidate ncRNAs in close proximity to each other within 300 bp were merged into one region and anchored by the right-most (3’-end) poly-U tail.

4) **Hfq-binding ncRNA structural analysis.** As explained earlier and demonstrated in Figure 4.1, Hfq-binding ncRNAs share 3 common structure regions: the Seed region, Hfq-binding site and poly-U Terminator. Therefore, we used the ContextFold software to predict candidate ncRNA secondary structures and utilized the structure information for further filtration. ContextFold produces structure predictions with the dot-bracket format. The matched parentheses indicate base-pairing regions while the dots represent single-stranded stretches. One example of the predicted structures of ncRNAs is shown in Figure 4.2. The seed region should be at least 6 bp long, the Hfq binding site should be at least 7 bp long and the poly-U tail region should remain as single stranded and at least 7 bp long. Therefore, candidate ncRNAs were screened for these parameters and candidates that did not satisfy all these criteria were removed.
Figure 4.2. Example secondary structure of a predicted Hfq-binding ncRNA with poly-U tail. The RNA sequence and its associated ConxtexFold output are given as well. In ConxtexFold output, '(' and ')' denote stem and '. .' denotes unbound RNA nucleotide. The ConxtexFold predicted secondary structure is visualized using PseudoViewer [13].
5) **Second thermodynamic analysis.** After the structural analysis step, the seed regions of ncRNAs were predicted by PICKY again, since some candidate ncRNAs were merged or removed. ncRNA candidates were discarded if their PICKY predicted seed regions were not located on the single-stranded stretches of the secondary structures or their target sequences were not genes.

Predictions of ncRNAs on the *E. coli* and *A. tumefaciens* genomes can start with 7 bp minimal poly-U tails or the more stringent 8 bp minimal poly-U tail. The results are summarized in Table 4.1. Although differing by just 1 bp, the 7 bp screening produced 4 to 7 times more ncRNA candidates at each step than the 8 bp screening in *E. coli*. After passing all 5 prediction steps for *E. coli*, there were 410 ncRNA candidates with 7 bp or longer poly-U tail (abbrev. 7-bp ncRNAs), and 98 ncRNA candidates with 8 bp or longer poly-U tail (abbrev. 8-bp ncRNAs). Interestingly, the 98 8-bp ncRNAs were not a proper subset of the 410 7-bp ncRNAs because of ncRNA candidate merging in Step 3 and removal in Step 4. Our novel predictions were compared to known Hfq-binding ncRNAs reported in GenBank ([http://www.ncbi.nlm.nih.gov/nuccore/U00096.3](http://www.ncbi.nlm.nih.gov/nuccore/U00096.3)) and showed promising results — All 14 known Hfq-binding ncRNAs with 7 bp or longer poly-U tails were correctly identified by our prediction methods.

The 5 prediction steps were also conducted for *A. tumefaciens* C58. There were 84 7-bp ncRNAs in circular chromosome and 66 7-bp ncRNAs in linear chromosome, 11 7-bp ncRNAs in the At plasmid and 9 7-bp ncRNAs in the Ti plasmid. In addition, there were 16 8-bp ncRNAs in circular chromosome and 7 8-bp ncRNAs in linear chromosome. Our initial predictions were compared to known ncRNAs in the paper Lee *et al.* 2013 [14]. The paper did not indicate which ncRNAs are Hfq-binding ncRNAs, thus sequences of ncRNAs from
the paper were screened with poly-U regions and 3 ncRNAs with 7bp poly-U tail and 1
ncRNA with 8 bp poly-U tails in circular chromosome, and 1 ncRNA with 7bp poly-U tail in
linear chromosome were found. Known ncRNAs encoded by both plasmids, as reported in
the paper, do not have poly-U tails. — All 5 known ncRNAs with poly-U tails were correctly
identified by our prediction.

Table 4.1. Results of Hfq-binding ncRNA prediction

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>7-bp or longer poly-U</th>
<th>8-bp or longer poly-U</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. coli MG1655</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td># of poly-U regions</td>
<td>1414</td>
<td>224</td>
</tr>
<tr>
<td>2</td>
<td>Initial thermodynamic analysis</td>
<td>1148</td>
<td>221</td>
</tr>
<tr>
<td>3</td>
<td>ncRNA candidate merging</td>
<td>721</td>
<td>149</td>
</tr>
<tr>
<td>4</td>
<td>Secondary structural analysis</td>
<td>471</td>
<td>112</td>
</tr>
<tr>
<td>5</td>
<td>Second thermodynamic analysis</td>
<td>410</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Predicted ncRNA candidates</td>
<td>410</td>
<td>98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>7-bp or longer poly-U</th>
<th>8-bp or longer poly-U</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cc⁺       ln⁺</td>
<td>At⁺      Ti⁺</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. tumefaciens C58</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td># of poly-U regions</td>
<td>242</td>
<td>181</td>
</tr>
<tr>
<td>2</td>
<td>Initial thermodynamic analysis</td>
<td>239</td>
<td>129</td>
</tr>
<tr>
<td>3</td>
<td>ncRNA candidate merging</td>
<td>149</td>
<td>110</td>
</tr>
</tbody>
</table>
Table 4.1 continued

<table>
<thead>
<tr>
<th>4</th>
<th>Secondary structural analysis</th>
<th>105</th>
<th>76</th>
<th>16</th>
<th>13</th>
<th>18</th>
<th>9</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Second thermodynamic analysis</td>
<td>84</td>
<td>66</td>
<td>11</td>
<td>9</td>
<td>16</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Predicted ncRNA candidates</td>
<td>84</td>
<td>66</td>
<td>11</td>
<td>9</td>
<td>16</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

* ’cc’ denotes circular chromosome; ‘ln’ denotes linear chromosome; ‘Ti’ denotes Ti plasmid; ‘At’ denotes At plasmid

**Microarray validations**

To validate our predictions, we applied a whole-genome tiling microarrays to measure *E. coli* and *A. tumefaciens* transcriptome changes under 10 different stress conditions. The high-density tiling microarray was designed using Picky and has on average one genome-specific probe every 100 bp along the *E. coli* and *A. tumefaciens* genomes (H) Cho and HH Chou, in prep.). Therefore, all predicted ncRNAs and all genes in both strains could be monitored by the tiling microarray. The 10 different stress conditions were designed and fine-tuned to trigger large-scale transcriptomic changes, allowing us to measure the expression and correlation of predicted ncRNAs and their targets. Using R (http://www.r-project.org), we calculated the Pearson correlations of each predicted ncRNA with its targeted regions, genes that include those targeted regions, and operons that include the genes, respectively. Because prokaryotic genes are expressed in operons, ncRNAs should influence the whole operon, not just individual genes. The correlations between candidate ncRNAs and their predicted target regions, genes, and operons are shown in Figure 4.3. Bacterial small ncRNAs can have either positive or negative regulatory roles in a cell [15], so their correlation to target genes can be positive or negative.
To control false positives, i.e., to control the false discovery rate (FDR), several statistical steps were conducted. Fisher’s Z transformation was performed to transform correlation coefficient values to be normally distributed, and then a test for whether the correlation coefficient was equal to zero for each predicted pair of ncRNA and target region was conducted with the p-value calculated. Finally, the resulting p-values were used to compute q-values using the ‘qvalue’ package in R/Bioconductor [16] to control the FDR. While the FDR was controlled at 5%, we identified that 193 predicted 7-bp ncRNAs, or about 47% had significant non-zero correlation coefficients with their target regions in *E. coli*. For 8-bp ncRNAs in *E. coli*, we identified 56 (57%) predicted ncRNAs that had significant non-zero correlation coefficients with their target regions while controlling the
FDR at 5%. All FDR controlled ncRNA candidates in E. coli had a high correlation value above 55.1%, but 8-bp ncRNAs had cutoff values slightly higher than those of 7-bp ncRNAs and its percentage remaining after FDR was also higher than 7-bp. Therefore, 8-bp ncRNAs showed higher precision. One of the known 8-bp ncRNAs did not pass the FDR control, because FDR control certainly does not guarantee 100% power and hence cannot prevent false-negatives [17]. For A. tumefaciens, the results of FDR control are shown in Table 4.2 and Figure 4.4. While the FDR was controlled at 5%, only 5 predicted 7-bp ncRNAs in circular chromosome, 3 predicted 7-bp ncRNAs in linear chromosome, 5 predicted 7-bp ncRNA in At plasmid, and 1 predicted 7-bp ncRNAs in Ti plasmid remained and only 1 predicted 8-bp ncRNA in circular chromosome remained. However, ncRNAs include all known ncRNAs with poly-U tails and their correlation values were all higher than 56.7% while controlling the FDR at 5%. Interestingly, linear chromosome had about 6 times more predicted ncRNAs than At plasmid before FDR control, but predicted ncRNA numbers of linear chromosome after FDR control dropped sharply and became smaller than predicted ncRNA numbers of At plasmid. The tiling microarray results showed that our predicted Hfq-binding ncRNAs and their predicted targets have strong relationship either positively or negatively.

<table>
<thead>
<tr>
<th></th>
<th>7-bp ncRNAs</th>
<th>8-bp ncRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. coli MG1655</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original predictions</td>
<td>410</td>
<td>98</td>
</tr>
<tr>
<td>FDR control &lt; 5%</td>
<td><strong>193 (47%)</strong></td>
<td><strong>56 (57%)</strong></td>
</tr>
<tr>
<td>Match to known ncRNAs</td>
<td>14 out of 14</td>
<td>7 out of 8</td>
</tr>
</tbody>
</table>

Table 4.2. Results of predicted Hfq-binding ncRNAs after FDR control at 5%
Table 4.2 continued

<table>
<thead>
<tr>
<th></th>
<th>cc⁺</th>
<th>ln⁺</th>
<th>At⁺</th>
<th>Ti⁺</th>
<th>cc⁺</th>
<th>ln⁺</th>
<th>At⁺</th>
<th>Ti⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original predictions</td>
<td>85</td>
<td>67</td>
<td>11</td>
<td>9</td>
<td>16</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>After FDR control</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* ‘cc’ denotes circular chromosome; ‘ln’ denotes linear chromosome; ‘Ti’ denotes Ti plasmid; ‘At’ denotes At plasmid

Figure 4.4. Correlations between predicted Hfq-binding 7-bp or 8-bp ncRNAs and their predicted targets in *Agrobacterium tumefaciens*. 
Comparison to the other ncRNA identification tools

We compared our results with 3 other ncRNA identification tools: GraPPLE, RNAcon, and CPC. Each tool takes different approaches to identifying ncRNAs. GraPPLE is a web-based service to identify ncRNA based on graph properties [18]. The input sequences are folded into RNA secondary structures using the ViennaRNA package [19], and then the structures are converted to graphs using the igraph package in R [20]. Subsequently, GraPPLE uses a Support Vector Machine (SVM) [21] to classify the graphs into 46 Rfam families [22], which can be grouped to three broader functional categories: cis-regulatory elements, self-splicing RNAs and non-coding RNA — the last category is what we sought to confirm. GraPPLE confirmed 174 out of the 193 FDR-controlled 7-bp ncRNA predictions of *E. coli*. Also, GraPPLE confirmed 52 out of the 56 FDR-controlled 8-bp ncRNA predictions of *E. coli*. For *A. tumefaciens*, GraPPLE confirmed 12 out of 14 FDR-controlled ncRNA predictions.

RNAcon identifies ncRNA based on the sequence compositions of mono-nucleotide, di-nucleotide, tri-nucleotide, and tetra-nucleotide units [23]. RNAcon has good accuracy; its authors claimed that the sensitivity and specificity of RNAcon are higher than GraPPLE. Our 193 FDR-controlled 7-bp ncRNA predictions of *E. coli* were all confirmed by RNAcon as ncRNAs. Furthermore, 52 of our 56 FDR-controlled 8-bp ncRNA predictions of *E. coli* were confirmed as ncRNAs by RNAcon. For *A. tumefaciens*, RNAcon also confirmed 12 out of 14 FDR-controlled ncRNA predictions.

The third tool is the Coding Potential Calculator (CPC). CPC distinguishes coding RNAs from noncoding RNAs according to six biological sequence features using an alternative Support Vector Machine-based classifier [24]. CPC confirmed 191 of the 193
Hfq-binding 7-bp ncRNAs we have predicted in *E.coli*. Also, CPC confirmed 55 of the 56 Hfq-binding 8-bp ncRNAs in *E. coli*. In addition, CPC confirmed 12 of the predicted ncRNAs in *A. tumefaciens*. The results of all the comparisons are shown in Table 4.3; 172 out of 193 (89.12%) FDR-controlled 7-bp ncRNAs were confirmed as ncRNAs by all 3 software tools, and 47 out of 56 (83.93%) FDR-controlled 8-bp ncRNAs were confirmed as ncRNAs by all 3 tools in *E. coli*. 10 out of 15 (66.7%) FDR-controlled ncRNAs were confirmed as ncRNAs by all 3 tools in *A. tumefaciens*.

**Table 4.3. Confirming ncRNA predictions by 3 other ncRNA-identification tools**

<table>
<thead>
<tr>
<th></th>
<th>7-bp ncRNAs of <em>E. coli</em></th>
<th>8-bp ncRNAs of <em>E. coli</em></th>
<th><em>A. tumefaciens</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR-controlled ncRNA predictions</td>
<td>193</td>
<td>56</td>
<td>15</td>
</tr>
<tr>
<td>GraPPELE confirmed</td>
<td>174</td>
<td>90.2%</td>
<td>52</td>
</tr>
<tr>
<td>RNAcon confirmed</td>
<td>193</td>
<td>100%</td>
<td>52</td>
</tr>
<tr>
<td>CPC confirmed</td>
<td>191</td>
<td>98.96%</td>
<td>55</td>
</tr>
<tr>
<td>ncRNAs confirmed by all above tools</td>
<td>172</td>
<td>89.12%</td>
<td>46</td>
</tr>
</tbody>
</table>

**ncRNA target gene predictions**

Our Hfq-binding ncRNA prediction approach analyzes all potential hybridizations between ncRNA candidates and the whole *E. coli* genome. A distinctive advantage of this approach is that we can predict ncRNAs as well as their target regions on the genome. It is not easy to find imperfect but stable target bindings with predicted ncRNAs using sequence comparison tools like BLAST [25], because such regions are too short (7–12 bp). PICKY allows such potential binding sites to be identified according to thermodynamics, not just sequence similarity.
Therefore, we can also test if our ncRNA predictions make sense by looking at the biological functions of their predicted target genes, i.e., to see if their functions are consistent with the biological functions associated with regulatory ncRNAs. One of our predicted ncRNAs matches to the known micC which has been identified to regulate the expression of the ompC gene [26]. OmpC is an important outer membrane protein related to tetracycline resistance in E. coli and is involved in its export mechanism [27], [28]. Our predicted targets for micC include ygdD, acrB, and yoaE but not OmpC itself. However, ygdD is an inner membrane protein with four predicted transmembrane domains [29], acrB is an RND-type inner-membrane associated proton-substrate antiporter [30], and yoaE is an inner membrane protein with seven predicted transmembrane domains [29]. Therefore, our predicted targets of micC are all localized to the membrane and some are involved in transport, which are similar to the OmpC function.

Our predicted ncRNAs also include fnrS which is related to iron ion binding and metal ion binding based on functional enrichment analysis as well as anaerobiosis metabolism [31]. FnrS down-regulates sodA, sodB, cydDC, and metE during anaerobic growth [32]. Instead of the known targets, our method predicted the targets of fnrS are entA, entE, and narY. EntA is a member of the short-chain oxidoreductase family [33] and EntE catalyzes the condensation of 2,3-dihydroxybenzoate (DHB) [34]. Both EntA and EntE are involves in iron uptake and iron transport [35]. Finally, narY encodes nitrate reductase Z [36]. The narY gene is activated when oxygen concentration is reach to low [37]. While our predictions do not include the known target genes, the functions of the predicted targets are related with iron ion binding or anaerobic metabolism that are the known functions of fnrS.
Only 14 known Hfq-binding ncRNAs are reported in GenBank. For the 193 FDR-controlled 7-bp ncRNAs, we applied network analyses to check their predicted targets. Using the network analysis and visualization software Mango (http://www.complex.iastate.edu/download/Mango), we first represent predicted ncRNAs and their target genes as collections of many small clusters in Figure 4.5. In each cluster, the center larger node represents an ncRNA candidate and the surrounding smaller nodes are its predicted targets. We intersected this network of ncRNA-target clusters with the known E. coli pathway networks obtained from KEGG database (http://www.genome.jp/kegg), but we found limited overlap among the target genes (i.e., the thicker lines connecting clusters). This is probably because the E. coli pathway information on KEGG is not complete. We then also applied a gene ontology network derived from E. coli Gene Ontology dataset (version 2.0; http://geneontology.org/gene-associations/gene_association.ecocyc.gz) to the ncRNA-target network. The gene ontology (GO) network connects E. coli genes that share at least one common GO term, thus genes that serve in similar molecular functions, biological processes or cellular components are connected in the GO network. Very interestingly, many of the predicted target genes for ncRNAs are linked by the GO network (i.e., the thinner lines among target genes within each cluster). For example, the candidate ncRNA332 at genome location 1339189_1339313_1 is surrounded by predicted targets entF, hscC, envZ and dnaX in Figure 4.5 (the circled cluster), and they all share GO terms 0000166 and 0005524 indicating that they are all nucleotide binding or ATP binding. Many similar target genes are connected this way even though the target genes for each ncRNA candidate is predicted independently. This suggests that many predicted ncRNAs may regulate genes in similar
activities at the molecular level or similar biological functions or the part of cellular components.

Figure 4.5. Network analyses of predicted ncRNA and their target genes. Each cluster represents an ncRNA candidate (larger center node) and its associated predicted target genes (smaller surrounding nodes). The few thicker lines that connect clusters are overlapped KEGG pathway links. The many thinner lines connecting targets within a cluster are overlapped gene ontology network links, suggesting targets predicted for some ncRNAs have similar biological functions.
Discussion

Trans-acting ncRNAs regulate target genes usually through imperfect base-pairing with their targets. These ncRNAs contain a short segment of 7 to 12 bp called the seed region that binds, perfectly or imperfectly, to the target mRNAs. The thermodynamic analysis software Picky allows such seed regions to be identified. However, the sizes of seed regions are short and variable, and their melting temperatures with targets vary greatly between perfectly and imperfectly matched targets. It is difficult to predict trans-acting ncRNAs solely based on the thermodynamic analysis. Therefore, we also have to consider known ncRNA structures to improve our predictions. We have chosen the Hfq-binding ncRNA structure with 3 well-defined regions, the poly-U tail, the Hfq-binding region and the seed region, to facilitate our prediction.

To assess the correctness of our predictions, three different analyses were performed. First we found that our predicted ncRNAs matched to the 14 known Hfq-binding ncRNAs reported in GenBank. Second, we measured the correlation between our predicted ncRNAs and their targets under 10 different E. coli treatments using a whole-genome tiling microarray and found many of them to be highly correlated. We also confirmed our results using 3 other non-coding RNA identification software and found them to be largely in agreement with our predictions. The same assessment procedures were applied to predict ncRNAs of A. tumefaciens. However, prediction results were not promising in A. tumefaciens C58 comparing to E. coli results. Hfq protein is also reported to play regulatory role combined with ncRNAs in A. tumefaciens [38], [39]. It is possible that Hfq-binding ncRNAs in A. tumefaciens do not share the three common properties.
These 3 other non-coding RNA identification software programs take different approaches to recognize ncRNA: GraPPLE considers secondary structure graphs, RNAcon considers sequence composition, and CPC considers six sequence features. Both RNAcon and CPC focus on sequence properties and that may explain why they agree a bit more with our predictions than GraPPLE, because our approach is based on sequence properties such as the poly-U tail in conjunction with thermodynamic analyses. All 3 ncRNA identification tools recognize broader ncRNA classes and are not specific to Hfq-binding ncRNA, thus their agreement to our predictions does not necessarily confirm that ours are Hfq-binding ncRNAs but just that they are likely ncRNAs. In addition, a network analysis was performed on the predicted target genes of *E. coli* showed that many target genes of each predicted ncRNA candidate in *E. coli* serve similar biological functions or molecular functions or are part of similar cellular components.

Our prediction method is distinctive — it not only predicts ncRNAs with just the genomic sequence and annotated target genes, but it also predicts ncRNA target genes. Actually, it is the input target genes that help screen out ncRNA candidates. Knowing both ncRNAs and targets may help researchers design more focused experiments to validate the predicted ncRNAs and their targets. Once sufficient predicted ncRNAs have been confirmed and their associations to target genes have also been validated, it is possible to recognize the common properties of working ncRNAs and to incorporate these properties into the design principles for artificial ncRNAs to target genes we choose. This may be useful in reverse genetics on bacteria using ncRNAs as tools.
Materials and Methods

Growth conditions for bacteria and microarray experiments

Details about how to grow bacterial cells of *E. coli* MG1655 and *A. tumefaciens* C58 are described in chapter 2-experiment protocol.

Thermodynamic analysis using PICKY

In both Steps 2 and 5, the software PICKY is used to discovery potential gene targets of the predicted ncRNA candidates. The whole *E. coli* and *A. tumefaciens* genome sequences were loaded twice each into PICKY as targets; one of the loaded instance was reverse-complemented to represent the reverse strand. The Examination function in PICKY was used to examine the predicted ncRNAs to determine if they match to any regions in the targets. The following parameters of the Examine function was used: maximum match length 20, minimum match length 7, minimum sequence similarity 66%, and minimum melting temperature difference 5°C; we screened forward strands only (because both strands were already represented), and used the better salt effect equation. All other PICKY parameters were taken at their default values. PICKY analysis results were saved with maximum reported nontargets set to 30 and maximum temperature difference set to 25°C.

RNA Secondary structure prediction

The ContextFold software was used for RNA secondary structure prediction in Step 4. ContextFold employs several scoring models that are trained to consider either more types of structural elements, or a larger sequence context [8]. To predict ncRNA structures, we chose the StHighCoHigh scoring model to set both the structural (St) and contextual
(Co) information high, because this model is shown to have the highest accuracy and sensitivity by ContextFold creators.

**Perl programs used in ncRNA prediction**

During the 5 ncRNA prediction steps, various Perl programs are used to process the data: `make_baits_polyAU.pl` is used to scan poly-U regions in both directions of the *E. coli* genome in Step 1; `analyze_polyU_prediction.pl` is used to screen initial PICKY thermodynamic analysis results in Step 2; `merge_trim_for_polyU.pl` is used in Step 3 to merge some ncRNA candidates; `make_seq.pl` is used extract RNA sequences from *E. coli* genome according to the predicted ncRNA positions in Step 4; `filter_w_structure.pl` is also used in Step 4 to analyze ContexFold output to further screen ncRNA candidates; and `refine_targets_singles.pl` is used to extract single-stranded stretches from predicted structures and `find_targets_polyU.pl` is used to produce the final ncRNAs in Step 5. All Perl programs used for this work and our predictions results are described in the Appendix A.

**References**


CHAPTER 5. GENERAL CONCLUSIONS

General Discussion

Bacterial small noncoding RNA (ncRNA) regulate gene expression. For example, some ncRNAs bind to house-keeping genes involved in transcription and protein synthesis [1]. Also, some ncRNAs are related to stress response [2]. Trans-acting ncRNAs are a class of bacterial small noncoding RNAs that form imperfect base-pairs with their target genes and are encoded separately from their target genes. Therefore, comparison of sequences alone is not enough to predict trans-acting ncRNAs in prokaryotes — one must use thermodynamics to improve the predictions. Many software tools using various approaches such as sequence compositions [3], sequence features [4], and machine learning on graph properties [5] were developed to identify ncRNAs. However, many of these tools use classification methods that can be impacted by the training data; the results of these prediction programs can be biased if their training data are small or biased. We have developed an ab initio ncRNA prediction method based on thermodynamic analysis and ncRNA structural features that does not require a priori training data sets.

In my early PhD study, we used only the thermodynamic approach to predict ncRNAs, considering only their melting temperatures with target regions. In this approach, we start ncRNA predictions with 100 bp fragments from annotated gene operon regions, screen these fragments against the whole genome by PICKY to detect closest trans-targets, and then rule out ncRNA candidates which do not meet a melting temperature threshold with their targets. However, results from this approach did not provide strong evidences to support our predictions after performing rigorous validation steps. Our initial predictions
did not match to already known database entries of ncRNAs in GenBank and BSRD [6].

Carefully designed quantitative real-time polymerase chain reaction (qPCR) experiments with over-expression plasmids containing some of our predicted ncRNA candidates also did not provide consistent results confirming that they were interacting with target genes. These results led us to devise a new approach to predicting trans-acting ncRNAs in prokaryotes.

Some trans-acting ncRNAs need Hfq proteins to bind to target genes [7]. Many Hfq binding ncRNAs have similar structural composition. Our new prediction method considers common structural properties of Hfq-binding ncRNAs: the seed region, Hfq-binding site, and poly-U tail. They should all be located on the single stranded stretches in these ncRNAs. Thermodynamic analysis using PICKY predicts target regions of ncRNA candidates and it helps discard ncRNA candidates that do not satisfy the basic sequence features. The predicted secondary structure of ncRNAs helps discard more ncRNAs candidates, if their predicted structural properties do not match known common structural properties of Hfq-binding ncRNAs. The structural properties of Hfq-binding ncRNAs allow us to produce more reliable prediction results than our first attempt of prediction, which is solely based on the melting temperature of ncRNAs to their targets.

High-density whole-genome tiling microarrays were design to validate the predicted ncRNAs. Recently microarrays are gradually being replaced by next-generation sequencing (NGS) techniques but they are still used in some other applications and have some advantages over NGS in prokaryotes, such as that microarrays require less RNA that is typically needed for NGS when a sequencing library is constructed [8]. NGS also did not perform well for the GC-rich CDSs with light normalization [9]. For NGS techniques to work
well on prokaryotic transcriptomes, some sort of rRNA removal must be performed, but these methods are often double-strand specific and can also remove non-coding regular RNAs which we are studying. Our design of tiling microarray covers all predicted ncRNAs and all annotated target genes. Therefore, our tiling microarray are useful especially for novel and non-model species that have not been annotated yet. In fact, tiling microarrays can help identify novel gene expression and facilitate the annotation effort of these species.

Even though, *Agrobacterium tumefaciens* C58 is known to be involved in Hfq protein to make some ncRNAs bind to target genes, prediction results were not promising compared to results of *Escherichia coli*. However, small numbers of predicted Hfq-binding ncRNAs after FDR control in *A. tumefaciens* show higher correlation coefficient values than in *E. coli*. In addition, most of predicted ncRNAs in *A. tumefaciens* are agree to 3 other identification tools. We may need to study if *A. tumefaciens* also shares the 3 common structural properties of Hfq-binding ncRNAs or what features make predicted ncRNAs fail to pass FDR controlling threshold in *A. tumefaciens* to improve prediction of ncRNAs.

Our predicted ncRNAs in *E. coli* MG1655 and *A. tumefaciens* C58 will help further study of functional and structural properties in bacterial noncoding RNAs. Our novel prediction method can easily apply to other Gram-negative bacteria since is is known that Hfq is usually required for the function and/or stability of this family of ncRNAs [10].

**Recommendations for Future Research**

The knowledge we will learn from naturally occurring ncRNAs may allow us to design artificial ncRNAs to knock out specific target genes. Although we cannot precisely
predict how likely this future work will be proceed at this moment, the following ideas may come in handy when the time is right to pursue the artificial ncRNA design effort. This future artificial ncRNA design is mainly based on predicted structures of confirmed ncRNAs and common properties of structures shared by the confirmed ncRNAs. To predict secondary structures of ncRNAs, we will use same software, called ContextFold [11], which is used for prediction of secondary structure of Hfq-binding ncRNAs. The kernel-based approach based on dual graph representation [12] or/and comparative RNA structure analysis such as LocARNA [13] may be used as machine learning methods to detect common structural properties among confirmed ncRNAs. To devise machine learning method to identify certain patterns, high quality training data are needed. Large enough sample size of training data set is important factor to build an unbiased model to recognize a certain pattern.

In addition, known RNA structural data can be downloaded from RNA related databases such as Rfam [14], fRNAdb [15], or BSRD [6] to test training data set. Including our novel ncRNA data set with existing ncRNA data has the benefit of reduced bias when applying the learning algorithms to detect common ncRNA structural properties. A separate but related factor is, the associations of ncRNAs to their target genes, which may also share some unique properties and our prediction method also helps to predict target genes of ncRNAs. Improvements to our prediction method will subsequently allow more precise artificial ncRNAs to be designed.

As we mentioned before, we are able to apply the novel prediction method to other species directly, if ncRNAs of the species regulate their target genes with Hfq protein. In addition, we can apply the novel prediction method to different classes of trans-acting
ncRNAs, which are not related in Hfq-binding, if we find some common shared features or structural properties with appropriate modification.

The overall goal of our project is to allow controlled gene regulation in prokaryotes to be carried out as easily as RNA interference is in eukaryotes. With the availability of modern NGS techniques and many novel prokaryotic genomes, this is the natural path to follow in biological research.

References


APPENDIX A. REPOSITORY OF SUPPLEMENTARY FILES

All related files of this study are added into my gitHub repository. The repository is available at https://github.com/cyprus91/Dissertation.git.

Newly assembled sequences of *E. coli* MG1655 and *Agrobacterium tumefaciens* C58 are named ‘MG1655_ngs.fasta’ and ‘c58_ngs.fasta’ at the repository, respectively. They are mentioned in Chapter 2 and Changer 4. W83.picky, as mentioned in Chapter 2, is stored at the repository with the related ‘W83.report’ file. All predicted results of *E. coli* and *Agrobacterium tumefaciens*, which are described in Chapter 4, are recorded in the file ‘predicted_ncRNA_target.xlsx’ at the repository. All Perl programmings described in Chapter 4 are located in ‘perl_programming’ folder at the repository.