IDENTIFICATION OF *TRANS*-ACTING SRNA TARGETS IN BACTERIA

by

JOYDEEP MITRA

(Under the Direction of SIDNEY R. KUSHNER)

ABSTRACT

Small regulatory non-coding RNAs (sRNAs) have emerged as an important class of regulators across all kingdoms of life. In prokaryotes, the majority of the known sRNAs bring about regulation by base pairing with their target mRNAs, resulting in either increased or decreased stability of the target transcripts. Based on their mode of action, these sRNAs are further sub-categorized into two categories: cis-acting and trans-acting. While cis-acting sRNAs are encoded on the antisense strand of their targets, trans-acting sRNAs bear no identifiable relationship with the loci of their targets. The lack of complementarity between trans-acting sRNAs and their target mRNA sequences; along with the added complexity that each sRNA can have multiple targets and some mRNAs are targets for multiple sRNAs, makes the discovery of such interactions a formidable challenge.

The research presented in this thesis describes a knowledge-based machine-learning model based on the popular random forest algorithm developed for the prediction of novel interactions in bacteria. The model was trained on a high quality dataset of experimentally verified sRNA-target interactions obtained from the literature. The prediction model is shown to be applicable on a genome-wide scale. The algorithm is further extended to filter predictions using random forest's intrinsic similarity measure. Finally, the selected predictions were

validated experimentally in *Escherichia coli* for several known *trans*-encoded sRNAs, leading to the identification of novel regulatory interactions.

INDEX WORDS: non-coding RNAs, regulatory RNAs, Gene Regulation in Bacteria, *Escherichia coli*, Machine Learning, Balanced Random Forest, Classification Algorithms

IDENTIFICATION OF *TRANS*-ACTING SRNA TARGETS IN BACTERIA

by

JOYDEEP MITRA

B.Sc, Bangalore University, 2004, India

M.Sc, University of Pune, 2006, India

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

IDENTIFICATION OF *TRANS*-ACTING SRNA TARGETS IN BACTERIA

by

JOYDEEP MITRA

Major Professor:	Sidney R. Kushner
Committee:	Jan Mrazek
	Liming Cai
	Russell Malmberg

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

DEDICATION

To my parents, who have taught me to strive on through the face of adversity.

ACKNOWLEDGEMENTS

What made me into a scientist today is not the exams, the paperwork and all the formalities that go with grad-school. When I look back, I will remember all the advice, constructive criticism and words of wisdom from those who shaped me into what I am today. I knew working in an experimental lab would be challenging as a bioinformatician. What I did not know is that the puzzles that come up in an experimental lab made me a better bioinformatician than what I had bargained for. Looking back, working with my professor Dr. Sidney Kushner is the best experience I could have got as a scientist. Words couldn't do justice to the gratitude I feel for the training he gave me. I thank my committee members for instilling some scientific sense in me. Dr. Liming Cai, Dr. Russell Malmberg and Dr. Jan Mrazek, you have been invaluable to my scientific development. If I ever do see further than the others before me, it is by standing on the shoulders of these giants.

I am also grateful to Dr. Jonathan Arnold, for his constant guidance and supervision throughout my tenure at the University of Georgia. I would like to thank my colleagues at the Kushner lab, who have helped me hone my skills in doing wet-lab experiments. Finally, I thank my loving wife, Arunima Singh, for being by my side and being my best and worst critique throughout. Without the encouragement and help from all my teachers, friends and family, I wouldn't be here today.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Non-coding RNAs have emerged as a major component of global regulatory networks in all kingdoms of life. Important roles have now been implicated for non-coding RNAs in almost all known pathways in various organisms, including several human diseases. As modern sequencing technologies enable the discovery of novel non-coding RNAs at an accelerated rate, annotation efforts are falling behind. This has created an urgent need for discovery and annotation of biological roles for the large majority of non-coding RNAs known. In bacteria, several types of non-coding RNAs are now known. A better understanding of the mechanistic details in prokaryotes is anticipated to broaden our current understanding of riboregulators in general.

This thesis presents a novel approach to facilitate the study of the largest class of bacterial non-coding RNAs. Specifically, the research is concerned with the challenging problem of computational identification of regulatory targets of these non-coding RNAs. The information is presented in the form of relevant literature reviews, manuscript communicating original research and concluding remarks as follows:

CHAPTER 2: REGULATORY NON-CODING RNAS IN BACTERIA

Chapter 2 provides an introduction to the various kinds of regulatory RNAs in bacteria. Special emphasis is laid on the base-pairing so called *trans*-acting non-coding RNAs, the central

theme of this thesis. Mechanistic details are discussed for *trans*-acting non-coding RNAs are elaborated upon.

CHAPTER 3: IDENTIFICATION OF REGULATORY RNA TARGETS IN BACTERIA

In chapter 3, a review is presented that elaborates the various techniques employed by investigators in the field to identify regulatory targets of non-coding RNAs. As with the majority of this thesis, primary emphasis is on the various experimental and computational techniques used for finding targets of *trans*-acting non-coding RNAs.

CHAPTER 4: SUPERVISED PREDICTION OF REGULATORY NON-CODING RNA TARGETS IN BACTERIA

Chapter 4 is an original research study describing a novel machine-learning approach for the prediction of *trans*-acting non-coding RNA targets. The classification algorithm developed here uses simple sequence information from the non-coding RNA and their mRNA target sequences. This reduces computational complexity significantly when compared to other state-of-the-art methods that rely on secondary structure prediction and multiple sequence alignments.

CHAPTER 5: KNOWLEDGE BASED IDENTIFICATION OF *TRANS*-ACTING REGULATORY SRNA TARGETS IN *ESCHERICHIA COLI*

Chapter 5 presents original research which is an extension of the predictive algorithm introduced in chapter 4. The machine learning classifier was applied on a genome-wide scale for the prediction of several non-coding RNA targets in *Escherichia coli*. The predictions were then validated experimentally, leading to the identification of several novel regulatory interactions.

CHAPTER 6: CONCLUSIONS

The thesis closes with concluding remarks about the research presented. Thoughts on future directions and improvements to the field are presented here.

CHAPTER 2

LITERATURE REVIEW: REGULATORY NON-CODING RNAS IN BACTERIA

Living organisms constantly have to adapt to changing environments. At the cellular level, these adaptations are realized in the form of a wide range of biomolecular interaction mechanisms. Up until recently, most of our understanding of these mechanisms revolved around the so-called central dogma i.e. genes encoded in the DNA are transcribed into messenger RNAs, which are then translated into functional proteins. Thus, proteins were thought to be the mediators of all or most biological functions in the cell, including gene regulation. Apart from the non-protein coding ribosomal RNA (rRNA) and transfer RNAs (tRNA), RNA molecules were thought to be mere intermediaries in the central dogma.

Despite findings of crucial roles of RNA in the fundamental processes of translation and splicing, the protein centric view has dominated the known regulatory circuitry of the cell. The realization of the prevalence of non-protein coding RNAs as the major class of functional biomolecules eluded detection even in the post-genomic era, since non-coding RNAs are not structured in the genomic sequences the way open-reading-frames are [1]. The paradigm shift in regulatory pathway research was brought about by the advent of high-throughput whole-transcriptome profiling technologies[2,3]. With the wide availability of whole-genome tiling microarrays and deep sequencing of the transcriptome (RNASeq), pervasive transcription was detected in many organisms [4]. These results meant that the majority fraction of the genomes that was previously thought to be “junk” DNA codes for RNA, and implies a far greater role for RNA as functional biomolecules than previously thought [4]. It is now thought that non-coding

RNAs far outnumber the protein-coding genes in genomes of higher organisms[4]. Several classes of functional non-coding RNAs (ncRNAs) have now been discovered, and it is now well established that the majority of them serve regulatory functions.

Although the majority of ncRNAs remain functionally unannotated, the regulatory roles for these molecules already span all levels of gene-expression, and affect almost all well understood pathways. The largest and best understood class of regulatory ncRNAs are ones that exert their effect by base-pairing interactions. Several aspects of the mechanisms of action of base-pairing ncRNAs are conserved across all kingdoms of life [5]. The work presented in this thesis involves bacterial members of this class of ncRNAs, hereby referred to as small regulatory RNAs (or simply sRNAs).

Regulatory RNAs in Bacteria

Examples of regulatory RNAs in bacteria were known long before their prevalence began to be appreciated. Shortly after the discovery of regulatory transcripts encoded in plasmids [6,7], the first chromosomally encoded sRNA was identified in *Escherichia coli* [8]. MicF was discovered serendipitously while screening a library of chromosomal fragments, and was subsequently found to inhibit the translation of outer membrane porin OmpF [8,9]. Like MicF, the first few sRNAs in bacteria were discovered rather fortuitously [10]. Computational searches for conserved non-coding regions flanked by orphan promoters and Rho-independent terminators led to the initial conjecture that many more non-coding RNAs possibly existed [11]. The confirmation of some of these candidates was initially aided by microarrays that probed intergenic regions, followed by direct detection by other methods [2]. Thereafter, advances in whole genome tiling-microarrays[12,13] and deep transcriptome sequencing (RNASeq) [14,15]

led to the identification of hundreds of novel non-coding transcripts across various bacterial species [16].

As with higher organisms, non-coding RNA regulators in bacteria are known to regulate their targets employing a variety of mechanisms [16]. Based on the type of targets and mode of action, several categories of these riboregulators have been identified in bacteria. Although the subsequent chapters in this thesis focus on just one major class of non-coding RNAs, a brief description of the other categories is included in this section for the sake of completeness.

Protein Binding Regulatory RNAs

Not many instances of protein binding ncRNAs are known in bacteria. However, the few that are known have far-reaching regulatory effects, since the protein targets are usually global regulators of metabolism [16] (**Figure 2.1** B). One of the better studied examples in this category is the regulation of carbon storage by the CsrA/CsrB system [17]. The RNA-binding protein CsrA globally regulates carbon storage/usage, biofilm formation and cell motility as cells transition into nutrient-deficient conditions. In nutrient rich conditions, CsrA binds to 5' UTR regions of mRNAs thereby affecting their stability. As nutrients deplete, the BarA-UvrB two-component system induces the transcription of the sRNAs CsrB and its homolog CsrC. These sRNAs bind with the CsrA protein to antagonize its activity until nutrients are available again. Each sRNA molecule is thought to be able to bind to up to 18 CsrA protein molecules, thus rapidly reducing the abundance of functional CsrA protein [17]. The CsrA/CsrB system is conserved across several species (e.g. RsmY/RsmZ system in several bacteria) where it regulates several key pathways [18].

Another example of a protein binding regulatory RNA with a similar mode of action is the *Escherichia coli* 6S RNA. The 6S non-coding RNA sequesters the RNA polymerase bound

to the σ^{70} sigma factor by mimicking σ^{70} promoter sequences [19]. Other examples of sRNAs interacting with proteins to mediate regulation exist, although the evidence is mostly circumstantial, and very little is known about the mechanisms. For instance, the protein YhbJ is known to interact with GlmZ (a non-coding RNA) to destabilize it by an unknown mechanism [16]. Another non-coding RNA, GlmY competes with GlmZ to bind to YhbJ. Under conditions where GlmY is abundant, GlmZ RNA is freed from YhbJ to carry out regulation of mRNAs [20].

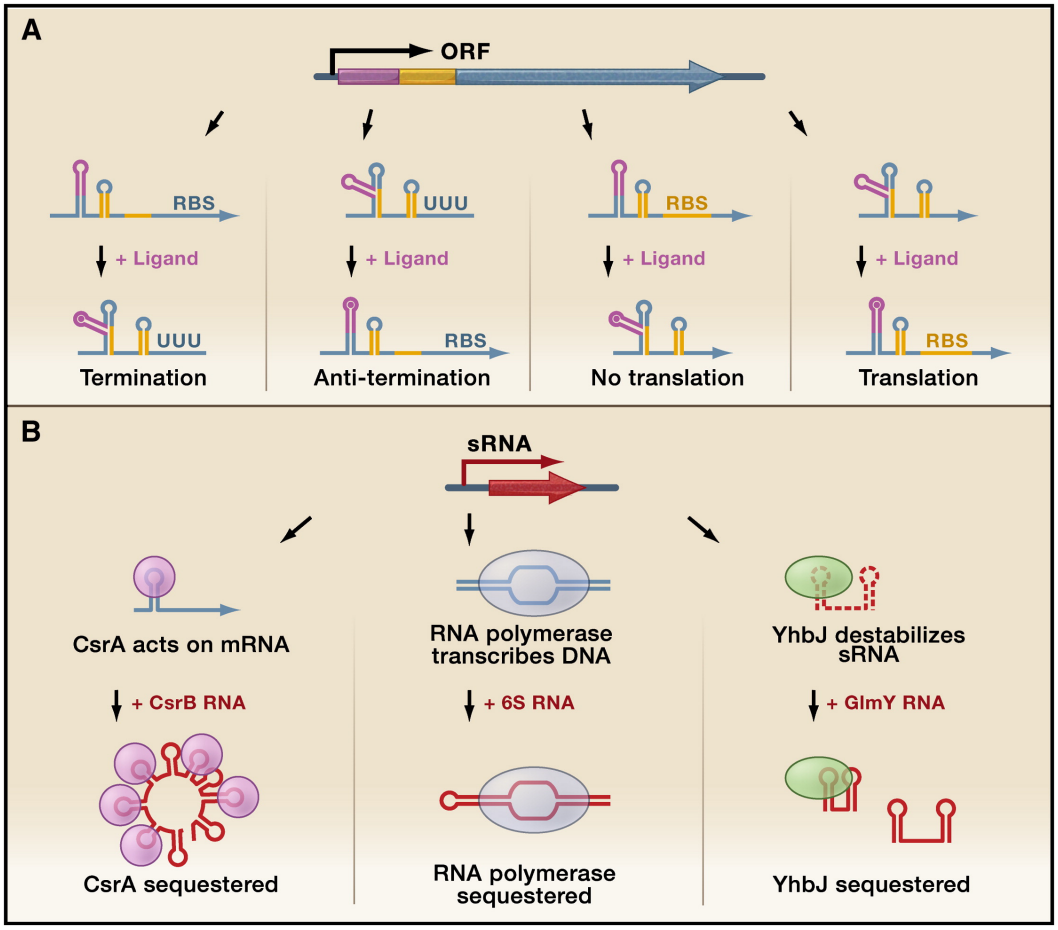


Figure 2.1: Illustrative depiction of non-RNA binding non-coding RNAs in bacteria. Depicted in (A) are the known mechanisms by which bacterial riboswitches act. The aptamer region (shown in pink) binds to the ligand and changes conformation, affecting the expression platform (shown in orange) to either form or disrupt terminator structures, or affecting the accessibility to the ribosome binding site (RBS). (B) Examples of non-coding RNAs that bind with multiple protein molecules through a repeating motif (CsrA/CsrB system) or sequester important regulators through molecular mimicry (6s RNA and GlmY/YhbJ). This figure was originally published in [16].

Riboswitches

Riboswitches are regulatory elements that are contained in the 5' regions of mRNAs that they regulate. They respond to changes in various environmental conditions and/or metabolite concentrations by adopting different conformations, affecting the expression of the mRNA [21,22]. The 5' UTR sequences can bind to small molecule metabolites and adopt different conformation in the presence of the metabolite. These structural changes result in regulation by switching between alternating RNA-hairpin structures that either allow the molecule to switch between terminator and anti-terminator elements (regulation of transcription), or disrupt ribosome binding sites within the transcript (regulation of translation) [23] (**Figure 2.1** A). The metabolite ligands are usually end products of the pathways that the corresponding riboswitch carrying genes are involved in, thereby forming a feedback loop [24].

Most of the known riboswitches have been primarily studied in the Gram-positive bacteria *Bacillus subtilis* [23]. These riboswitches generally regulate key junctures of metabolic pathways and are known to involve a wide variety of reaction intermediates acting as ligands such as flavinmononucleotide (FMN), guanine, and lysine [22,23].

Cis-Acting Regulatory RNAs

Most of the known regulatory RNAs known in bacteria act upon their mRNA targets through base pairing. Base-pairing sRNAs are further subclassified into two categories based on the extent of base pairing and mode of action. *Cis*-encoded or *cis*-acting regulatory RNAs are so-called because they are encoded in "*cis*", or at the same genomic locus as their targets, on the opposite strand (**Figure 2.2** A). As a result of this complementarity, *cis*-acting sRNAs can form extended stretches of contiguous base pairing with their targets, resulting in post-transcriptional inhibition of the target gene function. Although several putative endonucleases and other proteins have been purported to play a role in these mechanisms, the most commonly accepted

hypothesis involves degradation of the complex by ribonuclease III (RNase III). The long double-stranded regions formed from the base pairing allows RNase III, an endoribonuclease which specifically acts on double stranded RNA, to cleave the duplex [25]. An additional effect of base pairing by some sRNAs results in the occlusion of the ribosome binding site, thereby inhibiting translation [25].

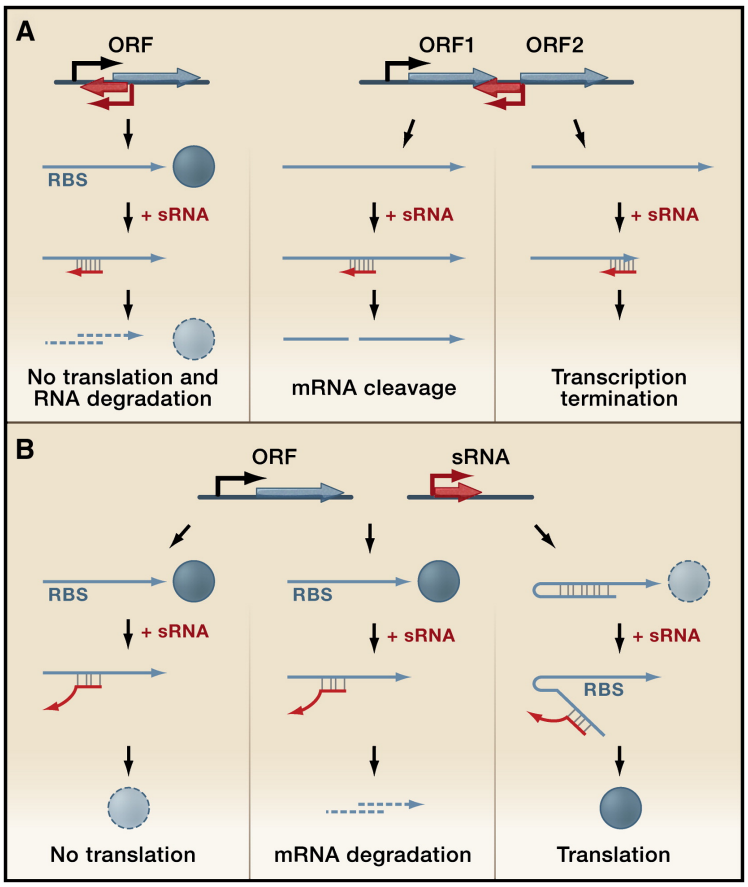


Figure 2.2: Gene arrangements in base-pairing regulatory non-coding RNAs. (A) *cis*-acting sRNAs occur on the antisense strand of their targets and share extensive complementarity with them. Regulatory effects are brought about by regions of perfect base pairing. (B) *trans*-acting sRNAs occur at different loci from their targets, and can have a variety of effects depending on the region of the target transcript they bind to. This figure was originally published in [16].

Cis-acting sRNAs have been found to be encoded in both chromosomes and plasmids [25]. The sRNAs and their target mRNAs are transcribed using independent promoters in either

direction. Several plasmid encoded *cis*-acting sRNAs regulate fundamental biological processes, such as the RNAI/RNAII system of ColE1 for replication control [26], and *FinP/traJ* for conjugation control [27]. Chromosomal *cis*-encoded sRNAs, on the other hand, are usually expressed only under specific physiological conditions, in many cases as part of a toxin-antitoxin system. For example, GadY (antisense to *gadX* gene in *E. coli*) shows increased abundance in the stationary phase [28], while IstR (antisense to *isiA* in *Synechocystis sp.*) is abundant during iron stress [29].

Trans-Acting Regulatory RNAs

The most prevalent class of regulatory RNAs that also act by base-pairing with their mRNA targets are the so called *trans*-encoded or *trans*-acting regulatory RNAs. Unlike the aforementioned *cis*-acting counterparts, these sRNAs are encoded "in *trans*" or at a different location from their target mRNAs (**Figure 2.2** B). In fact, each *trans*-acting sRNA can target multiple mRNA targets occurring at unrelated loci [5]. As a consequence of this mechanism, *trans*-acting sRNAs lack extensive complementarity with their mRNA targets and their interactions are established with short stretches of weak base pairing interspersed by unpaired regions. The interactions often involve non-canonical base pairs and are made possible by the involvement of the RNA chaperone Hfq [5,16].

Trans-acting sRNAs exert regulatory effects on their targets similar to their *cis*-acting counterparts, with most known interactions leading to post-transcriptional down-regulation of the target genes, although a few cases of up-regulation have also been reported. Most *trans*-acting sRNAs show increased transcript abundance under conditions of physiological stress and regulate multiple genes in well-defined regulons pertaining to those conditions [30].

DsrA was one of the first sRNAs that was found to interact with multiple mRNAs having both down-regulatory and up-regulatory effects [31]. When constitutively expressed on a multicopy plasmid, DsrA represses the H-NS protein, a global regulator of capsule genes and polysaccharide production, thus facilitating capsule formation [32]. Incidentally, DsrA was also found to upregulate the stationary phase sigma factor, RpoS, in contrast to the down-regulatory effects of all its other interactions [31,33,34]. Another well-studied sRNA RyhB, was found to be regulated by the ferric uptake repressor transcription factor Fur [35]. In turn RyhB regulates several genes responsible for iron storage homeostasis[36]. Several sRNAs have been implicated in the regulation of the bacterial outer membrane, quorum sensing and related pathways[1,37,38].

It is these widespread effects on a variety of biological pathways and many aspects of *trans*-acting sRNAs that make them an interesting and challenging area of research, which is the main theme of this thesis. Henceforth, unless otherwise specified, the term sRNA will refer to *trans*-acting sRNAs.

Mechanistic Aspects of Regulation

The *trans*-acting sRNAs act stoichiometrically to bring about regulation, with the catalytic component usually being provided by ribonucleases [16]. The underlying mechanisms at the molecular level that lead to regulation by *trans*-acting sRNAs are still an area of active research. This section outlines the working hypotheses based on our current knowledge about these mechanisms.

Hfq facilitates base-pairing interactions

Hfq was first identified as a host factor required for RNA phage Qβ replication (hence the nomenclature) [39]. Since then, our understanding of the role of Hfq in the RNA world has evolved with multiple studies suggesting its involvement in several pathways as a global RNA chaperone [40,41].

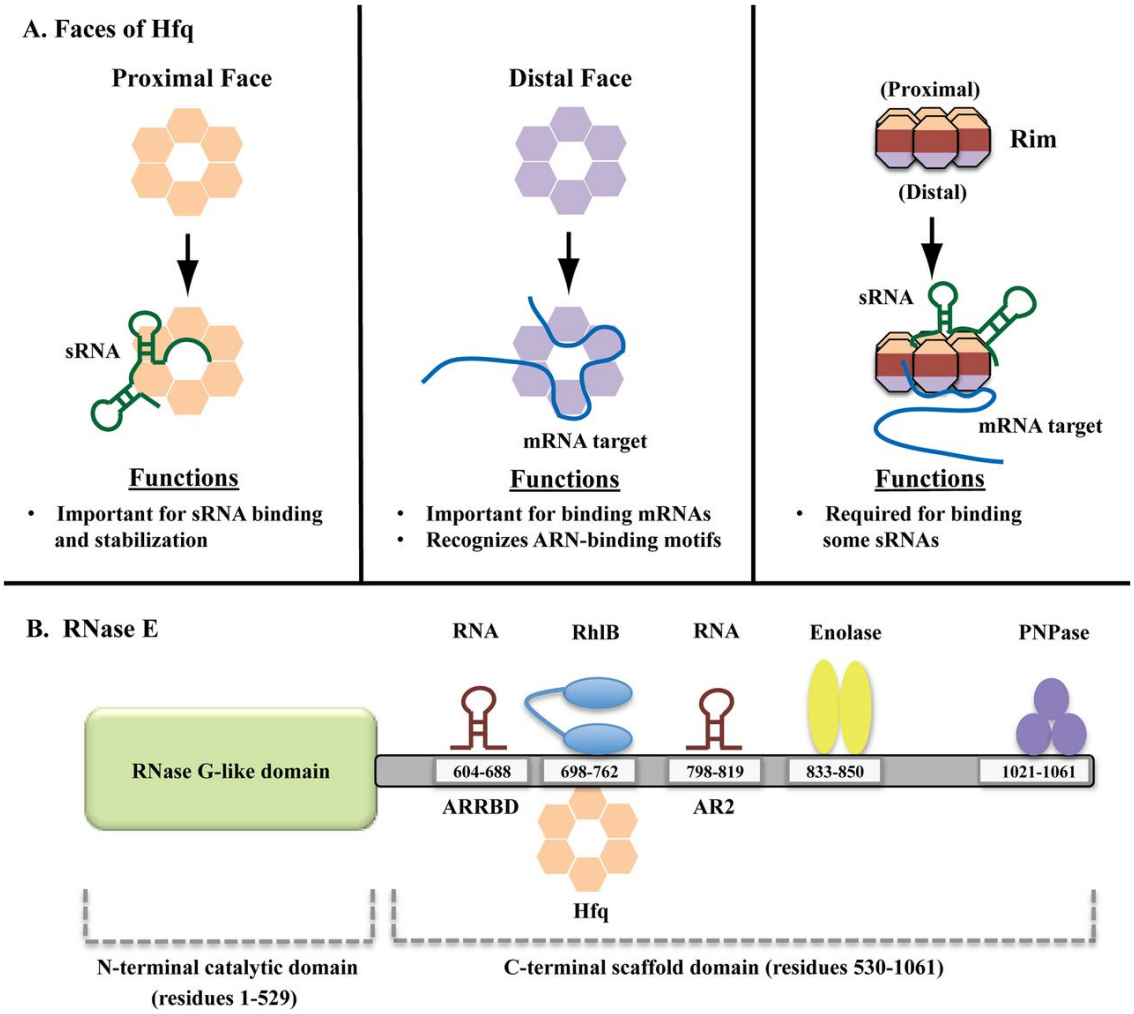


Figure 2.3: (A) Proximal face of Hfq (orange) binds to the sRNA and the distal face (purple) binds to the mRNA. The rim (right panel, in red) is thought to facilitate base pairing in a step wise manner. (B) The many domains of the RNase E protein that bind to other members of the degradosome. Interactions between Hfq and RNase E accelerates turnover of sRNA-mRNA duplexes. This figure was originally published in [42]

In the cell, Hfq exists as a hexameric structure with multiple interaction sites for RNAs and proteins [43]. Site-directed mutagenesis studies suggest that the so-called "proximal face" of the hexameric ring preferentially binds to sRNAs [44], while the "distal face" of the structure shows a strong affinity for mRNAs [44,45] (**Figure 2.3**). The preference for binding A/U rich regions was predicted based on the crystallographic data [43] and genomic SELEX experiments [46]. The distal face has a strong preference for repeating ARN motif (where R is a purine and N is any nucleotide) occurring in mRNAs [43].

Although the presence of multiple Hfq binding sites within RNAs complements the evidence that Hfq facilitates base pairing between sRNAs and mRNAs [47–49], the detailed mechanisms of this process remain obscure. A third region consisting of positively charged amino acids, termed the "lateral surface" or "rim" of the Hfq complex has been implicated in playing an important role in the steps leading to sRNA-mRNA duplex formation [44,50] (**Figure 2.3**).

Role of the degradosome and working hypotheses

Hfq physically interacts with ribonuclease E (RNase E), the primary ribonuclease in mRNA degradation pathways. RNase E is part of a multi-protein complex known as the degradosome, which also contains of polynucleotide phosphorylase (PNPase, a 3'->5' exonuclease), RNA helicase B DEAD-box motif (RhlB) and the glycolytic enzyme enolase [51–53]. It has been hypothesized that the interaction of Hfq with RNase E presents the bound sRNA-mRNA duplex to the degradosome, upon which the duplex undergoes rapid degradation [42] (**Figure 2.4**). In addition to the endonucleolytic activity of RNase E, PNPase acts as an exonuclease to further degrade the RNAs [54]. The RhlB helicase helps to disrupt secondary structures in the RNA molecules to allow PNPase activity to process those regions [54].

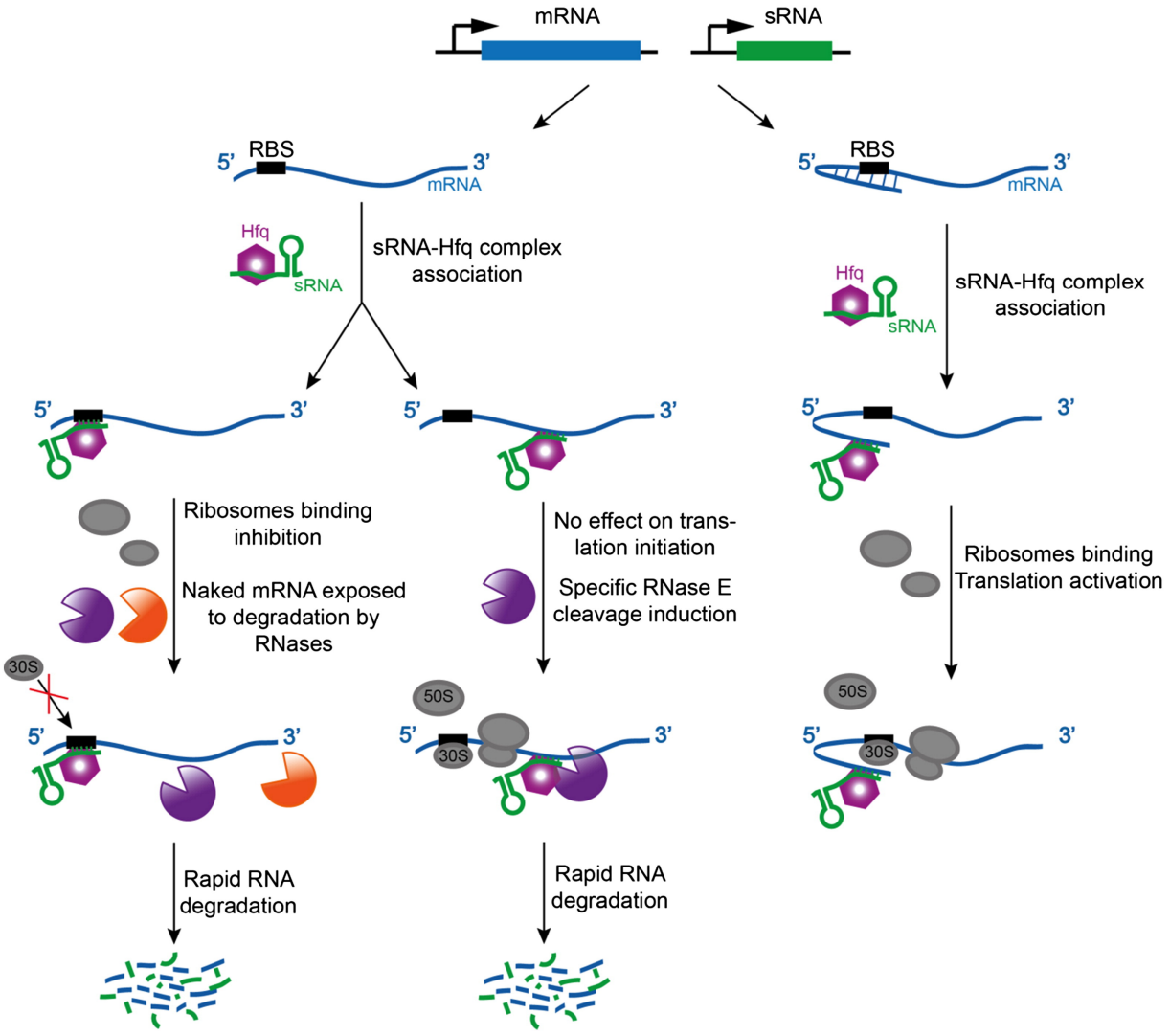


Figure 2.4: Multiple pathways to regulation by *trans*-acting sRNAs. Negative regulation is brought about by exposing cleavage sites to RNase E and other ribonucleases (left and center panel). In positive regulation (right panel), the sRNA binds upstream of the RBS to disrupt secondary structures that prevent ribosome binding. This figure was originally published in [55].

A second mechanism by which sRNAs mediate downregulation is by inhibition of translation (**Figure 2.4**). In bacteria, transcription and translation are tightly coupled [56,57], and ribosomes bind to transcripts as they are transcribed by RNA polymerase [56–58]. In many cases, the regions of interaction in the duplex coincide with the ribosome binding sites (RBS) on the mRNA, impeding loading of ribosomes required for translation. As a result, the mRNA, now

free from bound ribosomal proteins, is more exposed to endonucleases, resulting in loss of stability [59]. However, there is some experimental evidence that suggests that inhibition of translation alone is insufficient for the degradation of the mRNA and the pairing of sRNA to the mRNA is a requirement for the recruitment of RNase E and other ribonucleases [38].

Finally, a third mechanism of regulation involves upregulation of targets by translation activation. Unlike the more frequently found downregulatory interactions, fewer examples are known of interactions that lead to upregulation [60]. The working hypothesis described for these interactions suggests that the target mRNAs have existing secondary structures upstream of the transcription initiation site, around the RBS (**Figure 2.4**). The sRNA binds to a region upstream of the RBS, disrupting the existing secondary structure and thus making the RBS available for ribosome loading and protein synthesis [55,60]. As a side effect, the mRNA targets covered with ribosomes are no longer exposed to the endonucleases and have increased stability [60].

Other factors

Although RNase E is the primary enzyme involved in the turnover of sRNAs and their targets, other ribonucleases have also been found to be involved. A previous study on the genome-wide effects of RNase E and RNase III using tiling-microarrays revealed that both endonucleases are responsible for the processing of a number of sRNAs [13]. The role of RNase III in *cis*-acting sRNA regulation has been studied, but little is known about the mechanisms by which they act on the short regions of base-pairing in *trans*-acting sRNAs. The exonuclease PNPase is purported to play a role in the sRNA mediated degradation of certain categories of genes [61], and those sRNAs that do not associate with Hfq [62]. Another recently discovered endoribonuclease, YbeY, has been reported to modulate regulation by sRNA on several genes, in response to hydroxyurea stress [63].

RelA, a protein thought to be a central regulator of the stringent response, has also been reported to play a role in several Hfq mediated sRNA-mRNA interactions [64]. RelA apparently aids the oligomerization of the Hfq protein, and stimulates its binding efficiency to sRNAs in the process [64].

Yet another study determined that the triphosphate at the 5' end of a sRNA paired with its target is processed to a monophosphate to make the duplex more susceptible to RNase E [65]. The enzyme known to be responsible for the initiation of mRNA decay by processing the 5' triphosphate is RNA pyrophosphohydrolase (RppH) [66]. However, no change in this activity was observed in a RppH mutant in *S. enterica*, suggesting the presence of a second enzyme of this nature [66].

These studies show that there are several aspects of regulation by *trans*-acting sRNAs that further study. A deeper understanding of the mechanisms of sRNA gene regulation will be greatly facilitated with the discovery of more interactions. The following chapters of this thesis discuss the challenges and methods employed for the identification of novel interactions of *trans*-acting sRNAs with their targets.

CHAPTER 3

LITERATURE REVIEW: IDENTIFICATION OF REGULATORY RNA TARGETS IN BACTERIA

As was described in the previous chapter, sRNAs exert regulatory effects on genes involved in a wide variety of biological pathways. As high-throughput sequencing technologies continue to get cheaper and more accessible to researchers, several studies are being directed to sequence the transcriptome of organisms under different physiological conditions to discover novel non-coding RNA transcripts [4]. Although these studies continue to uncover novel transcripts in a wide variety of genomes, annotation efforts have struggled to keep up. There has been some ambiguity on whether these small RNAs are a result of pervasive transcription or biologically functional [4]. In fact, the number of known non-coding transcripts in many organisms already outnumber protein coding genes, with the biological functions unknown for most of them.

The widening gap in non-coding RNA discovery and annotation is particularly conspicuous in prokaryotes [38,67]. Going by the current trend, most intergenic non-coding RNAs in bacteria are likely to be *trans*-encoding sRNAs [68]. As previously mentioned, their counterparts in base pairing regulatory RNAs, *cis*-acting sRNAs, target mRNAs on the complementary strand. This makes identification of *cis*-acting RNAs relatively straightforward. In contrast, finding the potential mRNA targets of the growing number of *trans*-acting sRNAs is a non-trivial task. Firstly, the lack of extensive base pairing and correlation of genomic loci with their targets makes it difficult to identify targets. Furthermore, the short regions of base pairing

interactions are often comprised of weaker non-canonical base pairs, usually made possible by the RNA chaperone Hfq using mechanisms still not completely understood [60,69]. Just as each sRNA may target multiple mRNA targets, the target mRNAs may be regulated by multiple sRNAs [55,59]. Several experimental and computational tools have been developed to uncover this increasingly complex regulatory network of sRNAs and their targets. This chapter elaborates on the most successful and commonly used techniques in this area.

Experimental Approaches

The first target for a sRNA to be identified in bacteria was the mRNA for an outer membrane porin, *ompF* [8]. This finding was a result of the serendipitous discovery of the sRNA regulator itself, MicF, in a screen for genomic fragments that inhibited the OmpF protein [8]. Further characterization of this interaction suggested a 20 nucleotide region of imperfect base-pairing between MicF and *ompF* [70]. As more sRNAs were subsequently discovered [71,72], several experimental approaches were designed to identify their targets. In this section, an overview of the successful methodologies employed for this purpose is presented.

Classical genetic approaches

Early approaches designed to look for sRNA targets used genetically engineered bacterial strains to select for strains expressing the target gene. This was achieved by random insertions of μ phages carrying chromosomal fragments, with truncated *lacZ* genes [73]. When an sRNA is overexpressed in a library of cells, cells that carry the genes inhibited by the sRNA will show up as white colonies on X-gal indicator plates or as blue colonies when the sRNA is not expressed. Targeted genes were then identified by cloning these selected colonies. This approach was used to identify the effect of the sRNA OxyS on *fhlA* mRNA [74]. However, the classical genetic approach is labor and time consuming, and unsuitable for widespread detection of sRNA targets.

Biochemical approaches

At least two studies have reported to have successfully captured molecular species interacting with sRNAs using biochemical fishing techniques [75,76]. The first of these studies utilized the strong interaction between the sRNA RydC and the RNA chaperone Hfq. His-tagged Hfq molecules bound to RydC were allowed to incubate in total cellular RNA *in vitro*. The His-tagged complexes were then extracted with bound mRNA fragments, which could then be converted to cDNAs for analysis. RydC was found to regulate the *yejABEF* operon, encoding a predicted ABC permease [75].

Another study screened for genes targeted by the sRNA RseX, which would overcome the lethality associated with the deletion of the essential gene *rseP*, a global regulator of the extracytoplasmic stress response [77]. Since RseX is transcribed from a σ^E (envelope stress sigma factor) promoter in *E. coli*, it was predicted to regulate genes associated with this response. Biotinylated RseX bound to streptavidin magnetic beads was incubated with total RNA (**Figure 3.1** a). The bound mRNA fragments were then converted to in cDNA and hybridized to whole-genome microarrays. Microarray analysis revealed two targets for RseX, genes for the outer membrane proteins *ompA* and *ompC* [77]. However, capturing the molecular duplex biochemically relies on strong interactions, which are rarely seen with *trans*-acting sRNAs.

High throughput transcriptome screening

Recent improvements in transcriptome profiling technologies now enable researchers to survey affected transcripts across the entire genome. As mentioned in the previous section, microarrays have been successfully used to screen for putative targets that have been biochemically enriched [77]. Other studies have used whole-genome microarrays to look for affected mRNAs by comparing strains expressing low levels of the sRNA (or with the sRNA

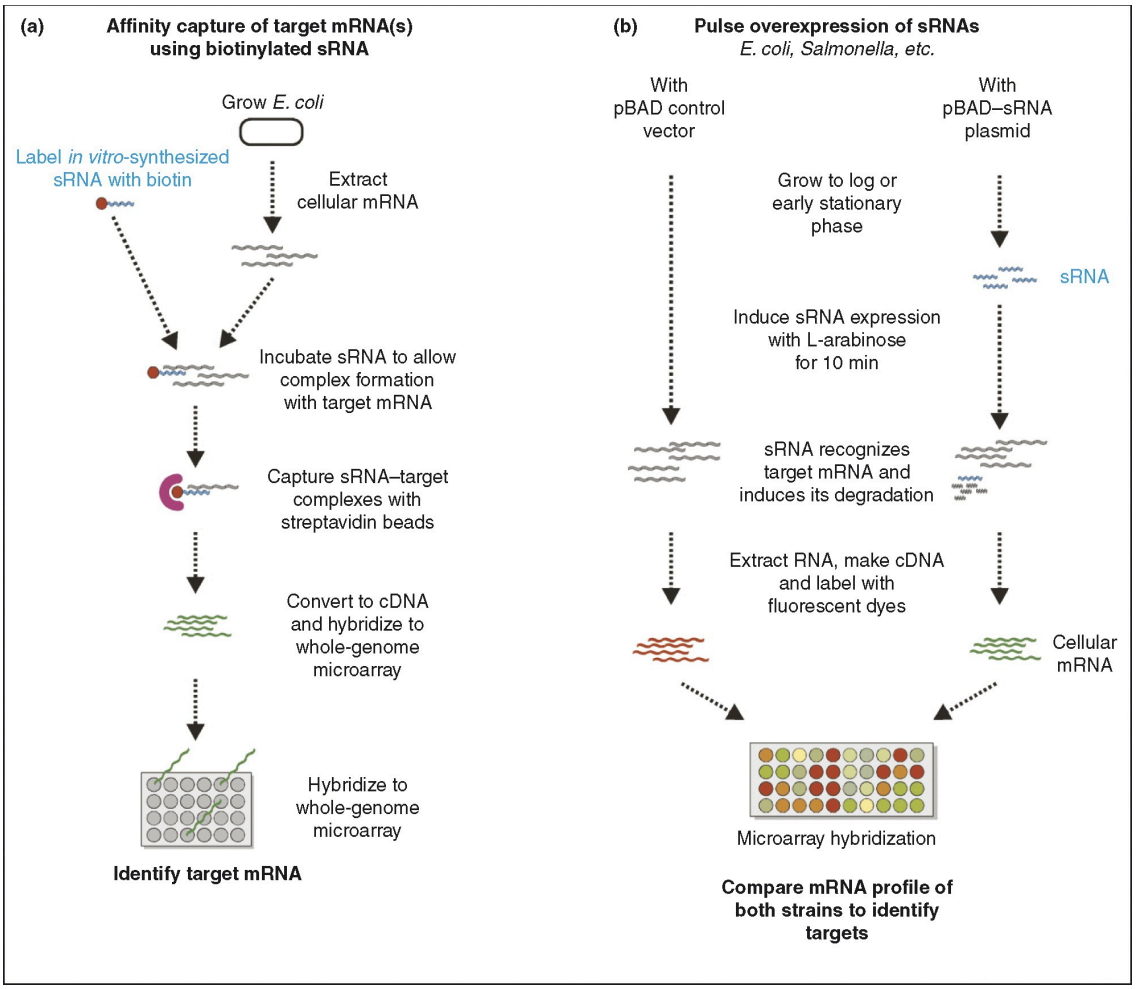


Figure 3.1: Identification of *trans*-acting sRNA targets using microarrays. Two commonly used strategies are used for this. (a) Biochemical capture of sRNA-target complexes bound to streptavidin beads are enriched and analyzed with microarrays. (b) Induced expression of the sRNA is followed by total RNA extraction. The sample is hybridized to a microarray along with the control to detect mRNAs with significantly changed abundance levels.

gene deleted) with strains expressing high levels of the sRNA (constitutively expressed from a plasmid) (**Figure 3.1** b). This strategy was used for the sRNA DsrA [78], suggesting many additional targets to the two that were already known. The issue with this method is that overexpressing the sRNA affects a large number of genes that are indirectly regulated by the true targets [68]. Subsequent validation of several true positives did reveal that DsrA plays a role in acid resistance [78].

A workaround to this problem has been sought by extracting the RNA for the microarray immediately following short-term expression of the sRNA from an inducible promoter [79]. The reasoning here is that a pulse induction of a tightly controlled inducible promoter (such as the arabinose induced pBAD promoter) [79] will affect the direct targets early on, minimizing the changes in indirectly affected genes when the total RNA is extracted. This strategy has been used for several sRNAs in *E. coli* and *S. enterica*. The sRNAs RybB and MicA, which are transcribed from σ^E (envelope stress sigma factor) promoters, were found to regulate several outer membrane proteins in *S. enterica* [80]. A few of these targets were subsequently also observed to be regulated in *E. coli* [81].

Computational Approaches

The experimental methods outlined in the previous section are labor intensive and time consuming. While high throughput methods have shown some promise, they require performing several replicates of high-cost experiments to minimize false positives. These issues can be greatly reduced by supporting the search for sRNA targets with computational methods. Several computational approaches have been developed, based on our current understanding of the regulatory mechanisms from the known interactions.

Methods based on Sequence Complementarity

The simplest approach that has been applied successfully to finding miRNA targets in eukaryotes [82,83] is to search for complementary regions in the mRNA and sRNA sequences. However, unlike miRNAs, bacterial *trans*-acting sRNAs do not have well defined "seed" regions of interaction. This makes the use of pure sequence search methods like BLAST [84] ineffective. A BLAST-like search method had been developed to account for non-Watson Crick base pairing in RNA called GUUGle [85]. Other pure sequence based approaches are the individual base-pair

model used by TargetRNA [86] where Watson Crick base pairs are scored uniformly, or a similar approach where GC pairs are given a higher score than AU pairs [87]. Although these approaches fall short when used for making target predictions, their simplicity allows the computation to be fast and they have been used to score the significance of matches in more advanced methods [88].

Energetic Scoring of Duplexed Regions

Somewhat more sophisticated approaches to searching for sequence complementary regions score the paired regions using scoring schemes used for evaluation of secondary structures. The energy parameters for this scoring scheme represent free energies (in Kcal/mol) that were derived from experimental data [89]. The scoring of a base pair in the interacting region depends on the immediately neighboring base pairs. This makes these approaches much more realistic than simplistic independent base-pair scoring, since such scoring schemes account for stacked base pairs and internal loops and bulges. Several algorithms incorporate this approach, notably RNAduplex [90] and RNAhybrid [91] from the Vienna RNA package, and the extended version of TargetRNA [86]. The advantage of using these simple energy models is that it is comparable in computational speed to the simple scoring method. However, since these methods ignore intra-molecular base pairs, these algorithms may end up predicting interactions at regions that are already involved in intramolecular secondary structures.

Secondary Structure Prediction of Concatenated Sequences

One of the main shortcomings of the previously described methods is their inability to account for intra-molecular secondary structural elements in the individual RNA molecules. Methods based on secondary structure prediction approaches were developed to account for these issues. The first category of such approaches aimed to predict the joint structure of the

sRNA and target RNA molecules by providing the concatenated sequence of the two molecules as input to a secondary structure prediction algorithm.

RNACofold [92] is based on this approach and accepts the input sequences concatenated and separated by a linker symbol. It applies a modified version of the RNAfold [93] algorithm, where the linker region is treated as a special bulge structure. As a result of this restriction, the modified algorithm only predicts secondary structures nested in the sequence of the two concatenated sequences [94].

The obvious advantage of concatenation based structure prediction is that the algorithms used for predicting secondary structure of a single RNA is easily extended for the joint structure prediction problem. This allows the computation of secondary structures of the individual RNA molecules (intramolecular base pairing) as well as the joint structure (inter-molecular base pairing) and thus significance statistics of the interactions predicted.

Accessibility based structural approaches

Although the concatenated secondary structure prediction methods overcome most of the shortcomings of the previous methods, they cannot predict non-nested joint structural elements like pseudoknots and kissing hairpins. Thus, a second category of secondary structure prediction based algorithms was developed to address this issue. As opposed to attempting to predicting a single joint secondary structure, the secondary structures of the individual sequences are taken into account first, in order to account for the accessible regions in each RNA molecule. Essentially, a region within a single RNA molecule must be free of intramolecular base-pairing in order to form an interactive base pair with another RNA molecule. This requires the calculation of the energy to free the regions in each RNA molecule of intramolecular base

pairing, followed by the energy required to make the interaction base pairs across an ensemble of structures.

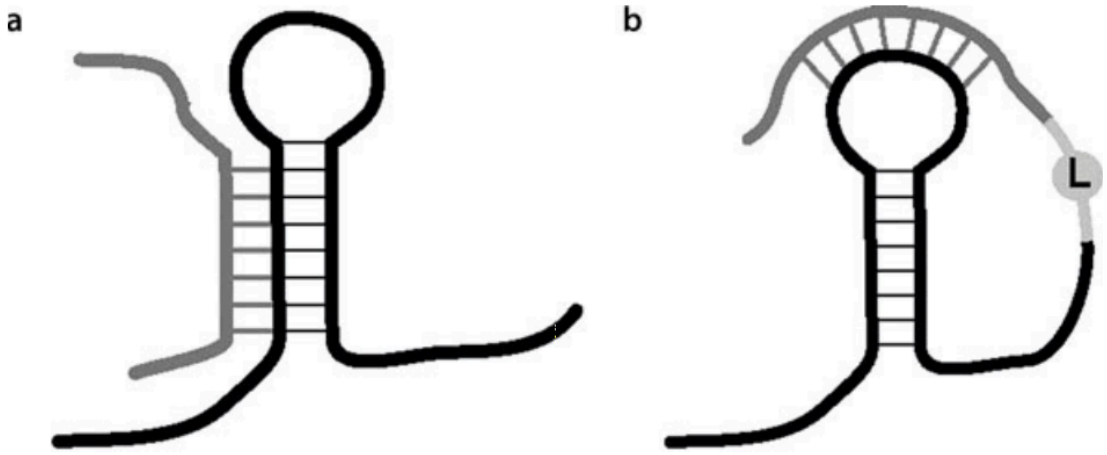


Figure 3.2: Common erroneous hybrid structures predicted by secondary structure based algorithms. (a) A biologically impossible structure that might get predicted by algorithms that use naïve scoring of duplexes using energy functions. (b) A commonly found non-nested structure that the concatenation based secondary structure methods will fail to predict.

Although this is computationally a lot more expensive than the previous methods, these methods can predict the non-nested pseudoknots and kissing hairpins that the concatenation based approaches would have missed. Because of the computational overload involved here, the algorithms that adopt this approach, RNAup [95] and IntaRNA [96] utilize precomputed energy values for all possible interaction regions.

Comparative Approaches

As an extension of the sequence based approaches that simply searched for complementary regions in the interacting RNA molecules, comparative approaches assume that the regions in the sequence involved in these interactions are evolutionarily conserved. The first published method that used evolutionary information for the prediction of sRNA targets was

PETcofold [97]. PETcofold uses multiple sequence alignments for both the sRNA and target mRNAs from multiple species. The multiple sequence alignments allow the incorporation of covariance information that result from compensatory base-pair mutations that preserve functional structural elements. This makes it more likely to find regions that are important for the interactions within the sequence. However, PETcofold's approach used positionally fixed alignments across multiple species, limiting the prediction of the sequence regions that are highly conserved[98]. A newer strategy used in CopraRNA [99] works around this limitation by allowing the sRNA and target mRNA interaction sites and patterns to be flexible. The driving hypothesis here is that while the target regulation should be conserved across related species, the base pairing patterns may be different. Thus, predictions are made independently in each species, before the evidence is combined to determine significance[99]. Although these methods require both the sRNA and the target mRNA sequences to be conserved in multiple species, they have successfully predicted targets for a few highly conserved sRNAs.

Target Validation

Irrespective of the methods employed for identification or prediction of sRNA targets, the regulatory interactions need to be validated individually using *in vivo* assays. For *trans*-acting RNAs, most known regulatory interactions result in change in stability of their targets [100]. Thus, the most common technique employed to test for regulatory effect is to assay the transcript abundances of the predicted targets in strains with the sRNA gene deleted, and/or with the sRNA being overexpressed from an inducible plasmid. As previously mentioned, overexpressing (or pulse inducing the expression) of sRNAs have been found to produce a large number of secondary effects. Therefore, individual assays of targets using this method is usually accompanied by assaying the same targets in sRNA deletion strains, comparing the mRNA levels

against appropriate wild type controls [68]. Since most regulatory interactions result in moderate changes in transcript abundance [30,50], these quantifications need to be carried out by sensitive experimental techniques like northern blot analysis or quantitative PCR.

Other validation methods have been reported to use translational fusion readouts of the putative target gene to a reporter gene. Commonly used reporter genes include the lacZ gene, encoding for β-galactosidase and the green fluorescent protein (GFP). While reporter systems are well established, if the fusion is driven by the target gene promoter, independent experiments need to be done in order to validate the effects on transcription [68]. On the other hand, inducing transcription of the fusion using an inducible promoter does not allow the identification of specific biological condition under which the regulation occurs.

Most of the confirmed regulatory interactions have been discovered in *E. coli* and *S. enterica* and very little is known about the sRNAs in other organisms [3]. More robust techniques are required to accelerate the discovery process across more species. Discovery of more regulatory interactions will significantly help in the understanding of the underlying mechanisms of riboregulation.

CHAPTER 4

SUPERVISED PREDICTION OF REGULATORY NON-CODING RNA TARGETS IN BACTERIA USING ALIGNMENT-INDEPENDENT SEQUENCE INFORMATION

Mitra. J, Kushner. S.R; Submitted to RNA Journal, 10/20/2015

Abstract

Small non-coding RNAs (sRNAs) are ubiquitous regulators of gene-expression across all kingdoms of life. Regulation by sRNAs in bacteria allows them to rapidly adapt to changing environmental and growth conditions. The vast majority of sRNAs in bacteria post-transcriptionally regulate levels of target mRNAs through molecular interactions. In contrast to *cis*-encoded sRNAs that are transcribed from the antisense strand of their targets, *trans*-encoded sRNAs regulate multiple mRNAs irrespective of their locations, interacting with them in unconventional ways that make these interactions difficult to predict. Computational methods for the prediction of sRNA targets have primarily focused on base-pairing interactions. This approach has only been modestly successful.

We have used Balanced Random Forests for the prediction of *trans*-encoded sRNA targets. The algorithm extends Random Forest's sampling strategy for achieving equal performance in terms of sensitivity and specificity. Numerical features used for the classification are calculated from the sequences, allowing the predictions to be made using sequence information in an alignment-independent manner. Our algorithm outperforms current methods in terms of classification performance, and can be applied for the prediction of targets for sRNAs across entire prokaryotic genomes. The source code and data are available at https://github.com/j-mitra/BRF-sRNA-target

Introduction

Bacteria adapt to changing environments and various stress conditions through intricate genetic regulatory pathways. These pathways involve diverse mechanisms at multiple levels of gene expression, enabling the cell to rapidly adjust its physiology. In recent years, non-protein coding small RNAs (sRNAs) have emerged as major post-transcriptional regulators of gene

expression in almost all known bacterial species, mediating control through molecular interactions with target mRNAs [101] or proteins [102]. These sRNAs control expression of genes involved in a wide range of pathways, including regulation of stress responses, carbon and iron metabolism, biofilm formation, cell motility and quorum sensing [100].

The more prevalent mRNA-pairing sRNAs post-transcriptionally regulate levels of target mRNAs through molecular interactions, and have been broadly classified into two groups; *cis*-encoded and *trans*-encoded [67,68,103]. The *cis*-acting sRNAs are encoded in the antisense strand of their target mRNAs, leading to the high sequence complementarity required for the interactions. T*rans*-encoded sRNAs, on the other hand, occur at genomic loci independent from their mRNA targets, share little sequence complementarity, and usually regulate multiple targets. Furthermore, the interactions between *trans*-encoded sRNAs and their targets involve short interspersed and often non-canonical base-pairing and are usually mediated by the RNA chaperone Hfq [67,104].

With the advent of high-throughput transcriptome profiling techniques, many new sRNA transcripts have been identified [13,15,105,106]. Additional sRNAs have been predicted based on *in silico* analysis [94,105]. While over 100 sRNAs have been identified in the gram-negative bacteria *Escherichia coli* and *Salmonella typhimurium*, many of their molecular functions have not yet been characterized [68]. Therefore, attempts have been made to develop new computational approaches for the prediction of sRNA targets in bacterial genomes [94,107].

Since sRNAs interact with their targets by base pairing, most previous methods for target prediction have relied on sequence and/or secondary structure based analysis. However, purely sequence-based approaches that have been successfully used in prediction of eukaryotic miRNA targets are not applicable to prokaryotic *trans*-encoded sRNAs due to the lack of a perfect

complementary region with mRNA targets [94]. Some improvements have been achieved using thermodynamic scoring of base pairs in short complementary regions [90,108].

Another target prediction method aims to identify the joint secondary structure of the interaction[109]. This technique faces the problem of identifying the native interaction from a combinatorially large number of possibilities, given that biologically functional interactions often do not correspond to the predicted structure with the minimum free energy [94,110]. This problem necessitates restricting the joint secondary structure prediction problem within a set of predetermined assumptions [94]. The earliest methods in this category applied a modified version of the single RNA secondary structure prediction algorithm to a concatenated sequence of the sRNA and the target mRNA to arrive at the joint secondary structure [92,111,112]. Subsequent secondary structure based methods have accounted for accessible regions in the secondary structures of the individual RNA molecules in formation of the interacting duplex [95,96].

Algorithms combining sequence conservation information with accessibility have shown promise [99,113,114]. These methods depend on the occurrence of the sRNA and its target mRNA in multiple bacterial species such that conserved motifs can be detected. This prerequisite limits the target search space to conserved genes only, and somewhat diminishes the advantage obtained by restricting the interaction to conserved, structurally accessible sequence positions.

While these bioinformatics methods have collectively demonstrated that various aspects of base-pairing interactions can be obtained from the sRNA and mRNA sequences alone, all predictions have been known to have an undesirably high false-positive rate (or low specificity). In most cases, the large number of false-positives result from the insufficiency of our current

understanding of the nondeterministic nature of these base-pairing interactions that allow each trans-encoded sRNA to regulate multiple mRNAs.

One possible way to improve the specificity of the predictions is to incorporate information from other factors that play a role in sRNA mediated regulation. For example, recent studies have suggested a role for accessory proteins involved in RNA metabolism in affecting mRNA steady state levels mediated by sRNAs [13,65,115,116]. However, the mechanistic details associated with the involvement of these proteins in the regulation are still largely undetermined.

An argument often exploited in bioinformatics is that since biological function is essentially encoded in nucleotide sequence, effective representation of sequence information should adequately enable predictive models for biological processes. This assumption has been particularly useful in the application of machine-learning algorithms for pattern recognition in biological sequences [117–119]. Machine-learning algorithms that are designed for non-linear pattern recognition capture relationships between descriptive features of biological sequences that are inaccessible by linear statistical methods.

Here we present a binary classifier for the prediction of sRNA targets based on the popular Random Forest algorithm [120]. The complex nature of biological systems has led to the widespread use of Random Forests for classification and regression problems in bioinformatics [121,122]. In the following sections, we elaborate on an iterative sampling based Random Forest model, hereby referred to as Balanced Random Forests (BRF) [123]. The BRF model uses a combination of numerically represented sequence features obtained from the sRNA and mRNA sequences to discriminate between sRNA and mRNA pairs known to have regulatory interactions and those that do not.

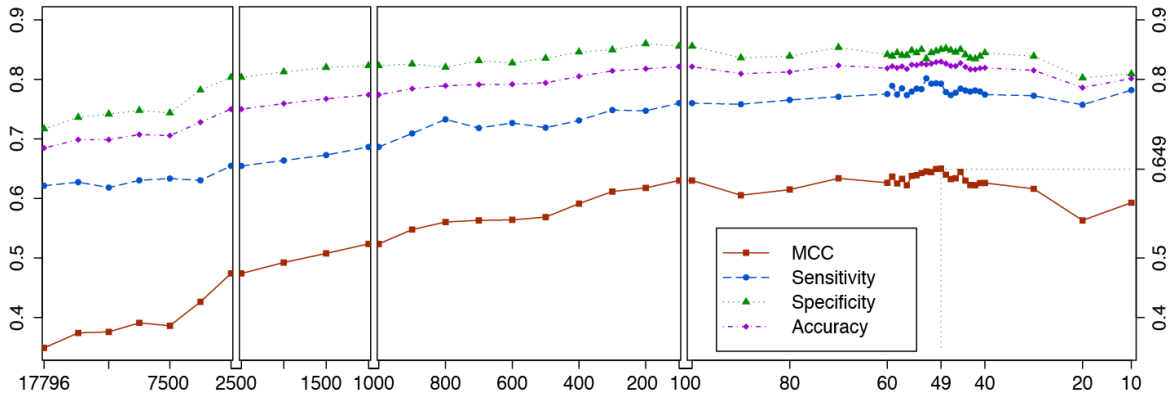


Figure 4.1: 10-fold cross validation (CV) performance over iterative feature selection: As feature selection progresses, varying number of features are removed at each iteration and evaluated over a 10-fold CV. For the sake of clarity, the plot is split into four subparts with varying ranges to include all the iterations.

Results

Evaluation of Selected Features

The complete feature set of 17760 features that was computed (as described in materials and methods) was subjected to iterative feature-selection. Each feature set was evaluated based on the 10-fold cross-validation Matthews Correlation Coefficient (MCC). The iterative removal of least important features progressively improved the 10-fold cross-validation MCC (**Figure 4.1**, Supplementary Table S2). The second round of selection for fine-tuning the feature-set was performed starting from a set of 60 features; a set of 49 features was settled upon as the optimal feature-set (Table S3, Figure S4). In the selected feature set, it is interesting to note that most of the k-mer frequencies obtained from the target sequences that differ only by the number of gaps can be consolidated into simple patterns. For instance, the two-letter patterns SWS$\{N\}_{1\text{-}3}$SWW, SSWW$\{N\}_{1\text{-}3}$SWWW, MMKK$\{N\}_{1\text{-}3}$KMMK and KKM$\{N\}_{1\text{-}3}$MKK encompass 12 of the 30 features obtained from mRNAs. Similar patterns SWSSSWW and KKK$\{N\}_{1\text{-}3}$MKK were

captured from the sRNA sequences amongst a number of low-complexity patterns. Among the low complexity patterns, stretches of 'R' were found to be a recurring pattern in sRNA sequences. In order to determine if these patterns had any positional preferences in the sequences, we counted the occurrence of each pattern in the source sequences. It is worth noting here, that RF feature selection does not necessarily select features that are enriched in the positive class. Rather, it is a non-linear combination of the frequencies that make the classification possible in the selected model. This is evident when we compare the distributions of the frequencies in the two classes (**Figure 4.2**). Furthermore, the position-wise counts of the kmer patterns show that some patterns are less abundant than others in the positive sequences (figure S4). Not unexpectedly, several kmer patterns show either increased or decreased counts in regions around the transcription start site.

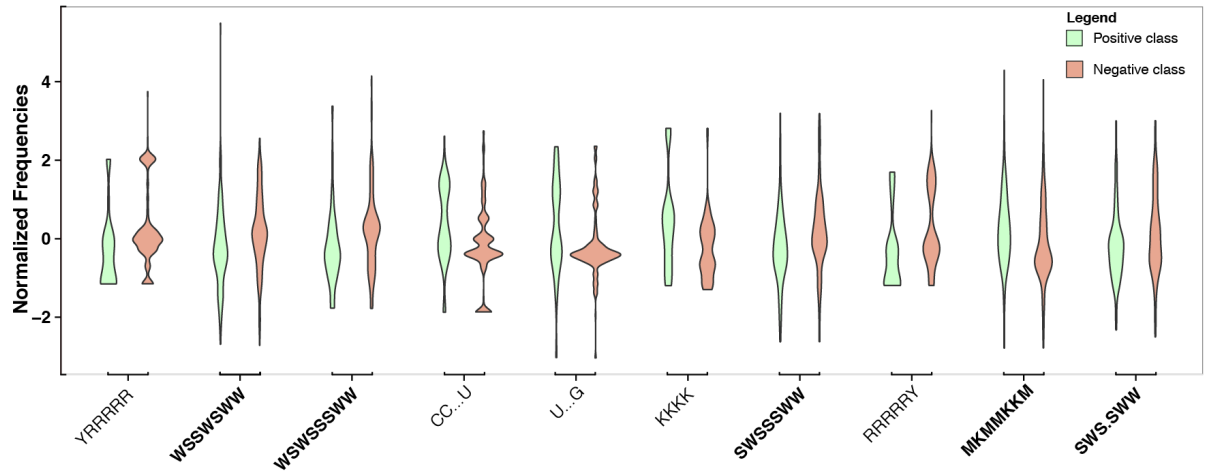


Figure 4.2: Violin-plots comparing the distribution densities of the top 10 selected features between the two classes. The frequencies for the patterns labeled in bold are contributed by the mRNA sequences.

A predictive BRF model was fitted using the selected feature set from the entire training data as described in the methods section. The resulting model, consisting of 1000 classification forests of 5 trees each, was evaluated using our blind test set. The blind test results for the final

model compared well with the 10-fold CV, indicating a good fit while retaining the balance in sensitivity and specificity.

Comparisons with other methods

To put the BRF's blind test performance in perspective, we made several comparisons with other models using the same test set. First, we established that performance was not compromised due to the sampling strategy employed by comparing against both weighted and unweighted conventional RFs constructed using the same training set. Both the models were tuned for *mtry* and grown to 5000 trees, keeping the parameters consistent with the BRF. The class weights were tuned for the weighted RF (WRF) for maximum OOB MCC. **Table 4.1** shows that although the three models were comparable in terms of overall accuracy, the BRF model did a better job of maximizing sensitivity (or true positive rate) without compromising on specificity (true negative rate).

Table 4.1: Comparison of performances of all the models on the blind test-set. Balanced Random Forest (BRF) compares favorably with both unweighted Random Forest (RF) and Weighted Random Forest (WRF) in terms of Matthews Correlation Coefficient (MCC), overall accuracy, sensitivity (or positive predictive rate) and specificity (negative predictive rate). All three RF models outperform the state-of-the-art TargetRNA2 and IntaRNA methods.

	Accuracy	Sensitivity	Specificity	MCC
BRF	79.21	80.49	78.33	0.58
RF	78.22	70.73	83.33	0.55
WRF	80.20	73.17	85.00	0.59
TargetRNA2	64.00	56.10	69.49	0.26
IntaRNA	41.00	92.68	5.08	-0.05

Table 4.2: Description of the alphabets used for sequence representation..*N*, representing any nucleotide is used in the patterns from both standard and two-letter alphabets.

Alphabet	Symbol	Origin of designation

Standard Alphabet	G	Guanine (G)
	A	Adenine (A)
	U	Uracil (U)
	C	Cytosine (C)
Two-letter alphabet 1	R	puRine (G or A)
	Y	pYrimidine (U or C)
Two-letter alphabet 2	M	aMino (A or C)
	K	Keto (G or U)
Two-letter alphabet 3	S	Strong interaction, 3 H bonds (G or C)
	W	Weak interaction, 2 H bonds (A or U)
Universal	N	aNy nucleotide (A or U or G or C)

Next, we determined how the BRF model compared against currently available state-of-the-art prokaryotic sRNA target predictors. For the analysis, the first software we selected was the recently published TargetRNA2 [113], which uses sequence conservation along with secondary structural features of the interacting sequences to make the predictions. Another recent algorithm, CopraRNA [114], have had recent success using conservation information from multiple sequence alignments. However, since CopraRNA functions in a fundamentally different way from our method, requiring multiple sequences for both sRNA and targets as input, a parallel comparison are difficult to make. CopraRNA's prediction results are corroborated by a previously published web server IntaRNA [96]. IntaRNA is a secondary structure based target prediction algorithm that account for accessible regions in the secondary structures of the targets. For the sake of comparison of contrasting methods, we included the IntaRNA webserver in our analysis. sRNA and mRNA sequences from our test set were submitted to the respective web-servers as described in the published articles with default parameters. BRF outperformed both IntaRNA and TargetRNA2 in terms of overall accuracy and MCC (**Table 4.2**). Of the two methods, only TargetRNA2 distinguishes the two classes with reasonable competence. With default settings, IntaRNA tended to find interactions with a negative MFE for most sequence

pairs and predict them to be true interactions. This issue can be somewhat circumvented when IntaRNA makes predictions for a given sRNA across all mRNAs on a genome and the interaction MFEs are fitted to an extreme-value distribution [99,114]. Nonetheless, it is apparent that additional sequence information is required in order to arrive at reasonably reliable predictions.

Genome-wide predictions for individual sRNAs

Finally, the BRF classifier model can be used for fast genome-wide prediction of targets for a given sRNA. Predictions can be sorted by their associated probability values, and high-confidence predictions may be used for downstream analyses. We applied the model for target predictions for three sRNAs, across the *E. coli* genome, namely RyhB, OmrA and IstR, as case studies (Table S5). After taking out the targets already present in the training set from the genome-wide predictions, the top 100 predictions for each sRNA were subjected to gene ontology term enrichment of the molecular function category using the DAVID server [124]. The most significant 5 terms for each sRNA are shown in **Table 4.3**. A number of predictions for the well studied RyhB and OmrA that group into well-defined functional categories, indicating that the trained model captures sequence information from these sRNA interactions included in the training set, and predicts new interactions from the genome that share these features. The functional categories obtained from the enrichment analysis is in accordance with previous studies on RyhB and OmrA [125,126]. While IstR is not as well represented in the training set as the other two sRNAs, the enriched GO-terms suggest that it might be involved in similar regulatory networks as RyhB.

Table 4.3: Gene Ontology (GO) term enrichment analysis results for the three sRNAs tested for genome-wide predictions in *E. coli*. Only the top 5 most significant terms from molecular function ontology are shown here.

sRNA	Gene Ontology Term	No. of Genes	P-value
RyhB	GO:0016765~transferase activity, transferring alkyl or aryl (other than methyl) groups	4	2.50E-02
	GO:0015932~nucleobase,nucleoside, nucleotide and nucleic acid transmembrane transporter activity	3	3.60E-02
	GO:0016151~nickel ion binding	3	4.30E-02
	GO:0043169~cation binding	19	5.10E-02
	GO:0043167~ion binding	19	5.30E-02
OmrA	GO:0016151~nickel ion binding	4	4.40E-03
	GO:0042626~ATPase activity, coupled to transmembrane movement of substances	6	1.40E-02
	GO:0043492~ATPase activity, coupled to movement of substances	6	1.40E-02
	GO:0015399~primary active transmembrane transporter activity	6	1.60E-02
	GO:0015405~P-P-bond-hydrolysis-driven transmembrane transporter activity	6	1.60E-02
IstR	GO:0016151~nickel ion binding	4	6.40E-03
	GO:0004555~alpha,alpha-trehalase activity	2	9.60E-03
	GO:0043169~cation binding	23	1.70E-02
	GO:0043167~ion binding	23	1.80E-02
	GO:0046872~metal ion binding	22	2.30E-02

Discussion

In this article, we present a novel algorithmic approach to the prediction of trans-acting sRNA targets in bacteria. By providing a classification approach using sequence information independent of alignments, this method makes an useful addition to the spectrum of computational tools available for sRNA target prediction. In the past, the scarcity of known sRNA-target interactions had prohibited the effective development of supervised algorithms for prediction. By assembling an up to date dataset consisting of only experimentally verified interactions from the literature, we were able to obtain competent classification performance on

the blind test-set. The balanced-sampling strategy employed for construction of the RF model addressed the many-to-many cardinality of the interactions in the dataset, and allowed for the optimization of both sensitivity and specificity for an imbalanced dataset at the same time. Compared on the test-set, the BRF algorithm fares favorably with current state of the art predictors of bacterial sRNA targets.

Feature selection using RF's intrinsic "variable importance" measure allowed us to narrow down an expansive list of sequence pattern frequencies to a set of 49 features with the highest discriminative power. This feature set is likely to incorporate information essential for the interactions apart from base-pairing alone. A few patterns appear to coincide with known signatures, such as AU-rich Hfq binding motifs [41,127], while others are more cryptic at the moment. This is primarily because RF uses the features in a combinatorial fashion, and individual features would impart little or no predictive power. It is likely that some features contributing to the model relate to yet to be discovered mechanisms of sRNA regulation. Recent findings in bacterial sRNA regulation have revealed new players in the pathway [13,115,116,128]. Sequence signatures have also been found to be associated with other aspects of gene-expression, such as mRNA stability and translational efficiency [129,130]. Whole-genome motif enrichment studies in the newly discovered aspects of sRNA regulation will shed some light on the roles of sequence patterns.

Finally, we show that the BRF model can be used on a genome-wide scale for fast prediction of global targets for a given sRNA. However, despite the competent sensitivity and specificity of our final model on the blind test set, when applied to the complete set of coding mRNA sequences in the genome, the model predicts a large number of interactions with a probability greater than 0.5 (the default threshold). Therefore, additional filtering steps may be

required for the selection of candidates for experimental validation. In the two case studies we presented in the results, the top 100 predictions in the sorted lists do conform to functional enrichment consistent with that of their known targets. Thus, a sorted list of predictions with highest probabilities may be subjected to functional enrichment, and/or correlation analysis with other state-of-the-art methods to arrive at a experimentally manageable list of high-confidence predictions. We anticipate this method will be a good starting point for prediction pipelines, and will lead to the discovery of new sRNA-mRNA interactions.

Methods

Datasets

Supervised learning algorithms such as RFs require a training set comprised of instances known to belong to the distinct categories in question. In most cases, training datasets that are comprised of experimentally verified instances offer the highest confidence for predictive modeling. In bacteria, interactions between sRNAs and mRNAs are commonly tested either through a genomic deletion of the sRNA or constitutively over-expressing it in a plasmid, and subsequent measurement of transcript level changes in potential mRNA target candidates.

We obtained an initial list of sRNA-gene pairs that have been reported in the literature to either interact or not have any effect, from previously published databases [131–133]. RNA regulatory and metabolism pathways vary considerably between gram-positive and gram-negative species, and divergent species of bacteria in general. Keeping this in mind, we restricted our dataset to the widely studied enterobacteriales *Escherichia coli* and *Salmonella typhimurium*. Additional instances for both classes were collected from the literature, resulting in a final

dataset up to 168 experimentally verified interacting pairs of sRNAs and mRNAs, and 248 pairs that do not interact (Table S1).

Feature Vectors

The RF algorithm learns to distinguish between the two classes based on the information content in an appropriate numerical representation of the dataset. In order to encode the RNA sequences in numerical form, we computed frequencies of occurrence of a substantially large set of k-mer patterns in the sequences to serve as numerical features for classification. The rationale for using a large starting feature set was that RF's inherent feature-selection methodology may be used to find the best non-linear combination of a subset of these features that best discriminates between the two classes.

Each instance of a sRNA-mRNA pair in the dataset was represented by three individual sequences, and the pattern frequencies described below were calculated for each. The full sRNA and mRNA sequences were augmented by the sequence regions around the mRNA translation start sites (defined here as subsequence starting 150 nt upstream to 100 nt downstream of the start codon), given that most interactions with sRNAs occur in this region.

Table 4.4: Summary of sequence patterns used for calculation of frequency features. Here, *L* is a given letter from the alphabet and *N* is any nucleotide (unspecified).

Alphabet	Pattern	Feature Description
Standard Alphabet	$L_{(1-4)}$	Nucleotide, di-, tri- and tetranucleotide frequencies.
	$L_iN_{(1-3)}L_{ii}$, $L_iN_{(1-3)}L_{ii}L_{iii}$ and $L_iL_{ii}N_{(1-3)}L_{iii}$	Di- and trinucleotide frequencies interspersed with stretches (1-3nt in length) of unspecified nucleotides (*N*).
Two-letter alphabets	$L_{(3-8)}$	Frequencies of 3-8mers of two-alphabet letters.
	$L_{(3-4)}N_{(1-3)}L_{(3-4)}$	Frequencies of 6-8mers of two-alphabet letters, interspersed with stretches (1-3nt in length) of unspecified nucleotides (*N*)

Frequencies from the standard four-letter RNA alphabet (A,U,G and C) were calculated for all pattern combinations ranging from mono- to tetra-nucleotides. In order to account for the short interspersed interactions commonly found in sRNA-mRNA duplexes, the patterns were

extended by incorporating 1-3 nucleotide stretches of a "wild-card" letter N, where N can match any nucleotide (**Table 4.4**).

The three RNA sequences for each instance were also translated from the four-letter code to two-letter alphabets as proposed by the IUPAC-IUB Commission on Biochemical Nomenclature (CBN) [134] (**Table 4.2**). The reduced alphabet encoding makes the calculation of non-zero frequency values of longer pattern lengths, ranging from tri- to octamers. As with the standard alphabet, the range of the patterns was expanded with the incorporation of 1-3 nucleotide stretches of letter N (**Table 4.4**). Combining all the frequency features for the representative sequences in each instance across all the alphabets used amounted to an extensive starting feature set consisting of 17796 features.

Balanced Random Forests

Random Forest (RF) [120] is an ensemble-learning algorithm for classification, regression and clustering based on decision trees. Being based on the theory of ensemble learning allows the algorithm to learn complex classification tasks, and allows it to identify non-linear interactions between features. Iterative sampling of prediction variables (or features) allows the use of a large number of features as compared to the number of observations and to assess the importance of individual features in an embedded feature-selection method. Since its inception, RFs have gained popularity in several learning problems in bioinformatics [121] owing to their innate properties that make the method adaptable to a variety of situations.

When using classification datasets that are "imbalanced", i.e. when one class is under-represented in number of observations, predictions from classifiers are often biased towards the majority class. To deal with this issue, RFs allow the most commonly adopted approach of cost-sensitive learning in the form of Weighted Random Forests (WRF), where a high "cost" is

assigned to the misclassification of the minority class [123,135]. Another way to counteract the class-imbalance problem is to down-sample the majority class, over-sampling the minority class, or both [123]. In most cases, sampling out observations to balance class sizes leads to loss of information, making cost-sensitive learning the preferred approach. However, since RFs use an ensemble of tree-based classifiers, iterative down sampling allows the incorporation of all observations from the majority class distributed among the individual trees. The Balanced Random Forest (BRF) approach we used in this work was based on this idea, where the BRFs were constructed as follows:

i. A bootstrap sample was drawn from the minority class (the positive class, in this case), which was roughly 90% of all the minority class instances.
ii. A second bootstrap sample was drawn from the majority class (the negative class), which was equal in size to the minority positive class sample obtained in i.
iii. A small classification RF with *n* trees was initiated using the data obtained in step i. and ii. In the first iteration, the BRF was initiated with this small RF. In subsequent iterations, the RF was added to the combined BRF.
iv. Steps i to iii were repeated *m* number of times, resulting in a BRF consisting of an ensemble $n \times m$ trees. Predictions were aggregated over all the trees in the ensemble to arrive at the final prediction.

It is worth noting that the dataset used for any given tree may contain semi-redundant features, since each sRNA may regulate multiple mRNA targets, and each mRNA may be regulated by multiple sRNAs (**Figure 4.3**). The bootstrap sampling from the minority positive

class in step i was included to control for this redundancy, even though the class imbalance was addressed by under-sampling the majority class alone.

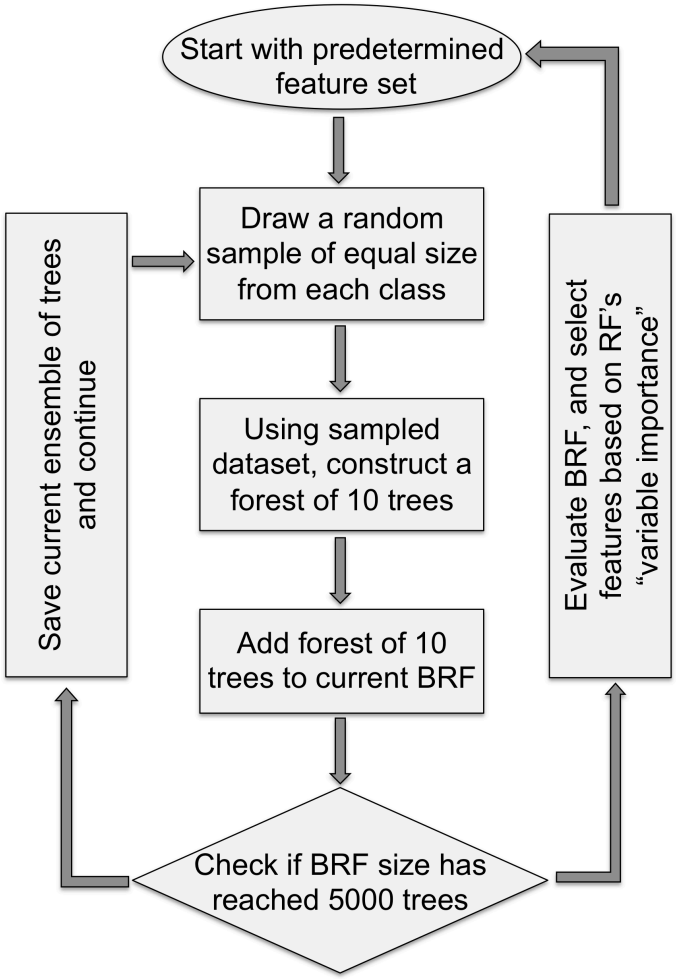


Figure 4.3: Flowchart illustrating training and feature selection steps for the Balanced Random Forest (BRF) algorithm. The left loop constructs the current Balanced Random Forest model, while the right loop extracts features based on variable importance values calculated on the current model.

Model training and feature selection

A blind test-set consisting of approximately 25% of the total number of instances was randomly sampled from each class. The remaining dataset was used for training, which involved iterative feature evaluation and selection using RF's inherent "variable importance" measure [120]. As

previously mentioned, all existing computational methods for the prediction of sRNA targets suffer from a high false positive rate [94]. Although our BRF methodology aims to boost the true-positive rate (sensitivity), it remained imperative that both sensitivity and specificity were optimized for model selection. In this regard, the Matthews correlation coefficient (MCC) offered one of the most balanced measures of a binary classifier's performance, by incorporating counts for true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) in its formulation:

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

Feature sets were evaluated based on their average 10-fold cross-validation MCC. Feature selection was performed in two stages: a number of features roughly proportional to the total number of features was eliminated in each iteration of the first round, followed by a second round of iterative removal of one least important feature. The feature set with the highest MCC was then used to train the final model, which was subsequently evaluated on the blind test-set.

The algorithm was implemented in the R programming language using package "randomForest". The original R source code, accessory scripts and data are freely available at https://github.com/j-mitra/BRF-sRNA-target.

CHAPTER 5

KNOWLEDGE BASED IDENTIFICATION OF TRANS-ACTING REGULATORY SRNA TARGETS IN *ESCHERICHIA COLI*

Mitra. J, Mohanty. B, Kushner. S.R; To be submitted to Nucleic Acids Research

Abstract

Small non-coding RNAs (sRNAs) have emerged as global regulators in major pathways in all organisms known to man. In bacteria, the largest class of riboregulators act by base-pairing to mRNA targets, regulating them post-transcriptionally. Among these, a subclass of sRNAs act in *cis*, being encoded on the opposite strand of their targets. The other, more prevalent class act in *trans*, being encoded at different locations from either single or multiple targets and do not share extensive sequence complementarity. Several aspects of regulation by *trans*-acting sRNAs make it challenging to identify their targets.

We have previously presented a machine learning algorithm based on the random forest classifier to predict mRNA targets for *trans*-acting in bacteria. Here, we describe how the algorithm can be applied on a genome-wide scale to make successful predictions in *E. coli*. A selection of the top scoring mRNA predictions informed by additional criteria were validated experimentally, revealing regulatory interactions by multiple sRNAs in *E. coli*.

Introduction

Recent years have witnessed the rise of non-coding RNAs as regulators of gene-expression in all kingdoms of life. The development of high-throughput transcriptome profiling technologies such as whole genome tiling-microarrays and RNASeq have led to the discovery of numerous non-coding RNA species. While the number of non-coding RNAs continue to grow rapidly, identification of their biological roles still remains a formidable challenge.

In prokaryotes, systematic studies for the identification and characterization of regulatory non-coding RNAs have largely focussed on the model bacteria *Escherichia coli* [5,16,103] and *Salmonella Enterica sp*. [5,16]. In *E. coli*, over a hundred small regulatory non-coding RNAs (hereby referred to as sRNAs) have been identified [98,136]. sRNAs that are encoded in *cis*, i.e.

the antisense strand of protein coding mRNAs, target the mRNAs on the opposite strand for regulation. In contrast, the so called "*trans*-acting" sRNAs are encoded in intergenic regions of the genome, and regulate multiple mRNA targets at uncorrelated genomic loci. The lack of complementarity between the sRNAs and their targets have made target identification very difficult.

The majority of these novel sRNA transcripts, however, have no biological function identified. This widening gap between the discovery and functional annotation of novel sRNAs has necessitated the development of several biocomputational approaches [68,98] to facilitate the annotation process. Early attempts in sRNA target predictions used simplistic sequence alignment based techniques [94,98] or secondary structure predictions of the ncRNA sequence concatenated with the mRNA [111,112]. The highly variable nature of these interactions have made it difficult for these early methods to have any noteworthy success. Subsequent efforts have incorporated structural accessibility and/or sequence conservation information of the interacting RNA molecules. Sequence conservation is generally accounted for by using multiple sequence alignments between sRNA and target sequences from related species [97,99,113]. Although these approaches are limited by the requirement that both the sRNA and the target be conserved in multiple species, the effective combination of comparative methods with secondary structural information has been successful in discovery of novel interactions [99,113].

Despite apparent improvements in prediction capabilities of current algorithms, a large number of sRNAs have no assigned function. *Trans*-acting sRNAs have been extensively studied in *E. coli*, but fewer than 20 sRNAs have known targets [136]. While it is difficult to predict whether novel non-coding RNA species mediate regulation through base pairing, the sRNAs known to act in *trans* on few targets are likely to interact with many more mRNAs [68].

Bioinformatic approaches have often utilized the assumption that the requisite information for biological function is encoded in the sequence. Thus, numerical quantification of sequence patterns have been used for predictions [117–119]. Previously, we have presented a machine learning classification based approach for the prediction of *trans*-acting sRNA targets in bacteria (Mitra and Kushner, in review). Our algorithm is based on the popular random forest (RF) classifier and uses k-mer and gapped k-mer frequencies as prediction features. Since our method performed competently on a blind test set and compared favorably with existing state-of-the-art methods, we anticipated that the application of the algorithm on a genome-wide scale would result in the identification of novel regulatory interactions.

Materials and Methods

Genome-wide predictions using Balanced Random Forest

A Balanced Random Forest (BRF) model was trained using the full training dataset as described in the previous chapter. Genome-wide predictions were made for various sRNA. For the purpose of experimental validation, genome-wide predictions were made for the *Escherichia coli* sRNAs MicC, RybB, RseX, OxyS, DicF and RprA. Functional enrichment was done using the top scoring 100 predicted genes using the DAVID web server [137], as described previously.

Filtering genome-wide results using BRF proximities

Our filtering strategy was done by computing RF's proximity measures for positively predicted interaction in the genome to a benchmark set of high confidence known interactions. RF proximities are a similarity measure computed between predictions made by a trained RF model [120,138]. Essentially, proximities are measured as the ratio of the number of trees in the

forest traversed by two independent instances (or predictions) following identical paths, to the total number of trees in the RF model [120].

While RF proximities are generally used for unsupervised learning applications using RF, we repurposed the measure as an additional criteria to prediction probabilities for selecting candidates for experimental validation. To accomplish this task, we performed leave-one-out (LOO) validation on the positive set using the BRF strategy described previously. Basically, each experimentally validated positive interaction was taken out of the training set while the remaining set was used for training. The trained models iteratively made predictions on the instance that was removed from its training data to output a probability. The variability introduced by the sampling from the larger negative set in the BRF algorithm was accounted for here by repeating the LOO evaluation 100 times with different seeds for random number generation. Positive interactions that were predicted with a probability higher than 0.75 (a predefined threshold for high-confidence) every single time were included in the benchmark set.

Bacterial Strains and Northern Blot Analysis

The BRF model was experimentally validated using northern blot analysis of predicted target mRNAs in *Escherichia coli* MG1655 and derivative strains. The sRNA knockout strains were generously provided by Gisele Storz at the National Institutes of Health (NIH). The mutant strains were generated by replacing each sRNA gene with a kanamycin resistance cassette [139].

Cells were grown on standard Luria broth and harvested for total RNA extraction during exponential (Klett 50, No. 42 green filter), late exponential (Klett 125), and stationary (Klett. 200) phases of growth. RNA was extracted using the RNA*snap*™ protocol [140]. The RNA

samples were further purified using acidic phenol/chloroform extraction and ethanol/Na-acetate precipitation as previously described [141].

Twenty μg of each RNA sample were separated on either 5% denaturing polyacrylamide gels, or 1.5% agarose gels (with glyoxal added to samples) [142]. The polyacrylamide gels were used for analyzing transcripts less than 1 kb in size. Radiolabeled probes for the northerns were prepared in one of two ways: short 20 nucleotide DNA oligomers end-labeled with ɣ-^{32}P-ATP using polynucleotide kinase [141], or PCR synthesized longer DNA fragments randomly labeled with α-^{32}P-dATP using the Klenow fragment of DNA polymerase I [141]. Northern hybridization was performed by incubating the membranes in ULTRAhyb™ hybridization buffer at T_m - 10°C when probed with radiolabeled oligonucleotides, or 65°C when probed with radiolabeled longer DNA fragments [141]. Hybridization was visualized on a Storm 840 PhosphorImager (GE Healthcare) and band intensities were quantified using ImageQuant™ software (GE Healthcare). Fold change values were calculated using the local median background correction method, and the reported values were an average of at least two independent biological replicate experiments.

Results

The number of predictions scoring higher than the default probability threshold varied greatly among the six sRNAs tested. Typically, we observed sRNAs that were well represented in the training dataset got higher scoring predictions overall. This result appeared to be a consequence of overfitting in the model, despite stringent measures having been taken to avoid this very problem. Thus, the prediction probability distribution RybB, which had 30 known interaction instances in the training dataset looks very different from that of DicF, with only one interaction, and OxyS, with only two interactions (histofigs). This makes the selection of

candidate predicted interactions for validation require additional constraints, in order to minimise false positives during experimental validation.

The benchmark-proximity filter employed in this study aimed to overcome this problem by identifying predicted interactions that were similar to high-confidence known interactions in the training dataset. The proximity filter was effective in two ways: first, it greatly reduces the total number of predictions across the genome; and second, the predictions could be sorted using additional criteria such as average proximity to all the benchmarks and/or the total number of benchmark interactions higher than a predefined proximity threshold. Around 4-5 interactions were selected from these sorted lists for each sRNA for validation. The selections were made from amongst the top 15 predicted results, based on the above mentioned criteria, and biological interest.

Table 5.1: Average linear fold change in expression for predicted targets. Negative values indicate down-regulation.

Predicted Target	sRNA	Fold Change in Expression		
		Log	Mid-log	Stationary
btuB	MicC	-1.51	-4.52	1.58
dppA	MicC	14.40	36.70	1.41
sdhC	MicC	1.88	-0.53	11.03
	RybB	0.87	0.44	1.62
	RseX	0.41	0.30	3.09
sthA	MicC	-1.16	-2.68	-1.04
	RybB	-1.69	-3.89	-1.37
	RseX	-4.49	-4.24	-1.04
chbC	DicF	0.97	12.86	162.65
	RprA	3.28	11.63	130.74
yqjG	MicC	1.07	4.42	0.84
yejA	MicC	0.42	0.84	4.92
	RprA	0.42	0.84	4.92

Based on these criteria, we identified a number of targets commonly predicted with high confidence in all or most of the sRNAs in this study. Interestingly, some of the genes predicted

to be regulated by multiple sRNAs were also predicted to be interaction hubs by CopraRNA [99] a target predicting program that utilizes a very different approach. In particular, both *sdhC* and *csrD* consistently scored high for multiple sRNAs. Amongst the potential hubs, we were able to detect *sdhC* in our northern analyses, and it showed differential expression in the three sRNA deletion mutants it was tested against. A regulatory effect was decided to be positive when the fold change in the northern analyses was greather than 1.5 fold, keeping it consistent with previous studies [99,114].

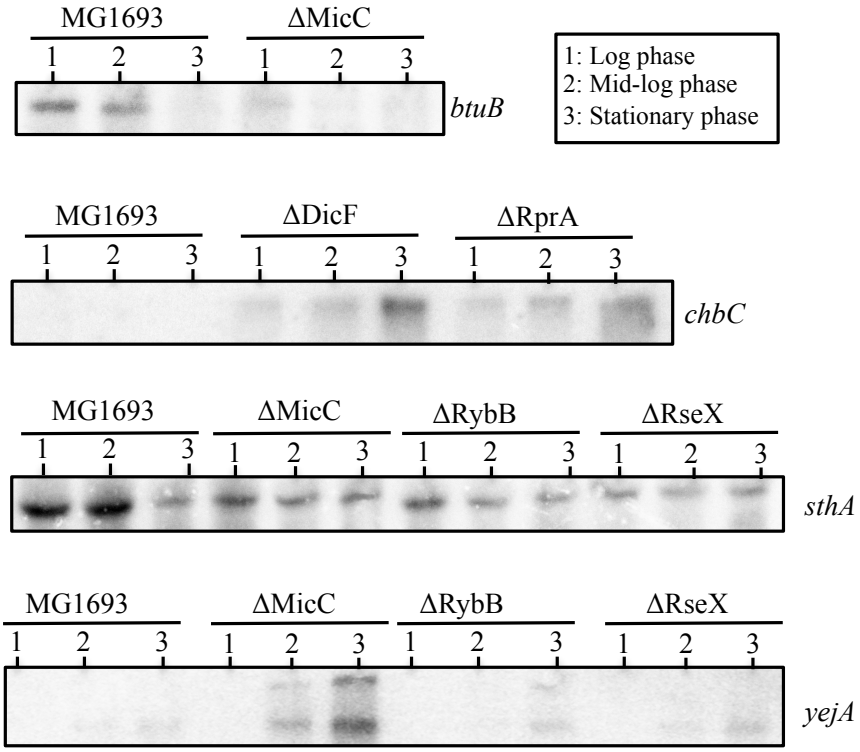


Figure 5.1: Northern blot analysis of predicted targets. Total RNA was extracted at Klett. 50, 125, and 200 for log-phase, mid-log phase and stationary phase, respectively, as described in the Materials and Methods. The blots shown here were performed on 1.5% agarose gels. All the genes were probed by longer DNA fragments as described in materials and methods, except for yejA, which was probed by an end-labeled oligomer.

The northern analyses revealed five novel regulatory interactions for MicC, which is known to regulate outer membrane proteins from previous studies [143]. *sdhC* was most strongly regulated by MicC, and only moderately by RybB and RseX. MicC regulated *dppA* negatively,

and *btuB* and *sthA* positively, the strongest effect usually seen in the mid-log phase (Table 5.1, Figure 5.1). Interestingly, *sthA* is the soluble protein alternative to the membrane bound pyridine nucleotide transhydrogenase (*pntAB*) [144,145]. This target was a departure from the known membrane associated genes that these sRNAs regulate. MicC also positively regulates *btuB*, another outer membrane porin that mediates the transport of cyanocobalamin across the membrane [146,147]. The operon *yejABEF*, previously found to be regulated by RspA [75], was strongly regulated by MicC (Figure 5.1, Table 5.1) and RprA (Table 5.1). The operon was previously predicted to be around 6 Kb in size. However, we detected two bands, both beyond the range of our riborulers. The smaller band was a little over 6 kb, and the larger band was between 6.5-7 kb.

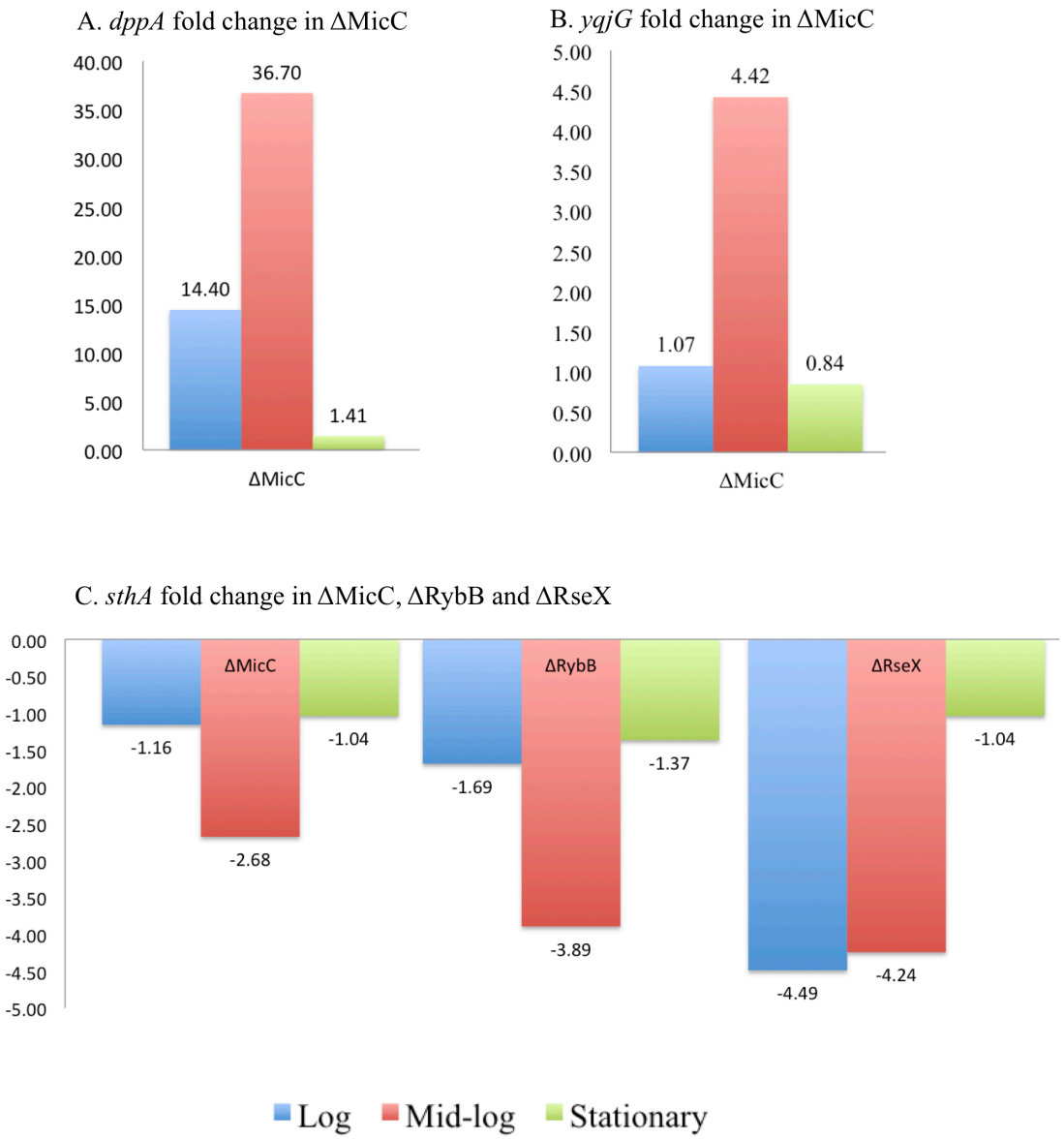


Figure 5.2: Barplots depicting fold changes in predicted targets for sRNAs tested.

Discussion

The balanced random forest (BRF) model used for the prediction of *trans*-acting sRNA targets described in this study was trained using numerical representation of the sRNA and mRNA sequences using k-mer and gapped k-mer pattern frequencies only. The iterative feature selection methodology described previously aims to capture information required for distinguishing interacting pairs of sequences from the pairs that do not interact. In doing so, we eliminated the

need for secondary structure predictions and multiple-sequence alignment, possibly accounting for sequence characteristics required for regulation that might be overlooked by other methods. The sampling strategy employed for training the BRF was intended to maintain an equilibration of sensitivity and specificity. The BRF had shown promising results on counts of both sensitivity and specificity when tested on the blind test set. The observation of variable number of total predictions between sRNAs is due to the unequal representation of the sRNAs in the training dataset. Moreover, it is safe to assume that even the current training dataset collected from literature is a gross under-representation of the underlying interaction network in the genome itself. This leads to a large number of positive predictions when the model is applied on a genome-wide scale. A high false positive rate is an issue faced by all approaches to this problem [105].

The proximity based filtering approach we applied to workaround these issues has made the selection of candidates for experimental validation easier. Since the model was trained on data from the closely related enterobacteria *E. coli* and *S. typhi*, we anticipated these experimental validations on predictions on the *E. coli* genome to reveal novel interactions. It is interesting to note, that although the training dataset did not distinguish between up-regulatory interactions from down-regulatory ones, our northern blot analyses have revealed both kinds of interactions from the predictions. The experimental results show that our prediction results lead to identification of novel interactions.

CHAPTER 6

CONCLUSIONS

Summary

Regulation by non-coding RNAs is a highly active and rapidly developing area of research. The widespread influence of these regulatory networks have made researchers working on many different biological pathways interested in them. In bacteria, *trans*-acting small regulatory RNAs (or sRNAs) have emerged as the largest class of non-coding RNAs that mediate regulation by base-pairing with target mRNAs. As was outlined in the introductory chapters 2 and 3, several aspects of regulation by *trans*-acting sRNAs make the prediction and identification of their regulatory targets a non-trivial challenge.

The original study presented in chapter 5 introduces a novel machine-learning based classifier to predict regulatory targets of *trans*-acting sRNAs. The model was trained on a high-quality dataset of experimentally validated pairs of sRNA and targets that interact, and those that do not interact. The algorithm incorporated an additional sampling step to the popular random forest algorithm [cite] to tackle the imbalance in the data, which usually biases supervised classifiers towards the majority class. By incorporating an additional sampling step that samples an equal number of instances from either class, the trained model performed equally well on counts of both sensitivity and specificity when tested on the blind test set. Random forests intrinsic feature selection capabilities were used to select the feature subset from a starting set of a combinatorially large number of features. Thus the model trained from the best performing set of sequence features allows the random forest algorithm to capture information required for distinguishing interacting pairs of sRNA and mRNA sequences from pairs of sequence that do

not interact. In doing so, we eliminated the need for secondary structure predictions and multiple-sequence alignment, possibly accounting for sequence characteristics required for regulation that might be overlooked by other methods. The performance on the blind test set also compared favorably against the latest state-of-the-art algorithms available.

The speed and performance of the trained model showed promise in discovery of novel interactions in bacteria. In chapter 6, this possibility is explored by making genome-wide predictions for six *trans*-acting sRNAs in *Escherichia coli*. In order to confirm predicted interactions, a manageable number of likely candidates needed to be picked for experimental validation. Unsurprisingly, we observed that genome scale predictions result in a larger number of positive predictions for each sRNA than is biologically probable. In order to increase the chances of selecting true-positives, an additional filter was incorporated using random forest's intrinsic similarity measure, the proximities that can be computed using a trained model. To use this as an additional selection criteria to the prediction probabilities (or votes), a high-confidence benchmark set was created, to which proximities of new predictions were computed from the BRF model. Several of the candidates that were tested using these criteria have been found to be regulated by the sRNAs in the study.

Future perspectives

The biggest advantage of knowledge based learning algorithms is that these models get better as they can be retrained, with new data added to the training set. Having stated that, the data being collected needs to incorporate more information. This may include stages or conditions that these sRNAs exert regulation, and/or the strength and type of regulation that they bring about. Experimental developments in eukaryotes have now enabled elucidation of RNA-

RNA interactions on a genome-scale. These technologies still fall short of capturing these interactions in bacteria, where the half-lives of most RNA species are very short.

Computational techniques rely on quality, quantity and design of experimental studies. It is safe to say, that most computational studies, including comparative methods and joint secondary structure prediction methods will benefit from experiments that address binding regions of regulatory interactions. In bacteria, performing these experiments is challenging, because most RNA molecules are extremely short lived in the cell. As various research groups in the field combine multiple techniques to arrive at new ideas to discover more about sRNAs in bacteria, mathematical modeling techniques will always be useful in further facilitating new discoveries.

BIBLIOGRAPHY

[1] G. Storz, "An expanding universe of noncoding RNAs.," *Science*, vol. 296, no. 5571, pp. 1260–3, May 2002.

[2] S. Altuvia, "Identification of bacterial small non-coding RNAs: experimental approaches," *Curr. Opin. Microbiol.*, 2007.

[3] C. M. Sharma and J. Vogel, "Experimental approaches for the discovery and characterization of regulatory small RNA," *Current Opinion in Microbiology*, vol. 12, no. 5. pp. 536–546, 2009.

[4] K. V Morris and J. S. Mattick, "The rise of regulatory RNA.," *Nat. Rev. Genet.*, vol. 15, no. 6, pp. 423–37, Jun. 2014.

[5] S. Gottesman, "Micros for microbes: non-coding regulatory RNAs in bacteria.," *Trends Genet.*, vol. 21, no. 7, pp. 399–404, Jul. 2005.

[6] P. Stougaard, S. Molin, and K. Nordström, "RNAs involved in copy-number control and incompatibility of plasmid R1.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 78, no. 10, pp. 6008–12, Oct. 1981.

[7] J. Tomizawa, T. Itoh, G. Selzer, and T. Som, "Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 78, no. 3, pp. 1421–5, Mar. 1981.

[8] T. Mizuno, M. Y. Chou, and M. Inouye, "A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA).," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 81, no. 7, pp. 1966–70, Apr. 1984.

[9] T. Suzuki, C. Ueguchi, and T. Mizuno, "H-NS regulates OmpF expression through *micF* antisense RNA in *Escherichia coli*," *J. Bacteriol.*, vol. 178, no. 12, pp. 3650–3653, Jun. 1996.

[10] K. Montzka Wassarman, “Small RNAs in *Escherichia coli*,” *Trends Microbiol.*, vol. 7, no. 1, pp. 37–45, Jan. 1999.

[11] J. Livny and M. Waldor, “Identification of small RNAs in diverse bacterial species,” *Curr. Opin. Microbiol.*, vol. 10, no. 2, pp. 96-101, 2007.

[12] S. Landt and E. Abeliuk, “Small non-coding RNAs in *Caulobacter crescentus*,” *Mol. microbiol.,* vol. 68, no. 3, pp. 600-614, 2008.

[13] M. B. Stead, S. Marshburn, B. K. Mohanty, J. Mitra, L. P. Castillo, D. Ray, H. Van Bakel, T. R. Hughes, and S. R. Kushner, “Analysis of *Escherichia coli* RNase E and RNase III activity *in vivo* using tiling microarrays,” *Nucleic Acids Res.*, vol. 39, no. 8, pp. 3188–3203, 2011.

[14] A. Sittka, S. Lucchini, K. Papenfort, and C. Sharma, “Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq,” *PLoS Genet*, vol. 4, no. 8, e1000163, 2008.

[15] R. Raghavan, E. A. Groisman, and H. Ochman, “Genome-wide detection of novel regulatory RNAs in E. coli.,” *Genome Res.*, vol. 21, no. 9, pp. 1487–1497, 2011.

[16] L. S. Waters and G. Storz, “Regulatory RNAs in bacteria.,” *Cell*, vol. 136, no. 4, pp. 615–628, 2009.

[17] P. Babitzke and T. Romeo, “CsrB sRNA family: sequestration of RNA-binding regulatory proteins,” *Curr. Opin. Microbiol.*, vol. 10, no. 2, pp. 156-163, 2007.

[18] K. Lapouge, M. Schubert, F. H.-T. Allain, and D. Haas, “Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour.,” *Mol. Microbiol.*, vol. 67, no. 2, pp. 241–53, Jan. 2008.

[19] K. M. Wassarman, “6S RNA: a regulator of transcription.,” *Mol. Microbiol.*, vol. 65, no. 6, pp. 1425–31, Sep. 2007.

[20] F. Kalamorz, B. Reichenbach, W. März, B. Rak, and B. Görke, “Feedback control of

glucosamine-6-phosphate synthase GlmS expression depends on the small RNA GlmZ and involves the novel protein YhbJ in *Escherichia coli*.," *Mol. Microbiol.*, vol. 65, no. 6, pp. 1518–33, Sep. 2007.

[21] F. J. Grundy and T. M. Henkin, "From Ribosome to Riboswitch: Control of Gene Expression in Bacteria by RNA Structural Rearrangements," *Crit. Rev. Biochem. Mol. Biol.*, vol. 41, no. 6, pp. 329-338, Oct. 2008.

[22] R. K. Montange and R. T. Batey, "Riboswitches: emerging themes in RNA structure and function," *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 117–133, Jun. 2008.

[23] E. Nudler and A. Mironov, "The riboswitch control of bacterial metabolism," *Trends Biochem. Sci.*, vol. 29, no. 1, pp. 11-17, 2004.

[24] J. A. Collins, I. Irnov, S. Baker, and W. C. Winkler, "Mechanism of mRNA destabilization by the *glmS* ribozyme.," *Genes Dev.*, vol. 21, no. 24, pp. 3356–68, Dec. 2007.

[25] S. Brantl, "Regulatory mechanisms employed by cis-encoded antisense RNAs.," *Curr. Opin. Microbiol.*, vol. 10, no. 2, pp. 102–9, Apr. 2007.

[26] Y. Eguchi, T. Itoh, and J. Tomizawa, "Antisense RNA.," *Annu. Rev. Biochem.*, vol. 60, pp. 631–52, Jan. 1991.

[27] M. J. Gubbins, D. C. Arthur, A. F. Ghetu, J. N. M. Glover, and L. S. Frost, "Characterizing the structural features of RNA/RNA interactions of the F-plasmid FinOP fertility inhibition system.," *J. Biol. Chem.*, vol. 278, no. 30, pp. 27663–71, Jul. 2003.

[28] J. A. Opdyke, J.-G. Kang, and G. Storz, "GadY, a Small-RNA Regulator of Acid Response Genes in *Escherichia coli*," *J. Bacteriol.*, vol. 186, no. 20, pp. 6698–6705, Oct. 2004.

[29] U. Dühring, I. M. Axmann, W. R. Hess, and A. Wilde, "An internal antisense RNA regulates expression of the photosynthesis gene *isiA*.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 18, pp. 7054–8, May 2006.

[30] K. Papenfort and J. Vogel, "Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level," *Res. Microbiol.*, vol. 160, no. 4, pp.

278–287, May 2009.

[31] R. A. Lease, M. E. Cusick, and M. Belfort, “Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 21, pp. 12456–61, Oct. 1998.

[32] D. Sledjeski and S. Gottesman, “A small RNA acts as an antisilencer of the H-NS-silenced *rcsA* gene of *Escherichia coli*.,” *Proc. Natl. Acad. Sci.*, vol. 92, no. 6, pp. 2003–2007, Mar. 1995.

[33] N. Majdalani, C. Cunning, D. Sledjeski, T. Elliott, and S. Gottesman, “DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 21, pp. 12462–7, Oct. 1998.

[34] D. D. Sledjeski, A. Gupta, and S. Gottesman, “The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in *Escherichia coli*.,” *EMBO J.*, vol. 15, no. 15, pp. 3993–4000, Aug. 1996.

[35] E. Massé and S. Gottesman, “A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 7, pp. 4620–5, Apr. 2002.

[36] E. Massé, C. K. Vanderpool, and S. Gottesman, “Effect of RyhB small RNA on global iron use in *Escherichia coli*.,” *J. Bacteriol.*, vol. 187, no. 20, pp. 6962–71, Oct. 2005.

[37] J. Vogel and K. Papenfort, “Small non-coding RNAs and the bacterial outer membrane,” *Curr. Opin. Microbiol.*, vol. 9, no. 6, pp. 605-611 2006.

[38] G. Storz, J. Vogel, and K. M. Wassarman, “Regulation by Small RNAs in Bacteria: Expanding Frontiers,” *Mol. Cell*, vol. 43, no. 6, pp. 880–891, 2011.

[39] M. T. Franze de Fernandez, L. Eoyang, and J. T. August, “Factor fraction required for the synthesis of bacteriophage Qbeta-RNA.,” *Nature*, vol. 219, no. 5154, pp. 588–90, Aug. 1968.

[40] B. Večerek, I. Moll, T. Afonyushkin, V. Kaberdin, and U. Bläsi, “Interaction of the RNA chaperone Hfq with mRNAs: direct and indirect roles of Hfq in iron metabolism of *Escherichia coli*,” *Mol. Microbiol.*, vol. 50, no. 3, pp. 897–909, Sep. 2003.

[41] C. T. Kåhrström, “Cellular microbiology: Ironing out Hfq regulation.,” *Nat. Rev. Microbiol.*, vol. 10, no. March, p. 2780, 2012.

[42] N. De Lay, D. J. Schu, and S. Gottesman, “Bacterial small RNA-based negative regulation: Hfq and its accomplices.,” *J. Biol. Chem.*, vol. 288, no. 12, pp. 7996–8003, Mar. 2013.

[43] T. M. Link, P. Valentin-Hansen, and R. G. Brennan, “Structure of *Escherichia coli* Hfq bound to polyriboadenylate RNA.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 46, pp. 19292–7, Nov. 2009.

[44] E. Sauer, S. Schmidt, and O. Weichenrieder, “Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 24, pp. 9396–401, Jun. 2012.

[45] P. J. Mikulecky, M. K. Kaw, C. C. Brescia, J. C. Takach, D. D. Sledjeski, and A. L. Feig, “*Escherichia coli* Hfq has distinct interaction surfaces for DsrA, RpoS and poly(A) RNAs.,” *Nat. Struct. Mol. Biol.*, vol. 11, no. 12, pp. 1206–14, Dec. 2004.

[46] C. Lorenz, T. Gesell, B. Zimmermann, U. Schoeberl, I. Bilusic, L. Rajkowitsch, C. Waldsich, A. von Haeseler, and R. Schroeder, “Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts,” *Nucleic Acids Res.*, vol. 38, no. 11, pp. 3794–3808, Mar. 2010.

[47] A. Fender, J. Elf, K. Hampel, B. Zimmermann, and E. G. H. Wagner, “RNAs actively cycle on the Sm-like protein Hfq.,” *Genes Dev.*, vol. 24, no. 23, pp. 2621–6, Dec. 2010.

[48] J. F. Hopkins, S. Panja, and S. A. Woodson, “Rapid binding and release of Hfq from ternary complexes during RNA annealing.,” *Nucleic Acids Res.*, vol. 39, no. 12, pp. 5193–202, Jul. 2011.

[49] W. Hwang, V. Arluison, and S. Hohng, “Dynamic competition of DsrA and RpoS fragments for the proximal binding site of Hfq as a means for efficient annealing.,”

Nucleic Acids Res., vol. 39, no. 12, pp. 5131–9, Jul. 2011.

[50] A. Zhang, D. J. Schu, B. C. Tjaden, G. Storz, and S. Gottesman, "Mutations in interaction surfaces differentially impact *E. coli* Hfq association with small RNAs and their mRNA targets.," *J. Mol. Biol.*, vol. 425, no. 19, pp. 3678–97, Oct. 2013.

[51] A. Carpousis, "Copurification of *E. coli* RNAase E and PNPase: Evidence for a specific association between two enzymes important in RNA processing and degradation," *Cell*, vol. 76, no. 5, pp. 889–900, Mar. 1994.

[52] A. Miczak, V. R. Kaberdin, C. L. Wei, and S. Lin-Chao, "Proteins associated with RNase E in a multicomponent ribonucleolytic complex.," *Proc. Natl. Acad. Sci.*, vol. 93, no. 9, pp. 3865–3869, Apr. 1996.

[53] B. Py, C. F. Higgins, H. M. Krisch, and A. J. Carpousis, "A DEAD-box RNA helicase in the *Escherichia coli* RNA degradosome," *Nature*, vol. 381, no. 6578, pp. 169–172, May 1996.

[54] S. R. Kushner, "Messenger RNA Decay," *EcoSal Plus*, vol. 1, no. 4, Dec. 2013.

[55] K. J. Bandyra, M. Bouvier, A. J. Carpousis, and B. F. Luisi, "The social fabric of the RNA degradosome.," *Biochim. Biophys. Acta*, vol. 1829, no. 6–7, pp. 514–522, 2013.

[56] F. J. Iborra, D. A. Jackson, and P. R. Cook, "Coupled transcription and translation within nuclei of mammalian cells.," *Science*, vol. 293, no. 5532, pp. 1139–42, Aug. 2001.

[57] J. Gowrishankar and R. Harinarayanan, "Why is transcription coupled to translation in bacteria?," *Mol. Microbiol.*, vol. 54, no. 3, pp. 598–603, Sep. 2004.

[58] S. Proshkin, A. R. Rahmouni, A. Mironov, and E. Nudler, "Cooperation between translating ribosomes and RNA polymerase in transcription elongation.," *Science*, vol. 328, no. 5977, pp. 504–8, Apr. 2010.

[59] D. Lalaouna, M. Simoneau-Roy, D. Lafontaine, and E. Massé, "Regulatory RNAs and target mRNA decay in prokaryotes," *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1829, no. 6–7. pp. 742–747, 2013.

[60] F. Repoila and F. Darfeuille, “Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects,” *Biol. Cell*, 2009.

[61] J. M. Andrade and C. M. Arraiano, “PNPase is a key player in the regulation of small RNAs that control the expression of outer membrane proteins,” *RNA*, vol. 14, no. 3, pp. 543–551, Jan. 2008.

[62] J. M. Andrade, V. Pobre, A. M. Matos, and C. M. Arraiano, “The crucial role of PNPase in the degradation of small RNAs that are not associated with Hfq,” *RNA*, vol. 18, no. 4. pp. 844–855, 2012.

[63] S. P. Pandey, J. a Winkler, H. Li, D. M. Camacho, J. J. Collins, and G. C. Walker, “Central role for RNase YbeY in Hfq-dependent and Hfq-independent small-RNA regulation in bacteria.,” *BMC Genomics*, vol. 15, p. 121, 2014.

[64] L. Argaman, M. Elgrably-Weiss, T. Hershko, J. Vogel, and S. Altuvia, “RelA protein stimulates the activity of RyhB small RNA by acting on RNA-binding protein Hfq.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 12, pp. 4621–6, Mar. 2012.

[65] K. J. Bandyra, N. Said, V. Pfeiffer, M. W. Górna, J. Vogel, and B. F. Luisi, “The Seed Region of a Small RNA Drives the Controlled Destruction of the Target mRNA by the Endoribonuclease RNase E,” *Mol. Cell*, vol. 47, no. 6, pp. 943–953, 2012.

[66] D. J. Luciano, M. P. Hui, A. Deana, P. L. Foley, K. J. Belasco, and J. G. Belasco, “Differential control of the rate of 5’-end-dependent mRNA degradation in *Escherichia coli*.,” *J. Bacteriol.*, vol. 194, no. 22, pp. 6233–9, Nov. 2012.

[67] S. Gottesman and G. Storz, “Bacterial Small RNA Regulators: Versatile Roles and Rapidly Evolving Variations,” *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 12. pp. a003798–a003798, 2011.

[68] J. Vogel and E. G. H. Wagner, “Target identification of small noncoding RNAs in bacteria.,” *Curr. Opin. Microbiol.*, vol. 10, no. 3, pp. 262–70, Jun. 2007.

[69] H. Aiba, “Mechanism of RNA silencing by Hfq-binding small RNAs,” *Current Opinion in Microbiology*, vol. 10, no. 2. pp. 134–139, 2007.

[70] N. Delihas and S. Forst, “MicF: an antisense RNA gene involved in response of *Escherichia coli* to global stress factors.,” *J. Mol. Biol.*, vol. 313, no. 1, pp. 1–12, Oct. 2001.

[71] J. L. Rosner and G. Storz, “Effects of peroxides on susceptibilities of *Escherichia coli* and Mycobacterium smegmatis to isoniazid.,” *Antimicrob. Agents Chemother.*, vol. 38, no. 8, pp. 1829–33, Aug. 1994.

[72] R. A. Lease, M. E. Cusick, and M. Belfort, “Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci,” *Proc. Natl. Acad. Sci.*, vol. 95, no. 21, pp. 12456–12461, Oct. 1998.

[73] B. L. Wanner, S. Wieder, and R. McSharry, “Use of bacteriophage transposon Mu d1 to determine the orientation for three proC-linked phosphate-starvation-inducible (psi) genes in *Escherichia coli* K-12.,” *J. Bacteriol.*, vol. 146, no. 1, pp. 93–101, Apr. 1981.

[74] S. Altuvia, “The *Escherichia coli* OxyS regulatory RNA represses fhlA translation by blocking ribosome binding,” *EMBO J.*, vol. 17, no. 20, pp. 6069–6075, Oct. 1998.

[75] M. Antal, V. Bordeau, V. Douchin, and B. Felden, “A small bacterial RNA regulates a putative ABC transporter.,” *J. Biol. Chem.*, vol. 280, no. 9, pp. 7901–8, Mar. 2005.

[76] J. Johansen, M. Eriksen, B. Kallipolitis, and P. Valentin-Hansen, “Down-regulation of outer membrane proteins by noncoding RNAs: unraveling the cAMP-CRP- and sigmaE-dependent CyaR-ompX regulatory case.,” *J. Mol. Biol.*, vol. 383, no. 1, pp. 1–9, Oct. 2008.

[77] V. Douchin, C. Bohn, and P. Bouloc, “Down-regulation of porins by a small RNA bypasses the essentiality of the regulated intramembrane proteolysis protease RseP in *Escherichia coli*.,” *J. Biol. Chem.*, vol. 281, no. 18, pp. 12253–9, May 2006.

[78] R. A. Lease, D. Smith, K. McDonough, and M. Belfort, “The small noncoding DsrA RNA is an acid resistance regulator in *Escherichia coli*.,” *J. Bacteriol.*, vol. 186, no. 18, pp. 6179–85, Sep. 2004.

[79] L. Guzman, D. Belin, M. Carson, and J. Beckwith, “Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter,” *J. Bacteriol.*,

vol. 177, no. 14, pp. 4121–4130, Jul. 1995.

[80] K. Papenfort, V. Pfeiffer, F. Mika, S. Lucchini, J. C. D. Hinton, and J. Vogel, "Sigma E - dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global omp mRNA decay," *Mol. Microbiol.*, vol. 62, no. 6, pp. 1674–1688, Dec. 2006.

[81] J. Johansen, A. A. Rasmussen, M. Overgaard, and P. Valentin-Hansen, "Conserved small non-coding RNAs that belong to the sigmaE regulon: role in down-regulation of outer membrane proteins.," *J. Mol. Biol.*, vol. 364, no. 1, pp. 1–8, Nov. 2006.

[82] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel, "The impact of microRNAs on protein output.," *Nature*, vol. 455, no. 7209, pp. 64–71, Sep. 2008.

[83] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky, "Widespread changes in protein synthesis induced by microRNAs.," *Nature*, vol. 455, no. 7209, pp. 58–63, Sep. 2008.

[84] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, Oct. 1990.

[85] W. Gerlach and R. Giegerich, "GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing," *Bioinformatics*, vol. 22, no. 6, pp. 762–764, Jan. 2006.

[86] B. Tjaden, S. S. Goodwin, J. A. Opdyke, M. Guillier, D. X. Fu, S. Gottesman, and G. Storz, "Target prediction for small, noncoding RNAs in bacteria.," *Nucleic Acids Res.*, vol. 34, no. 9, pp. 2791–2802, 2006.

[87] P. Mandin, F. Repoila, M. Vergassola, T. Geissmann, and P. Cossart, "Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets.," *Nucleic Acids Res.*, vol. 35, no. 3, pp. 962–74, Jan. 2007.

[88] W. R. Hess and A. Marchfelder, *Regulatory RNAs in Prokaryotes*. Vienna: Springer Vienna, 2012.

[89] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence

of thermodynamic parameters improves prediction of RNA secondary structure," *J. Mol. Biol.*, vol. 288, no. 5, pp. 911–940, May 1999.

[90] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich, "Fast and effective prediction of microRNA/target duplexes.," *RNA*, vol. 10, no. 10, pp. 1507–1517, 2004.

[91] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker, "The Vienna RNA websuite.," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. W70–4, Jul. 2008.

[92] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Partition function and base pairing probabilities of RNA heterodimers," *Algorithms Mol. Biol.*, vol. 1, no. 1, p. 3, 2006.

[93] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshefte for Chemie Chem. Mon.*, vol. 125, no. 2, pp. 167–188, 1994.

[94] R. Backofen and W. R. Hess, "Computational prediction of sRNAs and their targets in bacteria.," *RNA Biol.*, vol. 7, no. 1, pp. 33–42, 2010.

[95] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker, "Thermodynamics of RNA-RNA binding.," *Bioinformatics*, vol. 22, no. 10, pp. 1177–82, 2006.

[96] A. Busch, A. S. Richter, and R. Backofen, "IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions," *Bioinformatics*, vol. 24, no. 24, pp. 2849–2856, 2008.

[97] S. E. Seemann, A. S. Richter, T. Gesell, R. Backofen, and J. Gorodkin, "PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences.," *Bioinformatics*, vol. 27, no. 2, pp. 211–9, Jan. 2011.

[98] R. Backofen, F. Amman, F. Costa, S. Findeiß, A. S. Richter, and P. F. Stadler, "Bioinformatics of prokaryotic RNAs.," *RNA Biol.*, vol. 11, no. 5, pp. 470–83, Jan. 2014.

[99] P. R. Wright, A. S. Richter, K. Papenfort, M. Mann, J. Vogel, W. R. Hess, R. Backofen, and J. Georg, "Comparative genomics boosts target prediction for bacterial small RNAs.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 37, pp. E3487–96, 2013.

[100] F. Repoila and F. Darfeuille, "Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects.," *Biol. cell under auspices Eur. Cell Biol. Organ.*, vol. 101, no. 2, pp. 117–131, 2009.

[101] V. Dinçbas-Renqvist, A. Engström, L. Mora, V. Heurgué-Hamard, R. Buckingham, and M. Ehrenberg, "A post-translational modification in the GGQ motif of RF2 from *Escherichia coli* stimulates termination of translation.," *EMBO J.*, vol. 19, no. 24, pp. 6900–7, Dec. 2000.

[102] M. Y. Liu, G. Gui, B. Wei, J. F. Preston, L. Oakford, U. Yüksel, D. P. Giedroc, and T. Romeo, "The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*.," *J. Biol. Chem.*, vol. 272, no. 28, pp. 17502–17510, 1997.

[103] S. Gottesman, "The small RNA regulators of *Escherichia coli*: roles and mechanisms*.," *Annu. Rev. Microbiol.*, vol. 58, pp. 303–328, 2004.

[104] Y. Zhang, S. Sun, T. Wu, J. Wang, C. Liu, L. Chen, X. Zhu, Y. Zhao, Z. Zhang, B. Shi, H. Lu, and R. Chen, "Identifying Hfq-binding small RNA targets in *Escherichia coli*.," *Biochem. Biophys. Res. Commun.*, vol. 343, no. 3, pp. 950–955, 2006.

[105] J. Vogel and C. M. Sharma, "How to find small non-coding RNAs in bacteria.," *Biol. Chem.*, vol. 386, no. 12, pp. 1219–1238, 2005.

[106] C. M. Sharma and J. Vogel, "Experimental approaches for the discovery and characterization of regulatory small RNA.," *Curr. Opin. Microbiol.*, vol. 12, no. 5, pp. 536–46, Oct. 2009.

[107] C. Pichon and B. Felden, "Small RNA gene identification and mRNA target predictions in bacteria.," *Bioinformatics*, vol. 24, no. 24, pp. 2807–13, Dec. 2008.

[108] H. Tafer and I. L. Hofacker, "RNAplex: a fast tool for RNA-RNA interaction search.," *Bioinformatics*, vol. 24, no. 22, pp. 2657–63, Nov. 2008.

[109] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Partition function and base pairing probabilities of RNA heterodimers.," *Algorithms Mol. Biol.*, vol. 1, no. 1, p. 3, Jan. 2006.

[110] C. Alkan, E. Karakoç, J. H. Nadeau, S. C. Sahinalp, and K. Zhang, "RNA-RNA interaction prediction and antisense RNA target search.," *J. Comput. Biol.*, vol. 13, no. 2, pp. 267–282, 2006.

[111] M. Andronescu, Z. C. Zhang, and A. Condon, "Secondary structure prediction of interacting RNA molecules.," *J. Mol. Biol.*, vol. 345, no. 5, pp. 987–1001, Feb. 2005.

[112] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce, "Thermodynamic Analysis of Interacting Nucleic Acid Strands," *SIAM Rev.*, vol. 49, no. 1, pp. 65–88, Jan. 2007.

[113] M. B. Kery, M. Feldman, J. Livny, and B. Tjaden, "TargetRNA2: Identifying targets of small regulatory RNAs in bacteria," *Nucleic Acids Res.*, vol. 42, no. W1, 2014.

[114] P. R. Wright, J. Georg, M. Mann, D. A. Sorescu, A. S. Richter, S. Lott, R. Kleinkauf, W. R. Hess, and R. Backofen, "CopraRNA and IntaRNA: Predicting small RNA targets, networks and interaction domains," *Nucleic Acids Res.*, vol. 42, no. W1, 2014.

[115] S. C. Viegas, I. J. Silva, M. Saramago, S. Domingues, and C. M. Arraiano, "Regulation of the small regulatory RNA MicA by ribonuclease III: A target-dependent pathway," *Nucleic Acids Res.*, vol. 39, no. 7, pp. 2918–2930, 2011.

[116] S. P. Pandey, J. A. Winkler, H. Li, D. M. Camacho, J. J. Collins, and G. C. Walker, "Central role for RNase YbeY in Hfq-dependent and Hfq-independent small-RNA regulation in bacteria," *BMC Genomics*, vol. 15, no. 1, p. 121, 2014.

[117] P. Larranaga, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, Feb. 2006.

[118] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, and J. A. Lozano, "Machine learning: an indispensable tool in bioinformatics.," *Methods Mol. Biol.*, vol. 593, pp. 25–48, 2010.

[119] C. Fletez-Brant, D. Lee, A. S. McCallion, and M. A. Beer, "kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.," *Nucleic Acids Res.*, vol. 41, no. Web Server issue, pp. W544–56, Jul. 2013.

[120] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[121] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 493–507, 2012.

[122] Y. Qi, "Random Forest for Bioinformatics," *Ensemble Mach. Learn.*, pp. 307–323, 2012.

[123] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," *Discovery*, no. 1999, pp. 1–12, 2004.

[124] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[125] H. Salvail and E. Massé, "Regulating iron storage and metabolism with RNA: an overview of posttranscriptional controls of intracellular iron homeostasis.," *Wiley Interdiscip. Rev. RNA*, vol. 3, no. 1, pp. 26–36, Jan. .

[126] M. Guillier and S. Gottesman, "Remodelling of the *Escherichia coli* outer membrane by two small regulatory RNAs.," *Mol. Microbiol.*, vol. 59, no. 1, pp. 231–47, Jan. 2006.

[127] A. Zhang, K. M. Wassarman, C. Rosenow, B. C. Tjaden, G. Storz, and S. Gottesman, "Global analysis of small RNA and mRNA targets of Hfq.," *Mol. Microbiol.*, vol. 50, no. 4, pp. 1111–24, Nov. 2003.

[128] R. Hussein and H. N. Lim, "Disruption of small RNA signaling caused by competition for Hfq," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 3, pp. 1110–1115, 2011.

[129] J. C. Guimaraes, M. Rocha, and A. P. Arkin, "Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*.," *Nucleic Acids Res.*, vol. 42, no. 8, pp. 4791–9, Apr. 2014.

[130] C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, and D. Koller, "Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation.," *Mol. Syst. Biol.*, vol. 10, no. 12, p. 770, Jan. 2014.

[131] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp, "EcoCyc: a comprehensive database resource for *Escherichia coli*," *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D334–D337, 2005.

[132] Y. Cao, J. Wu, Q. Liu, Y. Zhao, X. Ying, L. Cha, L. Wang, and W. Li, "sRNATarBase: A comprehensive database of bacterial sRNA targets verified by experiments," *Rna New York Ny*, vol. 16, no. 11, pp. 2051–2057, 2010.

[133] L. Li, D. Huang, M. K. Cheung, W. Nong, Q. Huang, and H. S. Kwan, "BSRD: A repository for bacterial small regulatory RNA," *Nucleic Acids Res.*, vol. 41, no. D1, 2013.

[134] A. Cornish-Bowden, "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.," *Nucleic Acids Res.*, vol. 13, no. 9, pp. 3021–30, May 1985.

[135] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An Empirical Study of Learning from Imbalanced Data Using Random Forest," in *19th IEEE International Conference on Tools with Artificial IntelligenceICTAI 2007*, 2007, vol. 2, pp. 310–317.

[136] L. Li, D. Huang, M. K. Cheung, W. Nong, Q. Huang, and H. S. Kwan, "BSRD: a repository for bacterial small regulatory RNA.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D233–8, Jan. 2013.

[137] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.," *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Jan. 2009.

[138] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and A. F. T. Sacha van Hijum, "Data mining in the life science swith random forest: A walk in the park or lost in the jungle?," *Brief. Bioinform.*, vol. 14, no. 3, pp. 315–326, 2013.

[139] E. C. Hobbs, J. L. Astarita, and G. Storz, "Small RNAs and Small Proteins Involved in

Resistance to Cell Envelope Stress and Acid Shock in *Escherichia coli*: Analysis of a Bar-Coded Mutant Collection," *J. Bacteriol.*, vol. 192, no. 1, pp. 59–67, 2010.

[140] M. B. Stead, A. Agrawal, K. E. Bowden, R. Nasir, B. K. Mohanty, R. B. Meagher, and S. R. Kushner, "RNAsnap^{TM}: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria.," *Nucleic Acids Res.*, vol. 40, no. 20, p. e156, 2012.

[141] B. K. Mohanty, H. Giladi, V. F. Maples, and S. R. Kushner, *RNA Turnover in Bacteria, Archaea and Organelles*, vol. 447. Elsevier, 2008.

[142] W. V Burnett, "Northern blotting of RNA denatured in glyoxal without buffer recirculation.," *Biotechniques*, vol. 22, no. 4, pp. 668–71, Apr. 1997.

[143] S. Chen, A. Zhang, L. B. Blyn, and G. Storz, "MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli*.," *J. Bacteriol.*, vol. 186, no. 20, pp. 6689–97, Oct. 2004.

[144] F. Canonaco, T. A. Hess, S. Heri, T. Wang, T. Szyperski, and U. Sauer, "Metabolic flux response to phosphoglucose isomerase knock-out in *Escherichia coli* and impact of overexpression of the soluble transhydrogenase UdhA.," *FEMS Microbiol. Lett.*, vol. 204, no. 2, pp. 247–52, Nov. 2001.

[145] U. Sauer, "The Soluble and Membrane-bound Transhydrogenases UdhA and PntAB Have Divergent Functions in NADPH Metabolism of *Escherichia coli*," *J. Biol. Chem.*, vol. 279, no. 8, pp. 6613–6619, Nov. 2003.

[146] C. Bradbeer, J. S. Kenley, D. R. Di Masi, and M. Leighton, "Transport of vitamin B12 in *Escherichia coli*. Corrinoid specificities of the periplasmic B12-binding protein and of energy-dependent B12 transport.," *J. Biol. Chem.*, vol. 253, no. 5, pp. 1347–52, Mar. 1978.

[147] J. R. Roth, J. G. Lawrence, and T. A. Bobik, "Cobalamin (coenzyme B12): synthesis and biological significance.," *Annu. Rev. Microbiol.*, vol. 50, pp. 137–81, Jan. 1996.

[148] P. Mandin and S. Gottesman, "Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA.," *EMBO J.*, vol. 29, no. 18, pp. 3094–107, Sep. 2010.

[149] K. Papenfort, N. Said, T. Welsink, S. Lucchini, J. C. D. Hinton, and J. Vogel, "Specific and pleiotropic patterns of mRNA regulation by ArcZ, a conserved, Hfq-dependent small RNA.," *Mol. Microbiol.*, vol. 74, no. 1, pp. 139–58, Oct. 2009.

[150] N. Figueroa-Bossi, M. Valentini, L. Malleret, F. Fiorini, and L. Bossi, "Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target.," *Genes Dev.*, vol. 23, no. 17, pp. 2004–15, Sep. 2009.

[151] M. Overgaard, J. Johansen, J. Møller-Jensen, and P. Valentin-Hansen, "Switching off small RNA regulation with trap-mRNA.," *Mol. Microbiol.*, vol. 73, no. 5, pp. 790–800, Sep. 2009.

[152] A. A. Rasmussen, J. Johansen, J. S. Nielsen, M. Overgaard, B. Kallipolitis, and P. Valentin-Hansen, "A conserved small RNA promotes silencing of the outer membrane protein YbfM.," *Mol. Microbiol.*, vol. 72, no. 3, pp. 566–77, May 2009.

[153] P. Mandin and S. Gottesman, "A genetic approach for finding small RNAs regulators of genes of interest identifies RybC as regulating the DpiA/DpiB two-component system.," *Mol. Microbiol.*, vol. 72, no. 3, pp. 551–65, May 2009.

[154] N. Figueroa-Bossi, M. Valentini, L. Malleret, F. Fiorini, and L. Bossi, "Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target.," *Genes Dev.*, vol. 23, no. 17, pp. 2004–15, Sep. 2009.

[155] N. De Lay and S. Gottesman, "The Crp-Activated Small Noncoding Regulatory RNA CyaR (RyeE) Links Nutritional Status to Group Behavior," *J. Bacteriol.*, vol. 191, no. 2, pp. 461–476, 2009.

[156] F. Tétart and J. P. Bouché, "Regulation of the expression of the cell-cycle gene ftsZ by DicF antisense RNA. Division does not require a fixed number of FtsZ molecules.," *Mol. Microbiol.*, vol. 6, no. 5, pp. 615–20, Mar. 1992.

[157] A. Boysen, J. Møller-Jensen, B. Kallipolitis, P. Valentin-Hansen, and M. Overgaard, "Translational regulation of gene expression by an anaerobically induced small non-coding RNA in *Escherichia coli*.," *J. Biol. Chem.*, vol. 285, no. 14, pp. 10690–702, Apr. 2010.

[158] S. Durand and G. Storz, “Reprogramming of anaerobic metabolism by the FnrS small RNA.,” *Mol. Microbiol.*, vol. 75, no. 5, pp. 1215–31, Mar. 2010.

[159] A. Boysen, J. Møller-Jensen, B. Kallipolitis, P. Valentin-Hansen, and M. Overgaard, “Translational regulation of gene expression by an anaerobically induced small non-coding RNA in *Escherichia coli*.,” *J. Biol. Chem.*, vol. 285, no. 14, pp. 10690–702, Apr. 2010.

[160] C. M. Sharma, F. Darfeuille, T. H. Plantinga, and J. Vogel, “A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites.,” *Genes Dev.*, vol. 21, no. 21, pp. 2804–17, Nov. 2007.

[161] C. M. Sharma, K. Papenfort, S. R. Pernitzsch, H.-J. Mollenkopf, J. C. D. Hinton, and J. Vogel, “Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA.,” *Mol. Microbiol.*, vol. 81, no. 5, pp. 1144–1165, 2011.

[162] M. G. Jørgensen, J. S. Nielsen, A. Boysen, T. Franch, J. Møller-Jensen, and P. Valentin-Hansen, “Small regulatory RNAs control the multi-cellular adhesive lifestyle of *Escherichia coli*.,” *Mol. Microbiol.*, vol. 84, no. 1, pp. 36–50, Apr. 2012.

[163] S. C. Pulvermacher, L. T. Stauffer, and G. V Stauffer, “Role of the sRNA GcvB in regulation of cycA in *Escherichia coli*.,” *Microbiology*, vol. 155, no. Pt 1, pp. 106–14, Jan. 2009.

[164] S. C. Pulvermacher, L. T. Stauffer, and G. V Stauffer, “The small RNA GcvB regulates sstT mRNA expression in *Escherichia coli*.,” *J. Bacteriol.*, vol. 191, no. 1, pp. 238–48, Jan. 2009.

[165] C. M. Sharma, F. Darfeuille, T. H. Plantinga, and J. Vogel, “A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites.,” *Genes Dev.*, vol. 21, no. 21, pp. 2804–17, Nov. 2007.

[166] J. H. Urban and J. Vogel, “Two seemingly homologous noncoding RNAs act hierarchically to activate *glmS* mRNA translation.,” *PLoS Biol.*, vol. 6, no. 3, p. e64, Mar. 2008.

[167] V. Pfeiffer, A. Sittka, R. Tomer, K. Tedin, V. Brinkmann, and J. Vogel, “A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the Salmonella core genome.,” *Mol. Microbiol.*, vol. 66, no. 5, pp. 1174–91, Dec. 2007.

[168] J. Vogel, L. Argaman, E. G. H. Wagner, and S. Altuvia, “The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide.,” *Curr. Biol.*, vol. 14, no. 24, pp. 2271–6, Dec. 2004.

[169] E. B. Gogol, V. A. Rhodius, K. Papenfort, J. Vogel, and C. A. Gross, “Small RNAs endow a transcriptional activator with essential repressor functions for single-tier control of a global stress regulon.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 31, pp. 12875–80, Aug. 2011.

[170] L. Bossi and N. Figueroa-Bossi, “A small RNA downregulates LamB maltoporin in *Salmonella.*,” *Mol. Microbiol.*, vol. 65, no. 3, pp. 799–810, Aug. 2007.

[171] K. I. Udekwu, F. Darfeuille, J. Vogel, J. Reimegård, E. Holmqvist, and E. G. H. Wagner, “Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA.,” *Genes Dev.*, vol. 19, no. 19, pp. 2355–66, Oct. 2005.

[172] A. Coornaert, A. Lu, P. Mandin, M. Springer, S. Gottesman, and M. Guillier, “MicA sRNA links the PhoP regulon to cell envelope stress.,” *Mol. Microbiol.*, vol. 76, no. 2, pp. 467–79, Apr. 2010.

[173] V. Pfeiffer, K. Papenfort, S. Lucchini, J. C. D. Hinton, and J. Vogel, “Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation.,” *Nat. Struct. Mol. Biol.*, vol. 16, no. 8, pp. 840–6, Aug. 2009.

[174] E. Holmqvist, C. Unoson, J. Reimegård, and E. G. H. Wagner, “A mixed double negative feedback loop between the sRNA MicF and the global regulator Lrp.,” *Mol. Microbiol.*, vol. 84, no. 3, pp. 414–27, May 2012.

[175] C. P. Corcoran, D. Podkaminski, K. Papenfort, J. H. Urban, J. C. D. Hinton, and J. Vogel, “Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA.,” *Mol. Microbiol.*, vol. 84, no. 3, pp. 428–45, May 2012.

[176] M. Guillier and S. Gottesman, “The 5’ end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator.,” *Nucleic Acids Res.*, vol. 36, no. 21, pp. 6781–94, Dec. 2008.

[177] E. Holmqvist, J. Reimegård, M. Sterk, N. Grantcharova, U. Römling, and E. G. H. Wagner, “Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis.,” *EMBO J.*, vol. 29, no. 11, pp. 1840–50, Jun. 2010.

[178] B. Tjaden, S. S. Goodwin, J. A. Opdyke, M. Guillier, D. X. Fu, S. Gottesman, and G. Storz, “Target prediction for small, noncoding RNAs in bacteria.,” *Nucleic Acids Res.*, vol. 34, no. 9, pp. 2791–802, Jan. 2006.

[179] A. Zhang, S. Altuvia, A. Tiwari, L. Argaman, R. Hengge-Aronis, and G. Storz, “The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein.,” *EMBO J.*, vol. 17, no. 20, pp. 6061–8, Oct. 1998.

[180] F. Mika, S. Busse, A. Possling, J. Berkholz, N. Tschowri, N. Sommerfeldt, M. Pruteanu, and R. Hengge, “Targeting of csgD by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in *Escherichia coli*.,” *Mol. Microbiol.*, vol. 84, no. 1, pp. 51–65, Apr. 2012.

[181] N. Majdalani, D. Hernandez, and S. Gottesman, “Regulation and mode of action of the second small RNA activator of RpoS translation, RprA.,” *Mol. Microbiol.*, vol. 46, no. 3, pp. 813–26, Nov. 2002.

[182] K. Papenfort, M. Bouvier, F. Mika, C. M. Sharma, and J. Vogel, “Evidence for an autonomous 5’ target recognition domain in an Hfq-associated small RNA.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 47, pp. 20435–40, Nov. 2010.

[183] M. Bouvier, C. M. Sharma, F. Mika, K. H. Nierhaus, and J. Vogel, “Small RNA binding to 5’ mRNA coding region inhibits translational initiation.,” *Mol. Cell*, vol. 32, no. 6, pp. 827–37, Dec. 2008.

[184] G. Desnoyers and E. Massé, “Noncanonical repression of translation initiation through small RNA recruitment of the RNA chaperone Hfq.,” *Genes Dev.*, vol. 26, no. 7, pp. 726–39, Apr. 2012.

[185] H. Salvail, P. Lanthier-Bourbonnais, J. M. Sobota, M. Caza, J.-A. M. Benjamin, M. E. S. Mendieta, F. Lépine, C. M. Dozois, J. Imlay, and E. Massé, “A small RNA promotes

siderophore production through transcriptional and metabolic remodeling.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 34, pp. 15223–8, Aug. 2010.

[186] B. Vecerek, I. Moll, and U. Bläsi, "Control of Fur synthesis by the non-coding RNA RyhB and iron-responsive decoding.," *EMBO J.*, vol. 26, no. 4, pp. 965–75, Feb. 2007.

[187] G. Desnoyers, A. Morissette, K. Prévost, and E. Massé, "Small RNA-induced differential degradation of the polycistronic mRNA iscRSUA.," *EMBO J.*, vol. 28, no. 11, pp. 1551–61, Jun. 2009.

[188] K. Prévost, H. Salvail, G. Desnoyers, J.-F. Jacques, E. Phaneuf, and E. Massé, "The small RNA RyhB activates the translation of shiA mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis.," *Mol. Microbiol.*, vol. 64, no. 5, pp. 1260–73, Jun. 2007.

[189] J. B. Rice and C. K. Vanderpool, "The small RNA SgrS controls sugar-phosphate accumulation by regulating multiple PTS genes.," *Nucleic Acids Res.*, vol. 39, no. 9, pp. 3806–19, May 2011.

[190] H. Kawamoto, Y. Koide, T. Morita, and H. Aiba, "Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq.," *Mol. Microbiol.*, vol. 61, no. 4, pp. 1013–22, Aug. 2006.

[191] K. Papenfort, D. Podkaminski, J. C. D. Hinton, and J. Vogel, "The ancestral SgrS RNA discriminates horizontally acquired Salmonella mRNAs through a single G-U wobble pair.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 13, pp. E757–64, Mar. 2012.

[192] K. Papenfort, Y. Sun, M. Miyakoshi, C. K. Vanderpool, and J. Vogel, "Small RNA-mediated activation of sugar phosphatase mRNA regulates glucose homeostasis.," *Cell*, vol. 153, no. 2, pp. 426–37, Apr. 2013.

[193] C. L. Beisel and G. Storz, "The base-pairing RNA spot 42 participates in a multioutput feedforward loop to help enact catabolite repression in *Escherichia coli*.," *Mol. Cell*, vol. 41, no. 3, pp. 286–97, Feb. 2011.

[194] T. Møller, T. Franch, C. Udesen, K. Gerdes, and P. Valentin-Hansen, "Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon.," *Genes Dev.*, vol. 16, no. 13, pp. 1696–706, Jul. 2002.

[195] K. Papenfort, V. Pfeiffer, S. Lucchini, A. Sonawane, J. C. D. Hinton, and J. Vogel, "Systematic deletion of Salmonella small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis.," *Mol. Microbiol.*, vol. 68, no. 4, pp. 890–906, May 2008.

[196] J. H. Urban and J. Vogel, "Translational control and target recognition by *Escherichia coli* small RNAs in vivo.," *Nucleic Acids Res.*, vol. 35, no. 3, pp. 1018–37, Jan. 2007.

[197] Y.-Y. Lee, H.-T. Hu, P.-H. Liang, and K.-F. Chak, "An *E. coli* lon mutant conferring partial resistance to colicin may reveal a novel role in regulating proteins involved in the translocation of colicin.," *Biochem. Biophys. Res. Commun.*, vol. 345, no. 4, pp. 1579–85, Jul. 2006.

[198] T. Møller, T. Franch, C. Udesen, K. Gerdes, and P. Valentin-Hansen, "Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon.," *Genes Dev.*, vol. 16, no. 13, pp. 1696–706, Jul. 2002.

[199] K. Ono, K. Kutsukake, and T. Abo, "Suppression by enhanced RpoE activity of the temperature-sensitive phenotype of a degP ssrA double mutant in *Escherichia coli*.," *Genes Genet. Syst.*, vol. 84, no. 1, pp. 15–24, Feb. 2009.

APPENDIX

SUPPLEMENTARY DATA FOR CHAPTER 4

SUPPLEMENTARY S1: SRNA-MRNA PAIRS SELECTED FOR THE STUDY

Table S1: The table below lists all pairs of sRNAs and mRNAs tested for interactions, as reported in literature. The entry in the column labeled "Interaction" was taken as the class label for the algorithm. References to the interactions are listed below the table.

sRNA	mRNA	Locus ID	Species	Interaction	Reference
ArcZ	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1701	Yes	[148]
ArcZ	*sdaC*	stm2970	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[149]
ArcZ	*stm3216*	stm3216	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[149]
ArcZ	*tpx*	stm1682	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[149]
ChiX	*celB*	stm1313	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	Yes	[150]
ChiX	*chbC*	b1737	*Escherichia coli* K-12 substr. MG1690	Yes	[151]
ChiX	*chiP*	b0681	*Escherichia coli str.* K-12 substr. MG1658	Yes	[152]
ChiX	*dpiB*	b0619	*Escherichia coli str.* K-12 substr. MG1657	Yes	[153]
ChiX	*ybfM*	stm0687	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[154]
CyaR	*luxS*	b2687	*Escherichia coli str.* K-12 substr. MG1655	Yes	[155]
CyaR	*nadE*	b1740	*Escherichia coli str.* K-12 substr. MG1655	Yes	[155]
CyaR	*ompX*	b0814	*Escherichia coli str.* K-12 substr. MG1655	Yes	[155]
CyaR	*sdhA*	b0723	*Escherichia coli str.* K-12 substr. MG1663	Yes	[99]
CyaR	*yqaE*	b2666	*Escherichia coli str.* K-12 substr. MG1655	Yes	[155]
DicF	*ftsZ*	b0095	*Escherichia coli str.* K-12 substr. MG1655	Yes	[156]
DsrA	*argR*	b3237	*Escherichia coli str.* K-12 substr. MG1655	Yes	[31]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
DsrA	*hns*	b1237	*Escherichia coli str.* K-12 substr. MG1655	Yes	[31]
DsrA	*ilvL*	b0077	*Escherichia coli str.* K-12 substr. MG1655	Yes	[31]
DsrA	*rbsD*	b3748	*Escherichia coli str.* K-12 substr. MG1655	Yes	[31]
DsrA	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	Yes	[33]
FnrS	*cydD*	b0887	*Escherichia coli str.* K-12 substr. MG1655	Yes	[157]
FnrS	*folE*	b2153	*Escherichia coli str.* K-12 substr. MG1694	Yes	[158]
FnrS	*folX*	b2303	*Escherichia coli str.* K-12 substr. MG1696	Yes	[158]
FnrS	*gpmA*	b0755	*Escherichia coli str.* K-12 substr. MG1669	Yes	[158]
FnrS	*iscR*	b2531	*Escherichia coli str.* K-12 substr. MG1698	Yes	[99]
FnrS	*maeA*	b1479	*Escherichia coli str.* K-12 substr. MG1683	Yes	[158]
FnrS	*marA*	b1531	*Escherichia coli str.* K-12 substr. MG1685	Yes	[99]
FnrS	*metE*	b3829	*Escherichia coli str.* K-12 substr. MG1655	Yes	[157]
FnrS	*nagZ*	b1107	*Escherichia coli str.* K-12 substr. MG1678	Yes	[99]
FnrS	*sdhA*	b0723	*Escherichia coli str.* K-12 substr. MG1664	Yes	[99]
FnrS	*sodA*	b3908	*Escherichia coli str.* K-12 substr. MG1655	Yes	[157]
FnrS	*sodB*	b1656	*Escherichia coli str.* K-12 substr. MG1655	Yes	[157]
FnrS	*yobA*	b1841	*Escherichia coli str.* K-12 substr. MG1693	Yes	[159]
GcvB	*argT*	stm2355	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*brnQ*	stm0399	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*csgD*	b1040	*Escherichia coli str.* K-12 substr. MG1674	Yes	[162]
GcvB	*cycA*	b4208	*Escherichia coli str.* K-12 substr. MG1655	Yes	[163]
GcvB	*dppA*	stm3630	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*gdhA*	stm1299	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*gltL*	stm0665	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*iciA*	stm3064	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*ilvC*	stm3909	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*ilvE*	stm3903	*Salmonella enterica subsp. enterica*	Yes	[161]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			serovar Typhimurium str. LT2		
GcvB	*livJ*	stm3567	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*livK*	stm3564	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*lrp*	stm0959	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*metQ*	stm0245	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*ndk*	stm2526	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*ompR*	stm3502	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*oppA*	stm1746.s	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*serA*	stm3062	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*sstT*	b3089	*Escherichia coli str. K-12 substr. MG1655*	Yes	[164]
GcvB	*stm4351*	stm4351	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[160]
GcvB	*thrL/thrA*	stm0001	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[165]
GcvB	*tppB*	stm1452	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*ybdH*	stm0602	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[161]
GcvB	*ygjU*	stm3225	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	Yes	[161]
GlmY	*glmS*	b3729	*Escherichia coli str.* K-12 substr. MG1655	Yes	[166]
GlmZ	*glmS*	b3729	*Escherichia coli str.* K-12 substr. MG1655	Yes	[166]
InvR	*nmpC*	stm1572	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[167]
IstR	*tisB*	b4618	*Escherichia coli str*. K-12 substr. MG1655	Yes	[168]
MicA	*ecnB*	b4411	*Escherichia coli str.* K-12 substr. MG1715	Yes	[169]
MicA	*fimB*	b4312	*Escherichia coli str.* K-12 substr. MG1713	Yes	[169]
MicA	*gloA*	b1651	*Escherichia coli str*. K-12 substr. MG1689	Yes	[169]
MicA	*lamB*	stm4231	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[170]
MicA	*lpxT*	b2174	*Escherichia coli str.* K-12 substr. MG1695	Yes	[169]
MicA	*ompA*	stm1070	*Salmonella enterica subsp. enterica serovar Typhi str.* CT18	Yes	[170]
MicA	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1673	Yes	[171]
MicA	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[76]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
MicA	*ompX*	b0814	*Escherichia coli str.* K-12 substr. MG1671	Yes	[169]
MicA	*pal*	b0741	*Escherichia coli str.* K-12 substr. MG1667	Yes	[169]
MicA	*phoP*	b1130	*Escherichia coli str.* K-12 substr. MG1655	Yes	[172]
MicA	*tsx*	b0411	*Escherichia coli str.* K-12 substr. MG1655	Yes	[169]
MicA	*ybgF*	b0742	*Escherichia coli str.* K-12 substr. MG1668	Yes	[169]
MicA	*ycfS*	b1113	*Escherichia coli str.* K-12 substr. MG1680	Yes	[169]
MicC	*nmpC*	stm1572	*Salmonella enterica subsp. enterica serovar Typhimurium str*. LT2	Yes	[173]
MicC	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	Yes	[143]
MicF	*cpxR*	b3912	*Escherichia coli str*. K-12 substr. MG1709	Yes	[174]
MicF	*lpxR*	stm1328	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[175]
MicF	*lrp*	b0889	*Escherichia coli str.* K-12 substr. MG1672	Yes	[174]
MicF	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	Yes	[143]
MicF	*phoE*	b0241	*Escherichia coli str.* K-12 substr. MG1655	Yes	[174]
MicF	*yahO*	stm0366	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[175]
OmrA	*btuB*	b3966	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]
OmrA	*cirA*	b2155	*Escherichia coli O127:H6 str.* E2348/69	Yes	[176]
OmrA	*csgD*	b1040	*Escherichia coli str.* K-12 substr. MG1655	Yes	[177]
OmrA	*fecA*	b4291	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]
OmrA	*fecD*	b4288	*Escherichia coli str.* K-12 substr. MG1655	Yes	[178]
OmrA	*fepA*	b0584	*Escherichia coli O127:H6 str.* E2348/69	Yes	[126]
OmrA	*folP*	b3177	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]
OmrA	*glmM*	b3176	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]
OmrA	*gntP*	b4321	*Escherichia coli str*. K-12 substr. MG1655	Yes	[126]
OmrA	*ompR*	b3405	*Escherichia coli O127:H6 str.* E2348/69	Yes	[176]
OmrA	*ompT*	b0565	*Escherichia coli O127:H6 str.* E2348/69	Yes	[176]
OmrB	*cirA*	b2155	*Escherichia coli O127:H6 str.* E2348/69	Yes	[176]
OmrB	*csgD*	b1040	*Escherichia coli str.* K-12 substr. MG1655	Yes	[177]
OmrB	*fecA*	b4291	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
OmrB	*folP*	b3177	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]
OmrB	*glmM*	b3176	*Escherichia coli str.* K-12 substr. MG1655	Yes	[126]
OmrB	*gntP*	b4321	*Escherichia coli str.* K-12 substr. MG1655	Yes	[178]
OmrB	*ompR*	b3405	*Escherichia coli O127:H6 str.* E2348/69	Yes	[176]
OmrB	*ompT*	b0565	*Escherichia coli O127:H6 str.* E2348/69	Yes	[126]
OxyS	*fhlA*	b2731	*Escherichia coli str.* K-12 substr. MG1655	Yes	[74]
OxyS	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	Yes	[179]
RprA	*csgD*	b1040	*Escherichia coli str.* K-12 substr. MG1675	Yes	[180]
RprA	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	Yes	[181]
RprA	*ydaM*	b1341	*Escherichia coli str.* K-12 substr. MG1682	Yes	[180]
RseX	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	Yes	[77]
RseX	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	Yes	[77]
RybB	*asr*	b1597	*Escherichia coli str.* K-12 substr. MG1687	Yes	[169]
RybB	*chiP*	stm0687	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]
RybB	*fadL*	stm2391	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]
RybB	*fimA*	b4314	*Escherichia coli str.* K-12 substr. MG1714	Yes	[169]
RybB	*fiu*	b0805	*Escherichia coli str.* K-12 substr. MG1670	Yes	[169]
RybB	*fumC*	b1611	*Escherichia coli str.* K-12 substr. MG1688	Yes	[169]
RybB	*hinT*	b1103	*Escherichia coli str.* K-12 substr. MG1676	Yes	[169]
RybB	*mraZ*	b0081	*Escherichia coli str.* K-12 substr. MG1655	Yes	[99]
RybB	*nmpC*	b0553	*Escherichia coli str.* K-12 substr. MG1656	Yes	[169]
RybB	*ompA*	stm1070	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]
RybB	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	Yes	[81]
RybB	*ompD*	stm1572	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]
RybB	*ompF*	stm0999	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]
RybB	*ompN*	stm1473	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[183]
RybB	*ompS*	stm1995	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
RybB	*ompW*	b1256	*Escherichia coli str.* K-12 substr. MG1655	Yes	[81]
RybB	*rbsB*	b3751	*Escherichia coli str.* K-12 substr. MG1707	Yes	[169]
RybB	*rbsK*	b3752	*Escherichia coli str.* K-12 substr. MG1708	Yes	[169]
RybB	*rluD*	b2594	*Escherichia coli str.* K-12 substr. MG1699	Yes	[169]
RybB	*rraB*	b4255	*Escherichia coli str.* K-12 substr. MG1711	Yes	[169]
RybB	*sdhC*	b0721	*Escherichia coli str.* K-12 substr. MG1660	Yes	[184]
RybB	*tsx*	stm0413	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[182]
RybB	*ycfL*	b1104	*Escherichia coli str.* K-12 substr. MG1677	Yes	[169]
RybB	*ydeN*	b1498	*Escherichia coli str.* K-12 substr. MG1684	Yes	[169]
RybB	*yhjJ*	b3527	*Escherichia coli str.* K-12 substr. MG1704	Yes	[169]
RydC	*yejA*	b2177	*Escherichia coli str.* K-12 substr. MG1655	Yes	[75]
RyhB	*acnA*	b1276	*Escherichia coli str.* K-12 substr. MG1655	Yes	[35]
RyhB	*bfr*	b3336	*Escherichia coli str.* K-12 substr. MG1655	Yes	[35]
RyhB	*cysE*	b3607	*Escherichia coli str.* K-12 substr. MG1706	Yes	[185]
RyhB	*erpA*	b0156	*Escherichia coli str.* K-12 substr. MG1655	Yes	[99]
RyhB	*ftn*	b1905	*Escherichia coli str.* K-12 substr. MG1655	Yes	[36]
RyhB	*fumA*	b1612	*Escherichia coli str.* K-12 substr. MG1655	Yes	[35]
RyhB	*fur*	b0683	*Escherichia coli str.* K-12 substr. MG1655	Yes	[186]
RyhB	*iscS*	b2530	*Escherichia coli str.* K-12 substr. MG1655	Yes	[187]
RyhB	*marA*	b1531	*Escherichia coli str.* K-12 substr. MG1686	Yes	[99]
RyhB	*nagZ*	b1107	*Escherichia coli str.* K-12 substr. MG1679	Yes	[99]
RyhB	*nirB*	b3365	*Escherichia coli str.* K-12 substr. MG1703	Yes	[99]
RyhB	*sdhA*	b0723	*Escherichia coli str.* K-12 substr. MG1665	Yes	[99]
RyhB	*sdhC*	b0721	*Escherichia coli str.* K-12 substr. MG1661	Yes	[184]
RyhB	*sdhD*	b0722	*Escherichia coli str.* K-12 substr. MG1655	Yes	[35]
RyhB	*shiA*	b1981	*Escherichia coli str.* K-12 substr. MG1655	Yes	[188]
RyhB	*sodB*	b1656	*Escherichia coli str.* K-12 substr.	Yes	[178]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			MG1655		
SgrS	*manX*	b1817	*Escherichia coli str.* K-12 substr. MG1692	Yes	[189]
SgrS	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	Yes	[190]
SgrS	*ptsL*	b2416	*Escherichia coli str.* K-12 substr. MG1697	Yes	[99]
SgrS	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	Yes	[127]
SgrS	*sopD*	stm2945	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[191]
SgrS	*yigL*	stm3962	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	Yes	[192]
Spot42	*fucL*	b2802	*Escherichia coli str.* K-12 substr. MG1702	Yes	[193]
Spot42	*galK*	b0757	*Escherichia coli O127:H6 str.* E2348/69	Yes	[194]
Spot42	*gdhA*	b1761	*Escherichia coli str.* K-12 substr. MG1691	Yes	[99]
Spot42	*gltA*	b0720	*Escherichia coli str.* K-12 substr. MG1659	Yes	[193]
Spot42	*icd*	b1136	*Escherichia coli str.* K-12 substr. MG1681	Yes	[99]
Spot42	*nanC*	b4311	*Escherichia coli str.* K-12 substr. MG1712	Yes	[193]
Spot42	*sdhC*	b0721	*Escherichia coli str.* K-12 substr. MG1662	Yes	[184]
Spot42	*slrA*	b2702	*Escherichia coli str.* K-12 substr. MG1700	Yes	[193]
Spot42	*sthA*	b3962	*Escherichia coli str.* K-12 substr. MG1710	Yes	[193]
Spot42	*sucC*	b0728	*Escherichia coli str.* K-12 substr. MG1666	Yes	[99]
Spot42	*xylF*	b3566	*Escherichia coli str.* K-12 substr. MG1705	Yes	[193]
ArcZ	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	No	[148]
ChiX	*opgG*	b1048	*Escherichia coli str.* K-12 substr. MG1655	No	[99]
ChiX	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	No	[148]
CsrB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
CsrC	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
cyaR	*phoP*	b1130	*Escherichia coli str.* K-12 substr. MG1655	No	[172]
cyaR	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	No	[148]
DicF	*rpoS*	b2741	*Escherichia coli str.* K-12 substr. MG1655	No	[148]
DsrA	*galK*	b0757	*Escherichia coli str.* K-12 substr. MG1655	No	[196]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
DsrA	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
DsrA	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
DsrA	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
DsrA	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
DsrA	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
FnrS	*narK*	b1223	*Escherichia coli str.* K-12 substr. MG1655	No	[157]
GcvB	*dksA*	b0145	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GcvB	*hns*	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GcvB	*mltC*	stm3112	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[161]
GcvB	*mraZ*	b0081	*Escherichia coli str.* K-12 substr. MG1655	No	[99]
GcvB	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GcvB	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GcvB	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GcvB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
GcvB	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GcvB	*sodB*	b1656	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
GlmY	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
GlmZ	*mraZ*	b0081	*Escherichia coli str.* K-12 substr. MG1655	No	[99]
GlmZ	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	No	[195]
InvR	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[167]
InvR	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
IstR	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
MicA	*asr*	b1597	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*dppA*	b3544	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*fadL*	b2344	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*fimA*	b4314	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*fiuL*	b0805	*Escherichia coli str.* K-12 substr.	No	[169]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			MG1655		
MicA	*ftsB*	b2748	*Escherichia coli str.* K-12 substr. MG1655	No	[114]
MicA	*fumC*	b1611	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*galK*	b0757	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*hinT*	b1103	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*hns*	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*htrG*	b3055	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*lamB*	b4036	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*lpp*	b1677	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*nmpC*	b0553	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ompT*	b0565	*Escherichia coli str.* K-12 substr. MG1655	No	[172]
MicA	*ompW*	b1256	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*rbsB*	b3751	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*rbsK*	b3752	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*rluD*	b2594	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*rraB*	b4255	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*sodB*	b1656	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicA	*tsx*	b0411	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ycfL*	b1104	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*ydeN*	b1498	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*yfeK*	b2419	*Escherichia coli str.* K-12 substr. MG1655	No	[169]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
MicA	*yhcN*	b3238	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
MicA	*yhjJ*	b3527	*Escherichia coli str*. K-12 substr. MG1655	No	[169]
MicA	*yneM*	b4599	*Escherichia coli str.* K-12 substr. MG1655	No	[172]
MicC	*dppA*	b3544	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicC	*galK*	b0757	*Escherichia coli str*. K-12 substr. MG1655	No	[196]
MicC	*hns*	b1237	*Escherichia coli str*. K-12 substr. MG1655	No	[196]
MicC	*mraZ*	b0081	*Escherichia coli str.* K-12 substr. MG1655	No	[99]
MicC	*ompA*	b0957	*Escherichia coli str*. K-12 substr. MG1655	No	[196]
MicC	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicC	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
MicC	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicC	*sodB*	b1656	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicF	*dppA*	b3544	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicF	*galK*	b0757	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicF	*hns*	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicF	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicF	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
MicF	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
MicF	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
OmrA	*cheZ*	b1881	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*clpB*	b2592	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*csgB*	b1041	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrA	*csiE*	b2535	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*cydD*	b0887	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*deoR*	b0840	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*fdoL*	b3892	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*fimF*	b4318	*Escherichia coli str.* K-12 substr.	No	[126]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			MG1655		
OmrA	*folA*	b0048	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*glcD*	b2979	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*gmhB*	b0200	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*hisM*	b2307	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*hokB*	b4428	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*hokD*	b1562	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*lit*	b1139	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*malK*	b4035	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*mipA*	b1782	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*nadA*	b0750	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*narH*	b1225	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrA	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
OmrA	*osmB*	b1283	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrA	*rumA*	b2785	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*ssuC*	b0934	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*sufD*	b1681	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*uup*	b0949	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*xylH*	b3568	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yadD*	b0132	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yadL*	b0137	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yaeP*	b4406	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*ybcS*	b0555	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*ybeT*	b0647	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrA	*yccS*	b0960	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*ydbC*	b1406	*Escherichia coli str.* K-12 substr. MG1655	No	[178]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
OmrA	*ydhT*	b1669	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yeaZ*	b1807	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yfbT*	b2293	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*ygjN*	b3083	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yhbE*	b3184	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yrfC*	b3394	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrA	*yzgL*	b3427	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*csgB*	b1041	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*fepA*	b0584	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrB	*fimF*	b4318	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrB	*fldA*	b0684	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrB	*mutM*	b3635	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ompX*	stm0833	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrB	*osmB*	b1283	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
OmrB	*srlB*	b2704	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrB	*trxC*	b2582	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*yaeH*	b0163	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*yaiY*	b0379	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ybeT*	b0647	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ybhT*	b0762	*Escherichia coli str.* K-12 substr. MG1655	No	[126]
OmrB	*ybjE*	b0874	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ygaX*	b2013	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*yeeE*	b2680	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*yhdN*	b3293	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*yjhL*	b4299	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ykfL*	b0245	*Escherichia coli str.* K-12 substr.	No	[178]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			MG1655		
OmrB	*ykgJ*	b0288	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*ypdD*	b2383	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OmrB	*yphD*	b2546	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*dppD*	b3541	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*fabB*	b2323	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*gfcB*	b0986	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*moaD*	b0784	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
OxyS	*pmbA*	b4235	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*rpmG*	b3636	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*tolA*	b0739	*Escherichia coli str.* K-12 substr. MG1655	No	[197]
OxyS	*tolB*	b0740	*Escherichia coli str.* K-12 substr. MG1655	No	[197]
OxyS	*tolR*	b0738	*Escherichia coli str.* K-12 substr. MG1655	No	[197]
OxyS	*ybbB*	b0503	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*yccE*	b1001	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*yeaC*	b1777	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*yeaK*	b1787	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*yfdH*	b2351	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
OxyS	*yheN*	b3345	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RprA	*hns*	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
rprA	*mlrA*	b2127	*Escherichia coli str.* K-12 substr. MG1655	No	[180]
RprA	*ompA*	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RprA	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RprA	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RprA	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RprA	*phoU*	b3724	*Escherichia coli str.* K-12 substr. MG1655	No	[99]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
RprA	*sodB*	b1656	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RseX	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RybB	*lpp*	b1677	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
RybB	*ompT*	b0565	*Escherichia coli str.* K-12 substr. MG1655	No	[172]
RybB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RybB	*rbsB*	b3751	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
RybB	*yhcN*	b3238	*Escherichia coli str.* K-12 substr. MG1655	No	[169]
RydB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RydC	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RyeB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RyeC	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RyfA	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RygC	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RygD	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
RyhB	*citG*	b0613	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*entS*	b0591	*Escherichia coli str.* K-12 substr. MG1655	No	[185]
RyhB	*galK*	b0757	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RyhB	*gltA*	b0720	*Escherichia coli str.* K-12 substr. MG1655	No	[35]
RyhB	*hns*	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RyhB	*icd*	b1136	*Escherichia coli str.* K-12 substr. MG1655	No	[35]
RyhB	*icd*	b1136	*Escherichia coli str.* K-12 substr. MG1655	No	[185]
RyhB	*kdpA*	b0698	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*mdh*	b3236	*Escherichia coli str.* K-12 substr. MG1655	No	[35]
RyhB	*metH*	b4019	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*metL*	b0198	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*motA*	b1890	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*ompA*	b0957	*Escherichia coli str.* K-12 substr.	No	[196]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			MG1655		
RyhB	*ompC*	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RyhB	*ompF*	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RyhB	*perM*	b2493	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*pinH*	b2648	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*proA*	b0243	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*ptsG*	b1101	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
RyhB	*sucB*	b0727	*Escherichia coli str.* K-12 substr. MG1655	No	[35]
RyhB	*sucC*	b0728	*Escherichia coli str.* K-12 substr. MG1655	No	[35]
RyhB	*sucD*	b0729	*Escherichia coli str.* K-12 substr. MG1655	No	[35]
RyhB	*sugE*	b4148	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
ryhB	*tolC*	b3035	*Escherichia coli str.* K-12 substr. MG1655	No	[185]
RyhB	*yadS*	b0157	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*yagJ*	b0276	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*yagT*	b0286	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*ybjG*	b0841	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*ydaN*	b1342	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	*yecD*	b1867	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	yegK	b2072	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	ygeZ	b2873	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	ygiQ	b4469	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	ygiT	b3021	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	yheF	b3325	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	yiaM	b3577	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	ykgE	b0306	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB	ynfF	b1588	*Escherichia coli str.* K-12 substr. MG1655	No	[178]
RyhB-1	ompX	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
RyhB-2	ompX	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SgrS	dppA	b3544	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SgrS	hns	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SgrS	ompA	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SgrS	ompC	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SgrS	ompF	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SgrS	ompX	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SgrS	sodB	b1656	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SgrS	sopd2	stm0972	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[191]
SgrS	yigM	stm3963	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[192]
Spot42	ompX	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	No	[195]
Spot42	dppA	b3544	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
Spot42	fucK	b2803	*Escherichia coli str.* K-12 substr. MG1655	No	[198]
Spot42	hns	b1237	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
Spot42	ompA	b0957	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
Spot42	ompC	b2215	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
Spot42	ompF	b0929	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
Spot42	*sodB*	b1656	*Escherichia coli str.* K-12 substr. MG1655	No	[196]
SraB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SraF	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SraH	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SraL	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SroB	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SroC	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SsrA	*smpB*	b2620	*Escherichia coli str.* K-12 substr. MG1655	No	[199]
SsrS	*ompX*	stm0833	*Salmonella enterica subsp. enterica serovar Typhimurium str.* LT2	No	[195]
SsrS	*sspA*	b3229	*Escherichia coli str.* K-12 substr.	No	[19]

sRNA	mRNA	Locus ID	Species	Interaction	Reference
			MG1655		

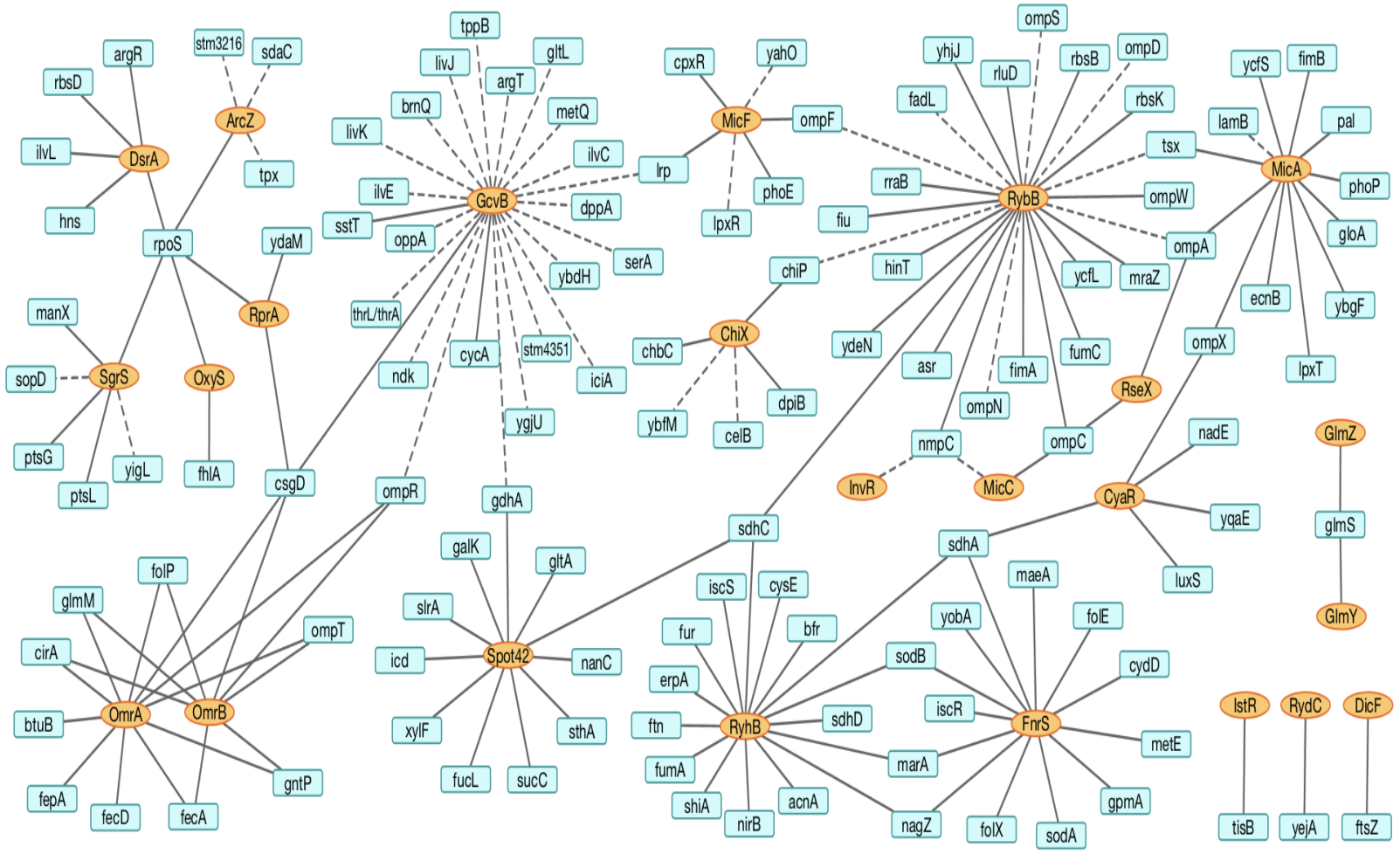


Figure S1: Known interacting pairs of sRNA and mRNAs used in the dataset. Solid connections have been experimentally

SUPPLEMENTARY S2: ITERATIVE FEATURE ELIMINATION RESULTS

Table S2: Performance evaluation of feature-sets through the iterative feature elimination process

No. of Features	Accuracy	Sensitivity	Specificity	MCC
17796	0.685	0.621	0.717	0.349
15000	0.699	0.627	0.736	0.374
12500	0.699	0.618	0.742	0.376
10000	0.707	0.630	0.748	0.391
7500	0.705	0.633	0.744	0.386
5000	0.728	0.630	0.782	0.426
2500	0.750	0.655	0.803	0.474
2000	0.759	0.664	0.813	0.492
1500	0.767	0.673	0.820	0.508
1000	0.774	0.686	0.823	0.524
900	0.784	0.709	0.826	0.548
800	0.789	0.733	0.820	0.560
700	0.791	0.718	0.831	0.563
600	0.792	0.727	0.827	0.564
500	0.794	0.719	0.835	0.569
400	0.805	0.731	0.846	0.591
300	0.814	0.748	0.850	0.612
200	0.818	0.747	0.860	0.618
100	0.821	0.760	0.856	0.630
90	0.809	0.758	0.836	0.605
80	0.812	0.765	0.839	0.615
70	0.823	0.771	0.853	0.634
60	0.819	0.775	0.842	0.626
59	0.822	0.789	0.839	0.637
58	0.819	0.774	0.844	0.625
57	0.822	0.785	0.840	0.633
56	0.817	0.773	0.840	0.622
55	0.824	0.779	0.848	0.638
54	0.824	0.784	0.844	0.639
53	0.826	0.784	0.850	0.643
52	0.825	0.802	0.834	0.645
51	0.826	0.793	0.844	0.644
50	0.829	0.793	0.847	0.649
49	0.830	0.793	0.850	0.650
48	0.826	0.779	0.852	0.640
47	0.823	0.773	0.849	0.632

No. of Features	Accuracy	Sensitivity	Specificity	MCC
46	0.822	0.778	0.846	0.634
45	0.827	0.784	0.850	0.644
44	0.821	0.781	0.841	0.630
43	0.817	0.779	0.836	0.622
42	0.817	0.781	0.834	0.622
41	0.819	0.779	0.839	0.626
40	0.819	0.774	0.844	0.626
30	0.815	0.772	0.839	0.616
20	0.786	0.757	0.802	0.563
10	0.801	0.782	0.809	0.593

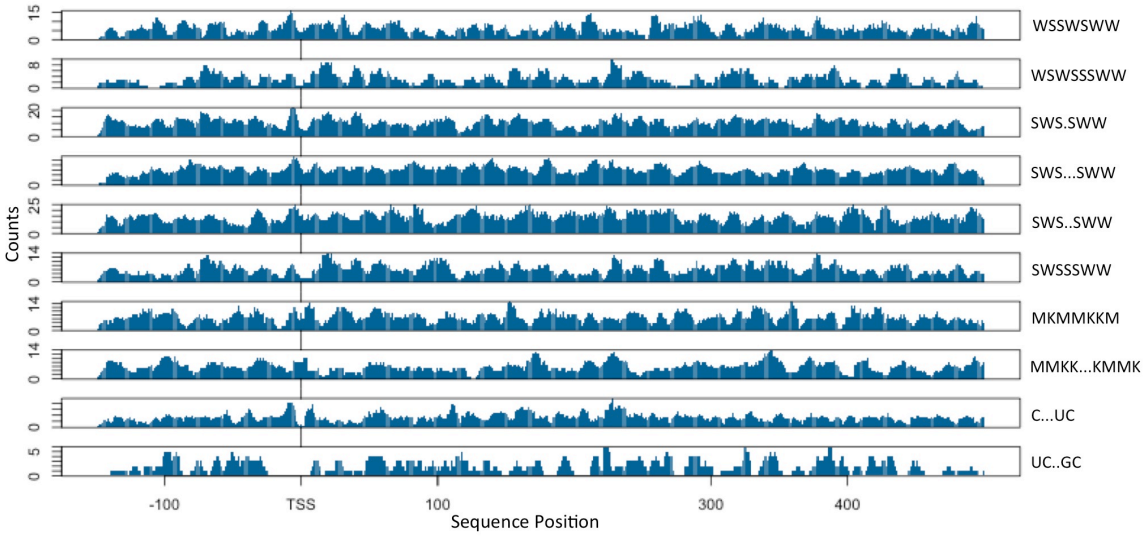
SUPPLEMENTARY S3: SET OF FEATURES SELECTED BY ITERATIVE FEATURE ELIMINATION

Table S3: Feature set selected for the final model. For new predictions using the model provided, column headers must have the names in the first column. The second column lists the actual patterns that were used to calculate the frequencies. The third column lists the source sequence from which the frequency was computed.

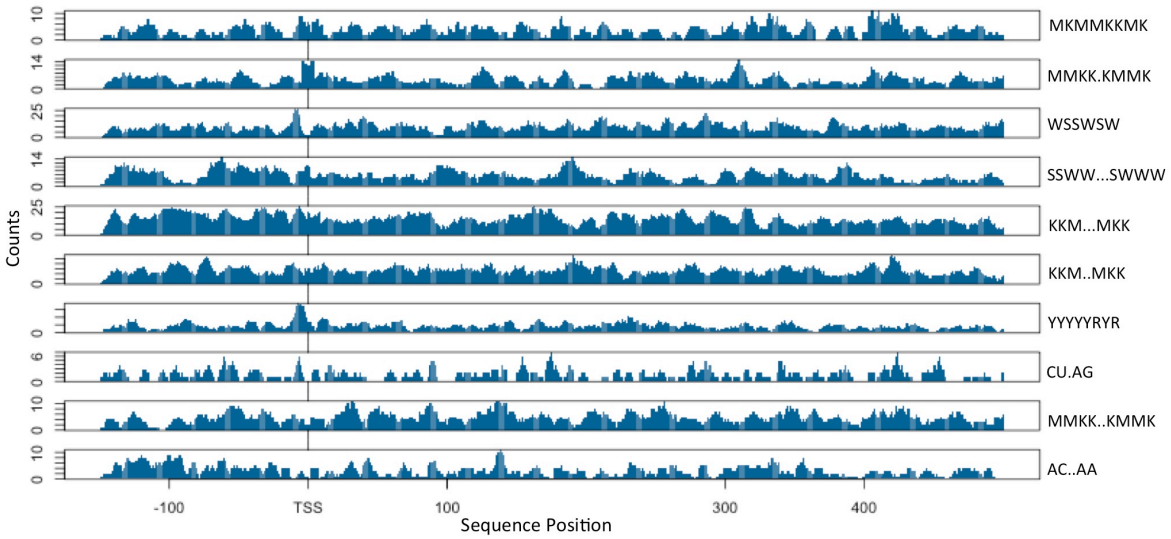
Column Header	Pattern	Sequence
stl1f6_33	YRRRRR	sRNA
mFtl2f7_53	WSSWSWW	Full mRNA
mFtl2f8_93	WSWSSSWW	Full mRNA
s3gf3_56	CC...U	sRNA
s3gf2_15	U...G	sRNA
stl3f4_16	KKKK	sRNA
mFtl2f7_93	SWSSSWW	Full mRNA
stl1f6_2	RRRRRY	sRNA
mFtl3f7_39	MKMMKKM	Full mRNA
mFgtl2f6_45	SWS.SWW	Full mRNA
mF3gtl2f6_45	SWS...SWW	Full mRNA
mF2gtl2f6_45	SWS..SWW	Full mRNA
stl1f5_1	RRRRR	sRNA
mF3gtl3f8_58	MMKK...KMMK	Full mRNA

Column Header	Pattern	Sequence
mF3gf3_46	C...UC	Full mRNA
mF2gf4_218	UC..GC	Full mRNA
mFtl3f8_78	MKMMKKMK	Full mRNA
mFgtl3f8_58	MMKK.KMMK	Full mRNA
mFtl2f6_27	WSSWSW	Full mRNA
mF3gtl2f8_201	SSWW...SWWW	Full mRNA
stl3f5_31	KKKKM	sRNA
stl1f7_95	YRYYYYR	sRNA
mF3gtl3f6_52	KKM...MKK	Full mRNA
mF2gtl3f6_52	KKM..MKK	Full mRNA
stl1f8_223	YYRYYYYR	sRNA
mFtl1f8_251	YYYYYRYR	Full mRNA
mFgf4_115	CU.AG	Full mRNA
mF2gtl3f8_58	MMKK..KMMK	Full mRNA
mF2gf4_17	AC..AA	Full mRNA
mFgtl2f8_201	SSWW.SWWW	Full mRNA
sgf3_3	A.AG	sRNA
s3gf3_39	C...CG	sRNA
sgtl3f6_60	KKK.MKK	sRNA
mFgtl3f6_52	KKM.MKK	Full mRNA
stl3f4_11	KMKM	sRNA
mF2gtl2f8_201	SSWW..SWWW	Full mRNA
s2gtl3f6_60	KKK..MKK	sRNA
mF2gtl2f8_114	WSSS..WWWS	Full mRNA
mF3gf3_85	GC...A	Full mRNA
stl3f7_124	KKKKMKK	sRNA
stl1f6_48	YRYYYY	sRNA
mFgtl2f8_114	WSSS.WWWS	Full mRNA
s3gf3_127	UU...G	sRNA
mF3gf4_220	UC...GU	Full mRNA
s3gtl3f6_60	KKK...MKK	sRNA
mF3gf4_30	AC...UC	Full mRNA
mF3gf4_209	UC...AA	Full mRNA
mFgtl2f8_89	WSWS.SWWW	Full mRNA
stl1f6_31	RYYYYR	sRNA

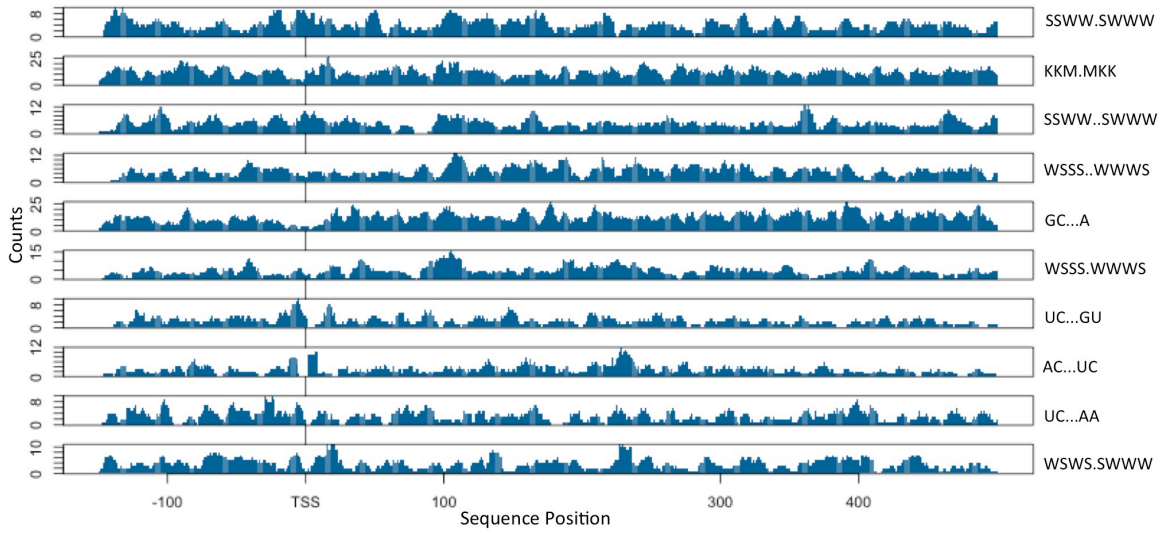
SUPPLEMENTARY S4: ANALYSIS OF SEQUENCE FEATURE POSITIONS IN SEQUENCES IN POSITIVE SET



A: Position wise counts of the top 1-10 k-mers for mRNA sequences

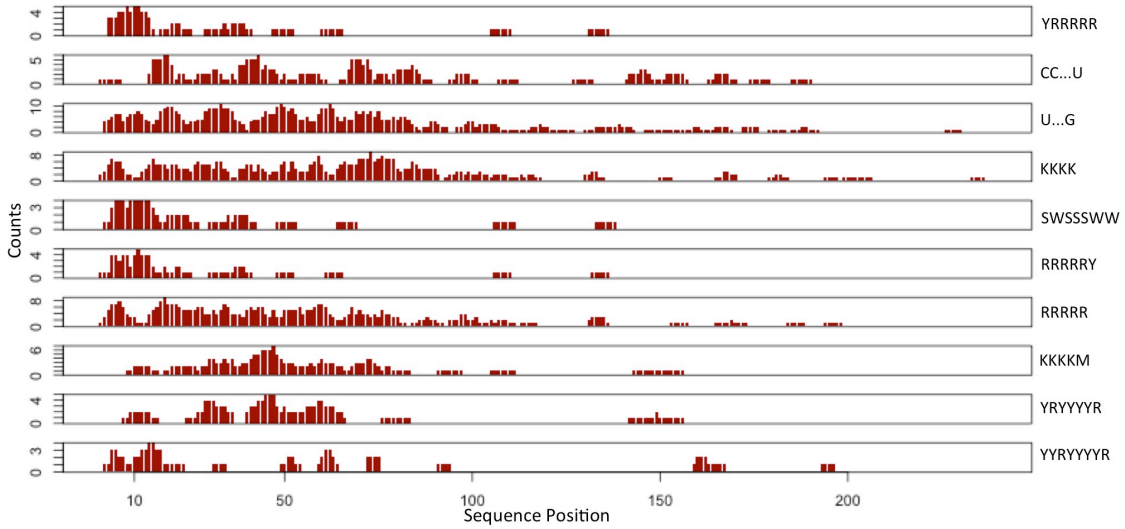


B: Position wise counts of the top 11-20 k-mers for mRNA sequences

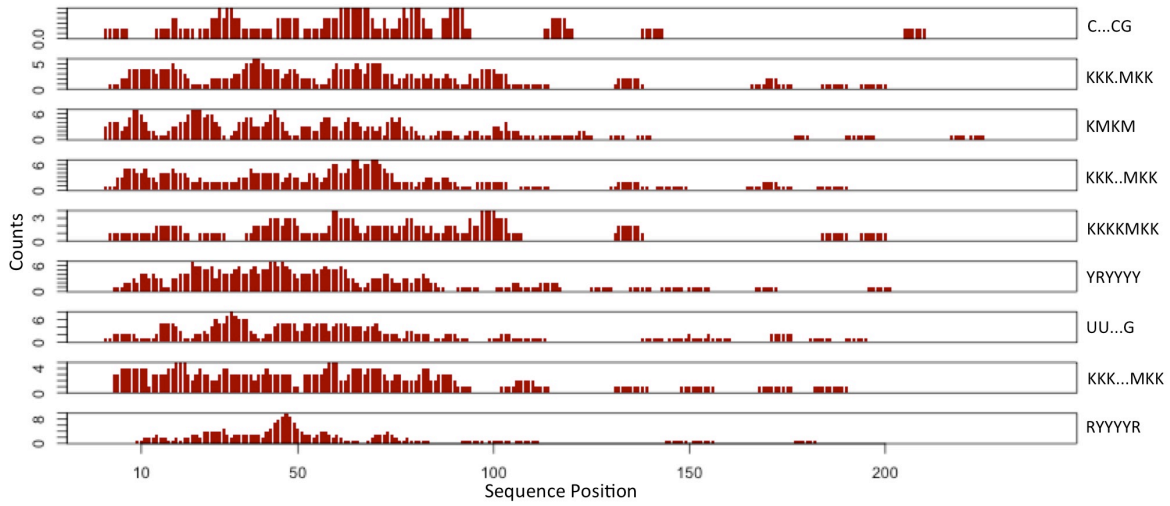


C:

Position wise counts of the top 21-30 k-mers for mRNA sequences



D: Position wise counts of the top 1-10 k-mers for sRNA sequences



E: Position wise counts of the top 1-10 k-mers for sRNA sequences

Figure S4: Positional occurrence of the 49 selected k-mer patterns in interacting sequences (positive set). The patterns are counted separately for mRNAs (A-C), showing 150nt upstream and 500nt downstream of the TSS; and the full sequences of sRNAs (D and E).

SUPPLEMENTARY S5: *E. COLI* GENOME-WIDE PREDICTIONS FOR SRNAS RYHB AND OMRA

Table S5: Genome wide prediction results for the RyhB and OmrA small RNAs across the *E. coli* genome. Only the top 100 results sorted by probability are shown here.

	RyhB Predictions			OmrA Predictions			IstR Predictions	
Rank	Gene	Probability		Gene	Probability		Gene	Probability
1	csgD	0.9276		phoP	0.8868		csgD	0.9538
2	gntP	0.9084		sdhC	0.8682		phoP	0.9494

3	phoP	0.9004		mhpA	0.8494		sdhC	0.9422
4	cirA	0.8684		gloA	0.8482		gntP	0.9392
5	gloA	0.858		murD	0.8468		fecA	0.9338
6	mhpA	0.8494		hybD	0.8354		rspA	0.9288
7	murD	0.8436		chbC	0.8272		polB	0.9234
8	chbC	0.8364		chiP	0.8248		hisG	0.922
9	ompR	0.822		erpA	0.812		hybD	0.921
10	hybD	0.821		cvrA	0.7978		murD	0.9202
11	fecA	0.8208		adiC	0.795		malZ	0.9176
12	nupG	0.817		coaA	0.7942		chbC	0.9168
13	chiP	0.8166		hisG	0.7928		mhpA	0.9164
14	ldtC	0.8104		degS	0.7928		cvrA	0.915
15	rspA	0.8038		menA	0.792		yhhJ	0.9146
16	btuB	0.7998		rspA	0.7902		ldtC	0.9142
17	yhhJ	0.7962		atoD	0.786		yejA	0.9132
18	coaA	0.7952		nupG	0.784		cirA	0.9126
19	polB	0.7906		gntU	0.7806		sthA	0.9126
20	atoD	0.7902		polB	0.7798		btuB	0.9126
21	cvrA	0.7886		yhhJ	0.776		yehX	0.9104
22	adiC	0.787		ybbW	0.7704		nupG	0.9102
23	menA	0.7866		sdhA	0.7694		gntU	0.9074

24	malZ	0.7854		ldtC	0.7684		chiP	0.9068
25	hisG	0.7812		prmA	0.7612		adiC	0.9066
26	gntU	0.781		yhfT	0.7608		gloA	0.9058
27	degS	0.7786		kbaZ	0.7596		bcsA	0.9058
28	kbaZ	0.7768		nfo	0.7542		menA	0.9046
29	yqjG	0.7742		ssuA	0.7474		ptrA	0.9044
30	ptrA	0.7714		lsrD	0.7468		glmM	0.904
31	ybbW	0.769		solA	0.7464		accA	0.9034
32	yejA	0.7686		dacB	0.7424		bglX	0.9018
33	lsrD	0.7638		mlaA	0.7414		sdhA	0.9012
34	yhfT	0.7626		yehX	0.7412		ybbW	0.9006
35	accA	0.7622		malZ	0.7382		yqjG	0.9006
36	nfo	0.7598		rutF	0.7378		dacB	0.8988
37	yehX	0.7592		arsB	0.7362		coaA	0.8978
38	mlaA	0.7568		yqjG	0.7326		gudD	0.8964
39	arsB	0.7554		gltS	0.7326		aroA	0.8944
40	nnr	0.7552		puuE	0.732		nikE	0.894
41	ssuA	0.751		yihT	0.731		cydD	0.8938
42	rutF	0.751		yhjA	0.7288		erpA	0.8926
43	ysgA	0.7508		ccmF	0.728		degS	0.8926
44	dacB	0.7492		ptrA	0.7264		arnA	0.8924

45	queC	0.7482		accA	0.7256		yhjA	0.892
46	aroA	0.7462		pbpG	0.7216		yraR	0.891
47	yqjE	0.7422		brnQ	0.7182		yoaD	0.8908
48	prmA	0.7418		queC	0.7164		ygcE	0.8904
49	puuE	0.7416		waaH	0.716		yhfT	0.8904
50	solA	0.741		cydD	0.7154		dgt	0.8898
51	yibF	0.7396		msbA	0.7152		prpR	0.8896
52	waaH	0.7388		yqjE	0.7136		modC	0.8896
53	yhjA	0.735		modC	0.7124		mhpD	0.8878
54	sthA	0.734		yejA	0.7098		treR	0.8856
55	lsrF	0.7316		yigL	0.7094		agaD	0.885
56	cydD	0.731		nagZ	0.7082		pbpG	0.8848
57	xapB	0.73		eutH	0.708		waaH	0.8844
58	fecD	0.73		otsA	0.707		rutF	0.8838
59	rluE	0.7268		bcsA	0.7056		cusB	0.8834
60	brnQ	0.7256		betT	0.7054		purL	0.8832
61	yigL	0.7222		glmM	0.7052		kbaZ	0.882
62	gsiA	0.7194		ascF	0.7044		yfhM	0.8816
63	ynfA	0.719		rsxD	0.7038		yafS	0.8802
64	pbpG	0.7156		nnr	0.7036		tdcE	0.88
65	kduD	0.7156		mnmG	0.7022		prmA	0.88

66	eutH	0.715		mhpD	0.702		solA	0.8786
67	otsA	0.7148		ycfT	0.702		mnmG	0.8786
68	mnaT	0.712		ynfA	0.701		nirB	0.8776
69	rsxD	0.7106		yibF	0.7004		ompR	0.8776
70	yoaD	0.7098		agaD	0.6982		metE	0.8776
71	glmM	0.7082		rluE	0.6966		nikA	0.8762
72	frlC	0.7072		aroA	0.6962		frlC	0.876
73	nikE	0.7068		ccmC	0.6944		frvB	0.876
74	mhpC	0.7062		sthA	0.6944		atoD	0.8754
75	msbA	0.7046		nikE	0.694		yehP	0.875
76	dgt	0.7042		bfr	0.6932		rlmI	0.8742
77	wcaE	0.7036		arnA	0.6924		yihT	0.8742
78	modC	0.7024		yraR	0.6924		yhjG	0.874
79	mhpD	0.7022		bglX	0.6906		rluA	0.8738
80	yqaE	0.6994		nikA	0.6896		eptC	0.8728
81	cdh	0.6986		lsrF	0.6892		arsB	0.8722
82	allC	0.6968		gsiC	0.6886		cycA	0.8722
83	asnC	0.6968		yiaN	0.6878		yibF	0.872
84	cusB	0.6966		ysgA	0.6878		csdA	0.8718
85	mnmG	0.694		xapB	0.6854		msbA	0.8716
86	uraA	0.693		wcaE	0.6852		treF	0.871

87	yccA	0.6926		uraA	0.6846		waaA	0.8704
88	ycfT	0.6926		amiB	0.6818		pfo	0.87
89	fhuF	0.6922		rlmI	0.6814		puuE	0.8698
90	bamE	0.6914		allC	0.6796		edd	0.8696
91	yidZ	0.6914		frvB	0.6794		ccmF	0.8694
92	agaD	0.6906		yoaJ	0.6786		lsrF	0.8692
93	treR	0.6906		prpC	0.678		sad	0.869
94	pldB	0.6896		marA	0.6772		yqjE	0.8682
95	arnA	0.6886		asnC	0.6762		ysgA	0.8682
96	rpsN	0.6878		gudD	0.676		treA	0.8676
97	yiaN	0.6862		dgt	0.6754		hycG	0.8676
98	yraR	0.6858		yeaN	0.6742		priA	0.8674
99	yihT	0.6856		yoaD	0.6738		gspL	0.8672
100	uidC	0.6832		mhpC	0.6726		gmd	0.867

SUPPLEMENTARY DATA FOR CHAPTER 5

Table S6: Genome wide prediction results for the 6 small RNAs studied across the *E. coli* genome. Only the top 100 results sorted by probability are shown here.

A. RybB

Rank	Gene Name	Gene ID	Locus ID	Prediction Probability
1	rspA	946126	b1581	0.986
2	sthA	948461	b3962	0.982
3	sdhC	945316	b0721	0.980
4	glmM	947692	b3176	0.978
5	mngB	945359	b0732	0.977
6	bglX	946682	b2132	0.977
7	purL	947032	b2557	0.975
8	narZ	945999	b1468	0.973
9	arnA	947683	b2255	0.973
10	srlA	947575	b2702	0.973
11	yqjG	947615	b3102	0.973
12	glmS	948241	b3729	0.972
13	yghJ	2847716	b4466	0.972
14	polB	944779	b0060	0.972
15	yfhM	947302	b2520	0.972
16	malZ	949131	b0403	0.971
17	yejA	946675	b2177	0.971
18	aegA	947383	b2468	0.971
19	sdhA	945402	b0723	0.971
20	priA	948426	b3935	0.971
21	rcsD	946717	b2216	0.970
22	nagC	945285	b0676	0.970
23	yehP	946652	b2121	0.970
24	tdcE	947623	b3114	0.969
25	ybhC	945381	b0772	0.969
26	btuB	948468	b3966	0.969
27	ddpA	946052	b1487	0.968
28	pfo	946587	b1378	0.968
29	tktA	947420	b2935	0.968
30	pfkB	946230	b1723	0.967
31	mscM	948676	b4159	0.967
32	fiu	946246	b0805	0.967
33	ptrA	947284	b2821	0.967

34	ftsK	945102	b0890	0.966
35	yhdP	2847740	b4472	0.966
36	selB	948103	b3590	0.966
37	metH	948522	b4019	0.966
38	etk	947409	b0981	0.965
39	yraR	947667	b3152	0.965
40	rluD	947087	b2594	0.965
41	nikA	947981	b3476	0.964
42	dmsA	945508	b0894	0.964
43	yfeW	946907	b2430	0.964
44	yhjJ	948040	b3527	0.964
45	cydD	949052	b0887	0.964
46	barA	947255	b2786	0.964
47	oxyR	948462	b3961	0.964
48	pepT	946333	b1127	0.963
49	dacB	947693	b3182	0.963
50	ycgV	945767	b1202	0.963
51	fadD	946327	b1805	0.963
52	menD	946720	b2264	0.963
53	lhr	946156	b1653	0.963
54	gspA	947825	b3323	0.963
55	rpoH	947970	b3461	0.962
56	yjiR	949089	b4340	0.962
57	cusB	945189	b0574	0.962
58	ldtC	945666	b1113	0.962
59	parC	947499	b3019	0.962
60	aroK	2847759	b3390	0.962
61	hisG	946549	b2019	0.962
62	fadI	948823	b2342	0.962
63	nadB	947049	b2574	0.962
64	cyaA	947755	b3806	0.962
65	yejF	946689	b2180	0.962
66	lpxK	945526	b0915	0.961
67	narG	945782	b1224	0.961
68	ispB	947364	b3187	0.961
69	menA	948418	b3930	0.961
70	rcsC	948993	b2218	0.961
71	bepA	947029	b2494	0.961
72	frlC	2847758	b4474	0.961
73	gntR	947946	b3438	0.960
74	ybhJ	945380	b0771	0.960

75	ycbZ	945569	b0955	0.960
76	xdhA	947116	b2866	0.960
77	tsaD	947578	b3064	0.960
78	ybhF	945413	b0794	0.960
79	tynA	945939	b1386	0.960
80	treF	948037	b3519	0.960
81	atoC	947444	b2220	0.960
82	puuA	946202	b1297	0.960
83	metE	948323	b3829	0.959
84	panE	945065	b0425	0.959
85	gspL	947842	b3333	0.959
86	allC	945150	b0516	0.959
87	ydbH	945949	b1381	0.959
88	lysA	947313	b2838	0.959
89	yjjI	948904	b4380	0.959
90	ileS	944761	b0026	0.959
91	kdpD	946744	b0695	0.959
92	recB	947286	b2820	0.959
93	poxB	946132	b0871	0.958
94	csgD	949119	b1040	0.958
95	ccmH	946623	b2194	0.958
96	yegD	947234	b2069	0.958
97	cirA	949042	b2155	0.958
98	ade	948210	b3665	0.958
99	entF	945184	b0586	0.958
100	nemA	946164	b1650	0.958

B. MicC

Rank	Gene Name	Gene ID	Locus ID	Prediction Probability
1	csgD	949119	b1040	0.908
2	fecA	946427	b4291	0.851
3	gntP	948848	b4321	0.847
4	cirA	949042	b2155	0.821
5	glmM	947692	b3176	0.819
6	sdhC	945316	b0721	0.812
7	mhpA	945197	b0347	0.811
8	ompR	947913	b3405	0.807
9	murD	944818	b0088	0.802
10	gloA	946161	b1651	0.801
11	chiP	945296	b0681	0.798
12	hybD	948982	b2993	0.781

13	phoP	945697	b1130	0.780
14	fecD	946816	b4288	0.780
15	sdhA	945402	b0723	0.779
16	btuB	948468	b3966	0.778
17	folP	947691	b3177	0.774
18	chbC	945982	b1737	0.768
19	degS	947865	b3235	0.767
20	gntU	2847760	b4476	0.766
21	cvrA	945755	b1191	0.755
22	erpA	944857	b0156	0.753
23	yhhJ	947991	b3485	0.752
24	coaA	948479	b3974	0.747
25	ybbW	945138	b0511	0.746
26	nupG	946282	b2964	0.742
27	bglX	946682	b2132	0.742
28	adiC	948628	b4115	0.740
29	atoD	947525	b2221	0.737
30	ldtC	945666	b1113	0.736
31	rspA	946126	b1581	0.736
32	ccmC	946703	b2199	0.735
33	nfo	946669	b2159	0.735
34	bcsA	948053	b3533	0.733
35	kbaZ	947637	b3132	0.733
36	betT	945079	b0314	0.731
37	mdtC	946608	b2076	0.730
38	mlaA	945582	b2346	0.729
39	glmS	948241	b3729	0.727
40	oppD	945802	b1246	0.725
41	polB	944779	b0060	0.724
42	yihT	948373	b3881	0.724
43	hisG	946549	b2019	0.723
44	menA	948418	b3930	0.723
45	selB	948103	b3590	0.720
46	ssuA	945560	b0936	0.719
47	queC	947034	b0444	0.717
48	emrB	947167	b2686	0.716
49	malZ	949131	b0403	0.715
50	sdhD	945322	b0722	0.713
51	prmA	947708	b3259	0.711
52	yhfT	947883	b3377	0.710
53	arsB	948011	b3502	0.710

54	lpxT	946693	b2174	0.710
55	gltS	948166	b3653	0.707
56	eptC	948458	b3955	0.707
57	hycG	947191	b2719	0.706
58	yoaE	946335	b1816	0.705
59	yqjG	947615	b3102	0.704
60	accA	944895	b0185	0.703
61	poxB	946132	b0871	0.703
62	msbA	945530	b0914	0.702
63	solA	944983	b1059	0.701
64	pbpG	946662	b2134	0.701
65	nagC	945285	b0676	0.701
66	ccmF	948783	b2196	0.700
67	eutH	944979	b2452	0.699
68	dacB	947693	b3182	0.699
69	yhjA	948038	b3518	0.698
70	ycfT	945679	b1115	0.697
71	yehX	946659	b2129	0.697
72	bcr	944808	b2182	0.696
73	prpR	944987	b0330	0.696
74	rluE	945701	b1135	0.695
75	xapB	946868	b2406	0.695
76	agaD	947649	b3140	0.694
77	purL	947032	b2557	0.691
78	rng	947744	b3247	0.691
79	puuE	945446	b1302	0.691
80	queG	948686	b4166	0.691
81	brnQ	945042	b0401	0.689
82	pal	945004	b0741	0.689
83	nirB	947868	b3365	0.688
84	rpoS	947210	b2741	0.688
85	yejA	946675	b2177	0.687
86	yigL	2847768	b3826	0.686
87	gcvP	947394	b2903	0.684
88	dxs	945060	b0420	0.684
89	tdcE	947623	b3114	0.683
90	sthA	948461	b3962	0.683
91	ldcC	944887	b0186	0.682
92	lsrD	946264	b1515	0.682
93	csdA	947275	b2810	0.682
94	frvB	948390	b3899	0.680

95	ptrA	947284	b2821	0.680
96	nikE	947987	b3480	0.679
97	ybbP	945118	b0496	0.679
98	cycA	948725	b4208	0.676
99	aqpZ	945497	b0875	0.676
100	fhlA	947181	b2731	0.676

C. RseX

Rank	Gene Name	Gene ID	Locus ID	Prediction Probability
1	sdhC	945316	b0721	0.969
2	rspA	946126	b1581	0.969
3	polB	944779	b0060	0.965
4	malZ	949131	b0403	0.959
5	sdhA	945402	b0723	0.959
6	sthA	948461	b3962	0.959
7	btuB	948468	b3966	0.959
8	glmS	948241	b3729	0.958
9	glmM	947692	b3176	0.958
10	csgD	949119	b1040	0.958
11	yejA	946675	b2177	0.955
12	bglX	946682	b2132	0.955
13	chbC	945982	b1737	0.954
14	hisG	946549	b2019	0.954
15	fecA	946427	b4291	0.951
16	cydD	949052	b0887	0.950
17	chiP	945296	b0681	0.950
18	cirA	949042	b2155	0.950
19	yfhM	947302	b2520	0.949
20	purL	947032	b2557	0.948
21	yghJ	2847716	b4466	0.947
22	ldtC	945666	b1113	0.947
23	tdcE	947623	b3114	0.947
24	pfo	946587	b1378	0.946
25	ptrA	947284	b2821	0.946
26	murD	944818	b0088	0.946
27	yqjG	947615	b3102	0.945
28	dacB	947693	b3182	0.945
29	menA	948418	b3930	0.945
30	phoP	945697	b1130	0.945
31	yraR	947667	b3152	0.944

32	dmsA	945508	b0894	0.944
33	gspA	947825	b3323	0.944
34	bcsA	948053	b3533	0.944
35	mhpA	945197	b0347	0.943
36	arnA	947683	b2255	0.942
37	accA	944895	b0185	0.942
38	yafS	944903	b0213	0.942
39	cusB	945189	b0574	0.942
40	nagC	945285	b0676	0.942
41	yehP	946652	b2121	0.942
42	hybD	948982	b2993	0.942
43	yehX	946659	b2129	0.941
44	nupG	946282	b2964	0.941
45	nagZ	945671	b1107	0.940
46	aroA	945528	b0908	0.940
47	priA	948426	b3935	0.940
48	cvrA	945755	b1191	0.939
49	tktA	947420	b2935	0.939
50	dxs	945060	b0420	0.938
51	ddpA	946052	b1487	0.937
52	nikA	947981	b3476	0.937
53	narZ	945999	b1468	0.937
54	yhjA	948038	b3518	0.937
55	treF	948037	b3519	0.937
56	mhpC	944954	b0349	0.936
57	recB	947286	b2820	0.936
58	gspL	947842	b3333	0.936
59	prpR	944987	b0330	0.936
60	etk	947409	b0981	0.936
61	ftsK	945102	b0890	0.936
62	waaH	948140	b3615	0.935
63	ybhC	945381	b0772	0.935
64	panE	945065	b0425	0.935
65	sapA	945873	b1294	0.935
66	mscM	948676	b4159	0.935
67	nikE	947987	b3480	0.935
68	ccmH	946623	b2194	0.934
69	adiC	948628	b4115	0.934
70	marC	947132	b1529	0.934
71	metE	948323	b3829	0.934
72	mngB	945359	b0732	0.933

73	edd	946362	b1851	0.933
74	clsC	947321	b1046	0.933
75	ygcE	946193	b2776	0.933
76	yibF	948113	b3592	0.933
77	dgt	947177	b0160	0.933
78	tsaD	947578	b3064	0.933
79	frlC	2847758	b4474	0.933
80	poxB	946132	b0871	0.932
81	pepT	946333	b1127	0.932
82	treA	945757	b1197	0.932
83	bepA	947029	b2494	0.932
84	ybhJ	945380	b0771	0.932
85	elfC	946934	b0940	0.932
86	selB	948103	b3590	0.932
87	lhr	946156	b1653	0.932
88	eptC	948458	b3955	0.931
89	yoaD	946336	b1815	0.931
90	srlA	947575	b2702	0.931
91	gntU	2847760	b4476	0.931
92	csdA	947275	b2810	0.931
93	yqeG	945028	b2845	0.931
94	lpxK	945526	b0915	0.931
95	msbA	945530	b0914	0.930
96	dapE	948313	b2472	0.930
97	ybbW	945138	b0511	0.930
98	mdtC	946608	b2076	0.930
99	prmA	947708	b3259	0.930
100	fadI	948823	b2342	0.930

D. OxyS

Rank	Gene Name	Gene ID	Locus ID	Prediction Probability
1	csgD	949119	b1040	0.882
2	fecA	946427	b4291	0.843
3	gntP	948848	b4321	0.828
4	cirA	949042	b2155	0.817
5	glmM	947692	b3176	0.816
6	sdhC	945316	b0721	0.793
7	ompR	947913	b3405	0.793
8	chiP	945296	b0681	0.788
9	mhpA	945197	b0347	0.779
10	murD	944818	b0088	0.772

11	fecD	946816	b4288	0.772
12	gloA	946161	b1651	0.768
13	folP	947691	b3177	0.763
14	sdhA	945402	b0723	0.758
15	btuB	948468	b3966	0.754
16	hybD	948982	b2993	0.753
17	phoP	945697	b1130	0.751
18	fhlA	947181	b2731	0.744
19	gntU	2847760	b4476	0.739
20	rpoS	947210	b2741	0.736
21	glmS	948241	b3729	0.735
22	chbC	945982	b1737	0.734
23	degS	947865	b3235	0.728
24	cvrA	945755	b1191	0.728
25	yhhJ	947991	b3485	0.721
26	erpA	944857	b0156	0.718
27	kbaZ	947637	b3132	0.716
28	betT	945079	b0314	0.713
29	bglX	946682	b2132	0.713
30	coaA	948479	b3974	0.711
31	mdtC	946608	b2076	0.709
32	mlaA	945582	b2346	0.707
33	ybbW	945138	b0511	0.705
34	nupG	946282	b2964	0.702
35	adiC	948628	b4115	0.702
36	ldtC	945666	b1113	0.701
37	hisG	946549	b2019	0.701
38	nfo	946669	b2159	0.698
39	atoD	947525	b2221	0.698
40	bcsA	948053	b3533	0.698
41	ccmC	946703	b2199	0.697
42	rspA	946126	b1581	0.695
43	yoaE	946335	b1816	0.693
44	ssuA	945560	b0936	0.692
45	yhfT	947883	b3377	0.692
46	yihT	948373	b3881	0.691
47	polB	944779	b0060	0.690
48	emrB	947167	b2686	0.689
49	selB	948103	b3590	0.686
50	prmA	947708	b3259	0.685
51	yqjG	947615	b3102	0.682

52	oppD	945802	b1246	0.680
53	agaD	947649	b3140	0.679
54	eptC	948458	b3955	0.677
55	menA	948418	b3930	0.676
56	queC	947034	b0444	0.676
57	ccmF	948783	b2196	0.675
58	nagC	945285	b0676	0.675
59	malZ	949131	b0403	0.674
60	gltS	948166	b3653	0.673
61	poxB	946132	b0871	0.672
62	hycG	947191	b2719	0.672
63	dacB	947693	b3182	0.671
64	arsB	948011	b3502	0.667
65	tdcE	947623	b3114	0.664
66	sdhD	945322	b0722	0.662
67	queG	948686	b4166	0.661
68	bcr	944808	b2182	0.660
69	yhjA	948038	b3518	0.660
70	dxs	945060	b0420	0.659
71	solA	944983	b1059	0.659
72	msbA	945530	b0914	0.658
73	puuE	945446	b1302	0.657
74	accA	944895	b0185	0.657
75	rng	947744	b3247	0.656
76	eutH	944979	b2452	0.655
77	frvB	948390	b3899	0.655
78	lsrD	946264	b1515	0.654
79	yehX	946659	b2129	0.654
80	purL	947032	b2557	0.653
81	ycfT	945679	b1115	0.652
82	nagZ	945671	b1107	0.652
83	sthA	948461	b3962	0.652
84	rluE	945701	b1135	0.651
85	mmuP	946284	b0260	0.651
86	brnQ	945042	b0401	0.651
87	nikE	947987	b3480	0.651
88	ompT	945185	b0565	0.650
89	pbpG	946662	b2134	0.650
90	gcvP	947394	b2903	0.648
91	glnD	944863	b0167	0.647
92	bfr	947839	b3336	0.647

93	ybbP	945118	b0496	0.646
94	cycA	948725	b4208	0.646
95	gdhA	946802	b1761	0.645
96	waaH	948140	b3615	0.643
97	rutF	946594	b1007	0.642
98	gntR	947946	b3438	0.641
99	csdA	947275	b2810	0.640
100	ilvY	948284	b3773	0.639

E. DicF

Rank	Gene Name	Gene ID	Locus ID	Prediction Probability
1	rspA	946126	b1581	0.877
2	sdhC	945316	b0721	0.876
3	csgD	949119	b1040	0.874
4	btuB	948468	b3966	0.868
5	polB	944779	b0060	0.868
6	ftsZ	944786	b0095	0.866
7	hisG	946549	b2019	0.861
8	menA	948418	b3930	0.861
9	malZ	949131	b0403	0.858
10	chbC	945982	b1737	0.858
11	phoP	945697	b1130	0.858
12	chiP	945296	b0681	0.857
13	nupG	946282	b2964	0.857
14	ldtC	945666	b1113	0.856
15	cirA	949042	b2155	0.855
16	yejA	946675	b2177	0.854
17	dacB	947693	b3182	0.854
18	yqjG	947615	b3102	0.853
19	sdhA	945402	b0723	0.852
20	cydD	949052	b0887	0.852
21	arnA	947683	b2255	0.851
22	ptrA	947284	b2821	0.851
23	yhjA	948038	b3518	0.851
24	yehX	946659	b2129	0.850
25	gntP	948848	b4321	0.849
26	yraR	947667	b3152	0.849
27	yafS	944903	b0213	0.848
28	aroA	945528	b0908	0.848
29	fecA	946427	b4291	0.847
30	ybbW	945138	b0511	0.846

31	mhpA	945197	b0347	0.846
32	pfo	946587	b1378	0.846
33	prpR	944987	b0330	0.846
34	edd	946362	b1851	0.846
35	dmsA	945508	b0894	0.844
36	ygcE	946193	b2776	0.844
37	tdcE	947623	b3114	0.844
38	murD	944818	b0088	0.843
39	modF	945368	b0760	0.843
40	gspA	947825	b3323	0.843
41	sthA	948461	b3962	0.842
42	ddpA	946052	b1487	0.842
43	kbaZ	947637	b3132	0.841
44	hybD	948982	b2993	0.840
45	mnmG	948248	b3741	0.840
46	accA	944895	b0185	0.840
47	yhhJ	947991	b3485	0.840
48	yfhM	947302	b2520	0.840
49	priA	948426	b3935	0.840
50	secM	944831	b0097	0.839
51	purL	947032	b2557	0.839
52	tsaD	947578	b3064	0.839
53	glmM	947692	b3176	0.839
54	yehP	946652	b2121	0.838
55	yidH	948190	b3676	0.838
56	yibF	948113	b3592	0.838
57	marC	947132	b1529	0.837
58	waaH	948140	b3615	0.837
59	sodA	948403	b3908	0.837
60	hsrA	948265	b3754	0.836
61	bcsA	948053	b3533	0.836
62	sfmA	945522	b0530	0.836
63	ygjJ	947597	b3079	0.836
64	clsC	947321	b1046	0.835
65	yeiI	946640	b2160	0.835
66	gntU	2847760	b4476	0.835
67	ybjP	945491	b0865	0.835
68	degS	947865	b3235	0.835
69	allC	945150	b0516	0.835
70	nikA	947981	b3476	0.835
71	mhpC	944954	b0349	0.834

72	rutF	946594	b1007	0.834
73	nagZ	945671	b1107	0.834
74	nirB	947868	b3365	0.834
75	ybgS	945356	b0753	0.834
76	gspL	947842	b3333	0.833
77	yfhH	947030	b2561	0.833
78	treF	948037	b3519	0.833
79	flhA	946390	b1879	0.833
80	elfC	946934	b0940	0.832
81	prmA	947708	b3259	0.832
82	glmS	948241	b3729	0.832
83	pepT	946333	b1127	0.832
84	cusB	945189	b0574	0.832
85	msbA	945530	b0914	0.831
86	tktA	947420	b2935	0.831
87	narZ	945999	b1468	0.830
88	cvrA	945755	b1191	0.830
89	gatZ	946641	b2095	0.830
90	dgt	947177	b0160	0.830
91	bglX	946682	b2132	0.830
92	ftsK	945102	b0890	0.829
93	pbpG	946662	b2134	0.829
94	adiC	948628	b4115	0.829
95	bcsE	948050	b3536	0.828
96	yiaO	948091	b3579	0.828
97	der	946983	b2511	0.828
98	flgE	945636	b1076	0.828
99	puuE	945446	b1302	0.828
100	rsxD	946134	b1630	0.827

F. RprA

Rank	Gene Name	Gene ID	Locus ID	Prediction Probability
1	csgD	949119	b1040	0.954
2	fecA	946427	b4291	0.890
3	cirA	949042	b2155	0.878
4	gntP	948848	b4321	0.876
5	sdhC	945316	b0721	0.864
6	sdhA	945402	b0723	0.862
7	chiP	945296	b0681	0.861
8	glmM	947692	b3176	0.860
9	mhpA	945197	b0347	0.858

10	murD	944818	b0088	0.847
11	hybD	948982	b2993	0.840
12	chbC	945982	b1737	0.840
13	gloA	946161	b1651	0.838
14	btuB	948468	b3966	0.831
15	gntU	2847760	b4476	0.827
16	cvrA	945755	b1191	0.822
17	bglX	946682	b2132	0.813
18	adiC	948628	b4115	0.811
19	fecD	946816	b4288	0.811
20	rpoS	947210	b2741	0.809
21	erpA	944857	b0156	0.809
22	ompR	947913	b3405	0.807
23	malZ	949131	b0403	0.806
24	rspA	946126	b1581	0.805
25	mlaA	945582	b2346	0.804
26	folP	947691	b3177	0.804
27	prmA	947708	b3259	0.803
28	polB	944779	b0060	0.800
29	nfo	946669	b2159	0.799
30	hisG	946549	b2019	0.799
31	degS	947865	b3235	0.799
32	yhhJ	947991	b3485	0.797
33	menA	948418	b3930	0.795
34	eptC	948458	b3955	0.794
35	ssuA	945560	b0936	0.793
36	ydaM	945909	b1341	0.793
37	ybbW	945138	b0511	0.792
38	kbaZ	947637	b3132	0.791
39	mdtC	946608	b2076	0.787
40	coaA	948479	b3974	0.786
41	selB	948103	b3590	0.784
42	glmS	948241	b3729	0.784
43	yihT	948373	b3881	0.784
44	yejA	946675	b2177	0.783
45	phoP	945697	b1130	0.783
46	bcsA	948053	b3533	0.783
47	queC	947034	b0444	0.782
48	emrB	947167	b2686	0.782
49	arsB	948011	b3502	0.781
50	betT	945079	b0314	0.781

51	agaD	947649	b3140	0.780
52	dacB	947693	b3182	0.779
53	yqjG	947615	b3102	0.779
54	poxB	946132	b0871	0.777
55	nagZ	945671	b1107	0.777
56	atoD	947525	b2221	0.774
57	nikE	947987	b3480	0.771
58	sthA	948461	b3962	0.771
59	ccmF	948783	b2196	0.769
60	ybbP	945118	b0496	0.767
61	solA	944983	b1059	0.766
62	ldtC	945666	b1113	0.766
63	accA	944895	b0185	0.765
64	ccmC	946703	b2199	0.764
65	nagC	945285	b0676	0.763
66	gdhA	946802	b1761	0.763
67	ptrA	947284	b2821	0.763
68	yhfT	947883	b3377	0.763
69	hycG	947191	b2719	0.762
70	rluE	945701	b1135	0.761
71	dxs	945060	b0420	0.761
72	nupG	946282	b2964	0.761
73	rng	947744	b3247	0.761
74	bcr	944808	b2182	0.761
75	yhjA	948038	b3518	0.760
76	yehX	946659	b2129	0.759
77	prpR	944987	b0330	0.759
78	tdcE	947623	b3114	0.759
79	queG	948686	b4166	0.759
80	oppD	945802	b1246	0.758
81	sdhD	945322	b0722	0.757
82	gcvP	947394	b2903	0.757
83	purL	947032	b2557	0.756
84	pbpG	946662	b2134	0.756
85	yoaE	946335	b1816	0.755
86	gltS	948166	b3653	0.755
87	eutH	944979	b2452	0.754
88	puuE	945446	b1302	0.752
89	dgt	947177	b0160	0.751
90	waaH	948140	b3615	0.750
91	brnQ	945042	b0401	0.748

92	msbA	945530	b0914	0.748
93	mnmG	948248	b3741	0.748
94	yoaD	946336	b1815	0.747
95	narZ	945999	b1468	0.746
96	nirB	947868	b3365	0.745
97	xapB	946868	b2406	0.745
98	sapA	945873	b1294	0.745
99	ycfT	945679	b1115	0.744
100	yibF	948113	b3592	0.743