1-2-2013

Investigating Hfq-Mrna Interactions In Bacteria

Martha Audra Faner
Wayne State University,

INVESTIGATING HFQ-mRNA INTERACTIONS IN BACTERIA

by

MARTHA A. FANER

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2013

MAJOR: CHEMISTRY (Biochemistry)

Approved by:

_____

Advisor Date

_____

_____

_____

DEDICATION

Dedicated to

my Husband

Todd Faner

and also to my Parents

Edward and Martha Mahaney

ACKNOWLEDGEMENTS

Since I was very young I have strived to achieve beyond the average. The culmination of that attitude and hard work is earning this degree. The support of many people allowed me to formulate and achieve this goal. I would like to acknowledge those people in approximate chronological order. Thanks to my parents, Edward and Martha Mahaney for instilling in me the confidence, curiosity, creativity, independence and perseverance that were required in order to complete this journey. Also to my parents, for being an ever present cheering section in all of my pursuits. Thanks to my sister, Bridgit Spielman for modeling strength of character and personal autonomy. I would like to thank my husband, Todd Faner who provides me with endless love, support, and encouragement in both my personal and professional life. Thanks to the Chemistry Department at Northern Michigan whose faculty helped me recognize and foster my love of biochemistry. I would like to acknowledge the excellent mentoring provided to me by my adviser Prof. Andrew Feig, who took me as his student for these six years and taught me how to be a scientist and teacher. Thank you to all of the Feig lab members that have made daily life in the lab a pleasure. Special thanks to Dr. Nilshad Salim and Dr. Stephanie (Kern) Kokoszka who donated their time to help me troubleshoot experiments and improve my general morale. Thank you to Prof. Tamara Hendrickson for providing needed counsel and laughs. I thank all of the Professors for whom I was teaching assistant, especially Dr. Barber and Dr. Zibuck who were particularly inspiring educators and fostered my desire to teach as well. I would like to acknowledge Melissa Barton for her guidance and encouragement through the program and Bernie Miesik for patiently helping me navigate the financial world of the

department. Dr. Emil Lozanov was instrumental to my success when teaching my own course by helping me create and implement fantastic chemical demonstrations. Many thanks to all of my friends who made work fun and who provided a supportive ear during tough times; in particular Ashley Floyd, Stephanie (Kern) Kokoszka and Rick Federley. A general thanks to Wayne State University for financial support and also for providing such a rich and diverse environment to be a part of for the last six years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND NAMING CONVENTIONS

Abbreviations

$(ARN)_x$	A is adenine, R is adenine or guanosine, and N is any nucleotide
5'RACE	5' Rapid amplification of cDNA ends
CLIP	Cross-linking and immunoprecipitation
CLR	Context likelihood of relatedness
co-IP	Coimmunoprecipitation
CPM	Counts per minute
EMSA	Electrophoretic mobility gel shift assay
GlcN6P	Glucosamine-6-phosphate
HTS	High throughput sequencing
ITC	Isothermal titration calorimetry
LB	Luria-Bertani
LC-MS	Liquid chromatography-mass spectrometry
LPS	Lipopolysaccharides
MALDI-TOF	Matrix-assisted laser desorption ionization time-of-flight
MAST	Motif alignment and search tool
MEME	Multiple em for motif elicitation
NAI	2-Methylnicotinic acid imidizolide
NMIA	*N*-methylisatoic anhydride
ORF	Open reading frame
qRT-PCR	Quantitative reverse transcription polymerase chain reaction
RBS	Ribosome Binding Site

RNP	Ribonucleoprotein
RT	Reverse transcription
SELEX	Systematic evolution of ligands by exponential enrichment
SHAPE	Selective 2'-hydroxyl acylation analyzed by primer extension
SPR	Surface plasmon resonance
sRNAs	Small regulatory RNAs
TBE	Tris/Borate/EDTA; RNP, ribonucleoprotein
TEX	5' Phosphate dependent terminator endonuclease
UTR	Un-translated region

Naming Conventions

Genes are lowercase (ex. rpoS)

Messenger RNAs are lowercase and italicized (ex. *rpoS*)

Proteins are capitalized (ex. RpoS)

sRNAs are capitalized (ex. DsrA)

CHAPTER ONE: IDENTIFYING AND CHARACTERIZING HFQ-RNA INTERACTIONS[†]

1.1 INTRODUCTION

Traditionally, a protein-centric interpretation of the central dogma has dominated the fields of molecular, cellular, and developmental biology. This idea suggests that RNA serves a predominately intermediate role in the flow of genetic information from DNA to protein, where proteins play the important functions in the cell. While the infrastructural functions of many RNAs (tRNA, rRNA, snRNA, snoRNA) have been recognized the potential role of RNA in regulation, although suggested several times by prominent scientists, was widely ignored or discredited [1-3]. The discovery of protein transcription factors contributed to the lack of interest in RNAs as functional gene end-products [4]. This view of RNA as an intermediated with limited function beyond coding for proteins is still pervasive but the discovery that many RNAs play a role in gene regulation, viral defense and catalysis has gained widespread attention and the traditional definition of a gene and the functions of RNA are changing [5-9].

RNA that does not code for a protein but has a cellular function is generally called a non-coding RNA (ncRNA). Non-coding RNAs exist in all forms of life from bacteria to humans. The amount of the genome that does not code for protein increases as the complexity of the life form increases (Figure 1) [9]. It is widely recognized that, while these non-coding regions are not translated, they are in fact transcribed, leading to an abundance of RNA in the cell. While the nature of these transcripts and their

[†] Portions of this work have been previously published. Faner, M.A.; Feig, A.L.Identifying and Characterizing Hfq-RNA Interactions. *Methods* 2013, in press. The work is reproduced here with permission of the copywrite holder.

ultimate functions is still not completely known many hypothesize that they contribute to the complexity of organisms by expanding the regulation of a repertoire of proteins that is common among them [9].

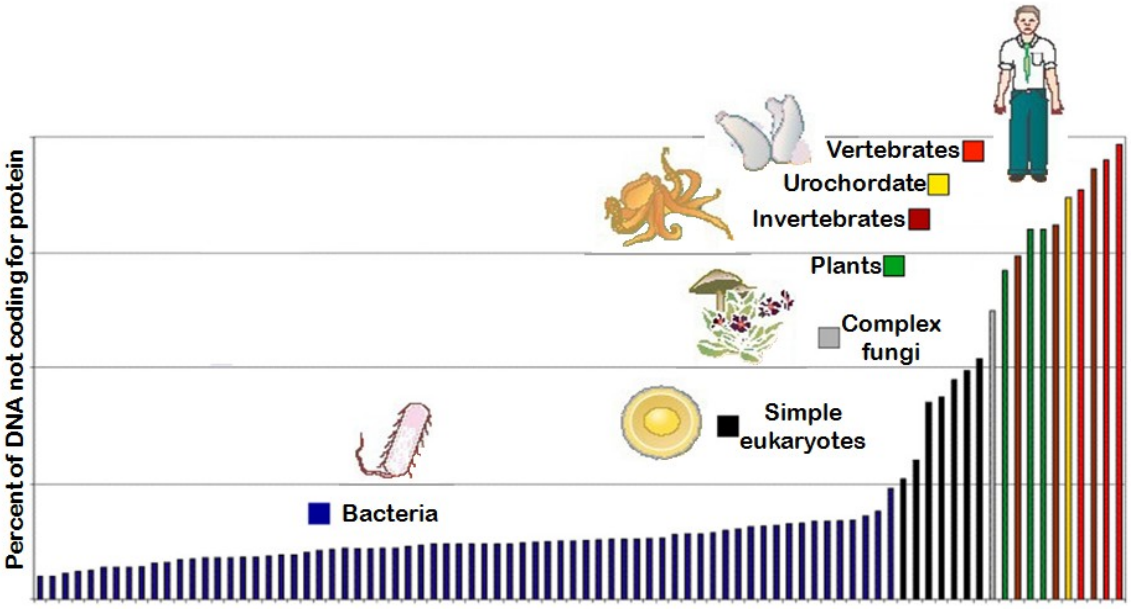


Figure 1. RNAs that do not Code for Proteins. The percent of DNA that does not code for proteins is represented for various organisms. Adapted from [9,13].

The field of ncRNAs is vast and endlessly interesting but our lab has chosen to focus specifically on ncRNAs in bacteria. There are a variety of ncRNAs in bacteria. The infrastructural ncRNAs are a field of their own and therefore will not be discussed further in this work. In addition to tRNA, rRNA, etc. 10-20% of the genes in bacteria code for ncRNAs involved in regulation [10]. Regulatory RNAs in bacteria (sRNAs) are important for the ability of bacteria to thrive in diverse environments and they also play a key role in virulence [11]. Regulatory RNAs can be divided into three main groups: ligand and protein binding, foreign DNA targeting, and base-pairing [6, 11-13]. Our work specifically focuses on base-pairing sRNAs and therefore our discussion will be limited to them.

There are two main types of base-pairing sRNAs in bacteria. *Cis*-encoded transcripts

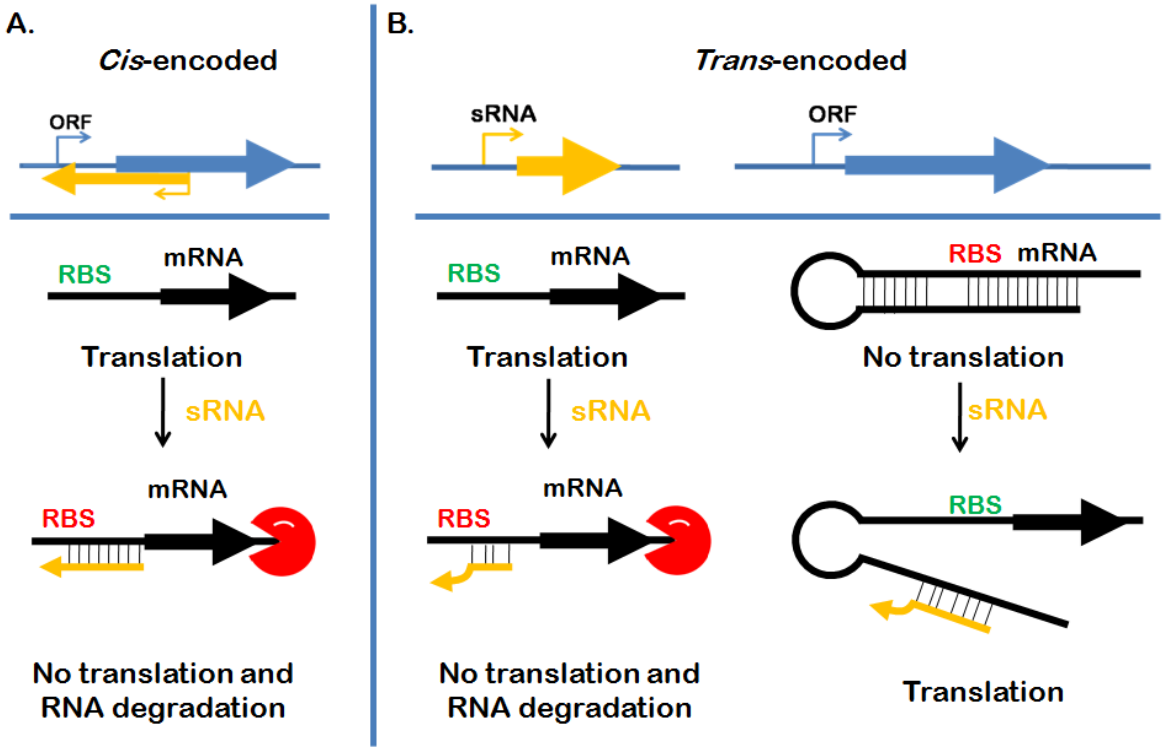


Figure 2. Base Pairing sRNAs in Bacteria. A. *Cis*-encoded sRNAs. These sRNAs are encoded on the opposite DNA strand as the mRNA that they regulate. The sRNA binds to its target with perfect complementarity, which disrupts translation and triggers RNA degradation. **B. *Trans*-encoded sRNAs.** These sRNAs are transcribed from a locus different from the mRNA that they regulate. For down regulation, the sRNA binds to the transcript (often near the RBS) to block translation and/or stimulate mRNA degradation. For upregulation, the sRNA binds to the mRNA which causes a structural change that releases a previously blocked RBS, allowing translation to occur. sRNAs bind to their targets with imperfect complementarity. Adapted from reference 15.

originate from the same locus as the genes or operons they regulate, and have 1:1 correspondence with them (Figure 2A). This class includes riboswitches and natural antisense transcripts. Riboswitches are RNA motifs encoded within the mRNA that modulate gene expression through structural rearrangements in response to a regulatory signal [14]. Natural antisense transcripts are RNAs transcribed from the opposite strand of the gene and act by base pairing with perfect complementarity to their target [15]. Unlike the *cis*-encoded sRNAs, *trans*-encoded sRNAs, which are the focus of this thesis, are transcribed from a different locus than their targets and act

through imperfect base pairing (Figure 2B). In this way they often regulate multiple mRNAs, forming a web of regulatory activities that occur in response to the environment of the bacterium [16]. Often, *trans*-sRNAs act to positively or negatively regulate the translation of target mRNAs by freeing or blocking the ribosome binding site or targeting a message for degradation [17]. Interactions that occur between a *trans*-sRNA and its targets often require the RNA binding protein Hfq [16]. Hfq facilitates these interactions by stabilizing RNA-RNA duplex formation, aiding in structural rearrangements, increasing the rate of structural opening or by increasing the rate of annealing (Figure 3A) [18-21].

Hfq is widely conserved in bacteria and about half of all gram-positive and gram-negative bacteria express it [22, 23]. In the case of hfq mutant or deletion strains, the regulatory effects of sRNAs fail to occur even though the sRNAs are transcribed in response to environmental cues. Phenotypes of these mutants typically include: slowed growth rates, increased cell size, and increased sensitivity to stress [24-26]. Hfq has also been recognized as a virulence factor in many bacteria including *Vibrio cholerae*, and *Salmonella typhimurium* where hfq deletion strains fail to colonize, regulate motility or regulate outer membrane protein expression [23, 27, 28].

Hfq forms a donut shaped homohexamer and has two well characterized RNA binding sites (Figure 3). In *E. coli*, sequences that are A/U rich and typical of sRNAs bind to the proximal surface of Hfq, while A rich sequences typical of mRNAs bind to the distal surface [29-32]. The proximal site was first characterized by a crystal structure of *S. aureus* Hfq bound to AU_5G RNA, which showed that the RNA wrapped itself around the central pore of the protein in a circular manner (Figure 3B) [32]. Biochemical

analyses later showed that Hfq binds to short A/U rich stretches that are preceded or followed by a stem-loop structure sometimes found in a central location of the RNA and

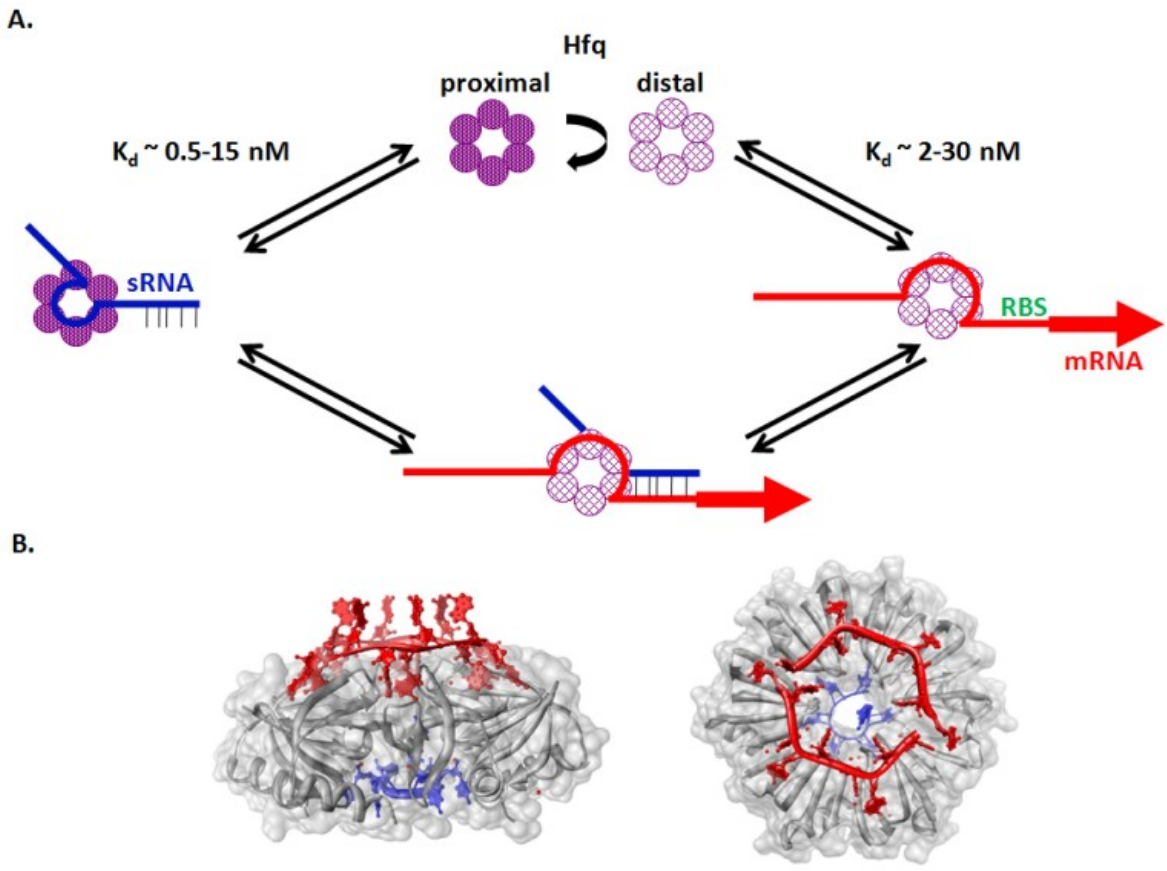


Figure 3. Hfq-RNA Complex Formation and Hfq Binding Faces. A. Hfq binds sRNAs and mRNAs with similar affinity. Hfq may bind either the mRNA or sRNA first before forming the ternary complex. **B.** Crystal structures of Hfq-AU_5G (PDB ID: 1KQ1) and Hfq-polyA (PDB ID: 1HK9) superimposed. AU_5G binds the proximal face and polyA binds the distal of the homohexamer [32, 37].

more recently at the rho-independent terminator [33-35]. The binding of A-rich sequences to the distal face was first defined by a series of mutations that led to disruption of polyA binding [30]. Some years later the specificity for the distal face interaction was further elucidated in a study of the interaction of Hfq with the mRNA *rpoS* as being an AAYAA motif, where Y is a C or a U [36]. These results were further verified by investigation of the interactions of Hfq with two more mRNAs, *fhlA* and *glmS*, genomic SELEX, as well as a crystal structure of *E. coli* Hfq bound to polyA RNA

(Figure 3B). The binding motif has been described as AAYAA, $(ARE)_x$, and most recently as $(ARN)_x$ [31, 36, 37]. The nomenclature for this motif has evolved to $(ARN)_x$ as the binding site was found to be less specific than AAYAA and the acronym ARE was already in use to describe A/U-Rich Elements in eukaryotic mRNAs [31, 38]. Hfq binds to sRNAs and mRNAs with similar affinity. *In vitro*, the order of binding does not appear to matter with respect to formation of tertiary complexes of duplex annealing (Figure 3A). Crystal structures of Hfq from other organisms reveal some species specific RNA interactions. While *S. aureus* Hfq binds A/U rich sequences in common with *E. coli* the distal site binds an (RL) motif, similar to *B. subtilis*, and in contrast with the *E. coli* $(ARN)_x$ motif [29, 31, 39]. The RL motif is a two nucleotide repeat where R is purine specific and L is a non-specific linker. Crystal structures and binding studies of 2 Hfq proteins from cyanobacteria suggest that the proximal site binding of these proteins is not specific for A/U rich RNAs as seen in other bacterial Hfqs. In addition to the well characterized proximal and distal surfaces, the lateral surface and the C-terminal extension also bind to RNA [40-42]. It has been proposed that the lateral surface binds to polyU tracts located in the body of an sRNA while the polyU tract at the 3' end of an sRNA anchors the transcript to the proximal face of Hfq [41]. The role of the C-terminal domain in RNA binding remains murky but structural and biochemical studies suggest that it may bind to longer RNA molecules and/or increase interaction specificity by recognizing additional motifs within an RNA [40, 42].

Identification of Hfq binding RNAs, characterization of their structure and interactions with Hfq, as well as unraveling their functions is fundamental to gaining an understanding of this complex regulatory network in bacteria. The complexity of sRNA

regulation has gradually come into focus over the last decade and now it is clear that this network is indeed vast. The ability of one sRNA to regulate multiple mRNAs and one mRNA to be regulated by multiple sRNAs, as well as the regulation of mRNAs that serve as transcriptional regulators themselves add to the complexity of the network

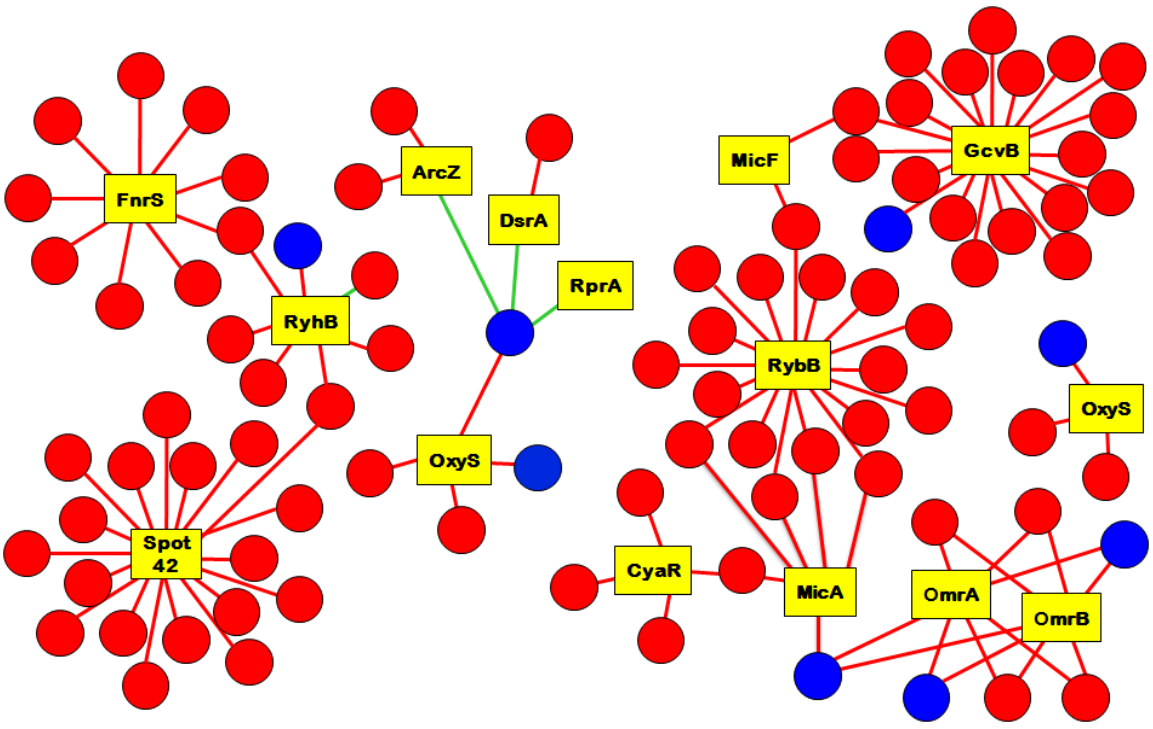


Figure 4. *Trans*-sRNA Network. The network created by a portion of sRNAs is represented above. Squares contain sRNAs and circles contain target mRNA. Red lines are a down regulation and green are an up regulation. Blue circles denote mRNA targets that are themselves transcriptional regulators. Adapted from [20].

(Figure 4) [43-45]. In *E. coli* and *S. typhimurium,* ~30-35 Hfq-binding sRNAs have been discovered and approximately ~ 25% of all *S. typhimurium* mRNAs bind Hfq *in vivo, mak*ing the number of potential RNA binding partners for Hfq in the cell very large [46-49]. Thus despite high levels of Hfq expression, it is believed that the availability of Hfq is often limiting in the cell [50, 51]. There is also evidence that Hfq and/or Hfq-RNA complexes may engage in protein-protein interactions with RNaseE, PNPase, poly(A) polymerase, RNA polymerase, the degradosome and the ribosome; these interactions

provide mechanistic insight but also imply additional complexity with respect to biological function [52-56]. While Hfq is abundant in the cell, observations tell us that it is a limiting factor which is not surprising given the plethora of RNA and protein binding partners possible for Hfq [50, 51, 57, 58]. Still, Hfq is able to coordinate a rapid cellular response to stress, in only 1-2 minutes [59-61]. How is Hfq able to successfully perform this job? While several plausible theories based on current evidence exist, many of which have been recently reviewed [62], it is critical to continue studying Hfq-RNA interactions at three different levels: discovery, biophysical characterization, and functional analysis. Finally, since so much of our understanding comes from a small set of organisms it is important to branch out into other bacterial species to increase our understanding of this complex and fascinating regulatory network.

The goal of this review is to provide a brief overview of some of the key techniques used to investigate and characterize Hfq-RNA interactions and to provide the reader with insight into the strengths of various methods and how they should optimally be applied. We have structured the article as if the reader were new to the field of Hfq-associated regulatory RNAs and needed to know what the fundamental questions are and how to go about answering them. In Section 2, the identification of binding partners will be discussed. The main question here is: To whom does Hfq bind? This section will also include insight into the function of the Hfq-RNA interaction. Section 3 focuses on the biophysical nature of Hfq-RNA interactions. Where do RNAs bind on Hfq and where does Hfq bind RNAs? What is the effect of Hfq binding on RNA secondary structure and duplex formation? What are the relative contributions of thermodynamics versus kinetics in Hfq-RNA interactions? The last section focuses on

questions surrounding the function of Hfq-RNA binding. What are the biologically relevant outcomes of Hfq-RNA interactions and how do they impact the fitness and virulence of bacteria?

1.2 IDENTIFICATION OF BINDING PARTNERS

The first step in studying Hfq-RNA interactions and gaining insight into their regulatory outcomes is to identify the binding partners. Strong binding between Hfq and its sRNA or mRNA partners and the effects of Hfq on transcript and protein levels can be used to identify novel sRNAs and their targets. Three main methods will be discussed: co-immunoprecipitation of RNAs with Hfq, proteomics and transcriptomics in *hfq* knockout strains, and SELEX.

1.2.1 CO-IMMUNOPRECIPITATION

Hfq co-immunoprecipitation (co-IP) is one of the most common methods used to identify Hfq binding RNAs. The co-IP step can be performed by isolating Hfq bound transcripts using an Hfq specific antibody, an epitope tagged Hfq, or by incubating cellular extracts or purified RNA pools with an affinity tagged Hfq. Once the binding partners have been isolated there are several methods for determining which RNAs have been pulled down. Early work used microarrays, shot gun cloning, and enzymatic sequencing [49, 63, 64]. More recently, the advent of inexpensive high-throughput sequencing (HTS) has altered the experimental landscape and is now the most common approach to deconvolute the pull-down components [47, 49, 63, 64]. One of the best features of co-IP is the ability to directly identify Hfq-RNA interactions in a high-throughput fashion, but some limitations occur due to the potential for non-specific

interactions. Another drawback is that the lengthy protocol can result in degradation of large mRNA transcripts.

A critical decision that the researcher has to make concerns the growth conditions of the bacteria. It is important because many transcripts are short lived or only expressed under specific growth conditions and thus may go undetected in one experiment while being highly abundant in another. In order to detect as many transcripts as possible several different conditions should be selected. Some researchers may wish to select a stress condition of particular importance in a pathogen of interest, or a growth phase that is known to exhibit significant expression changes in the absence of Hfq. Whatever the conditions, it is critical to recognize that it is most likely that many Hfq binding RNAs may not be present and thus go unnoticed.

Incorporation of a polyacrylamide gel electrophoresis size fractionation step is another key decision. It depends on whether the goal is to find sRNAs only or to also capture mRNA targets. The feasibility of sequencing a large number of isolated transcripts also plays into this equation. Size fractionation is helpful to enrich for sRNAs as well as to limit the size of the library that requires sequencing. The affordability of HTS makes the latter concern less relevant than in the past. It is ideal to use HTS without a size fractionation step so that both Hfq binding sRNAs and mRNAs are discovered simultaneously.

The choice to use an Hfq specific antibody or an affinity/epitope-tagged Hfq protein for the RNA pull down should be made with the following considerations. An Hfq specific antibody is available for *E. coli* but, to use this technique in interesting pathogens, either the *E. coli* antibody must cross react with that organism's Hfq or a

new antibody must be developed [65]. Sonnleitner *et al.* and Christiansen *et al.* have

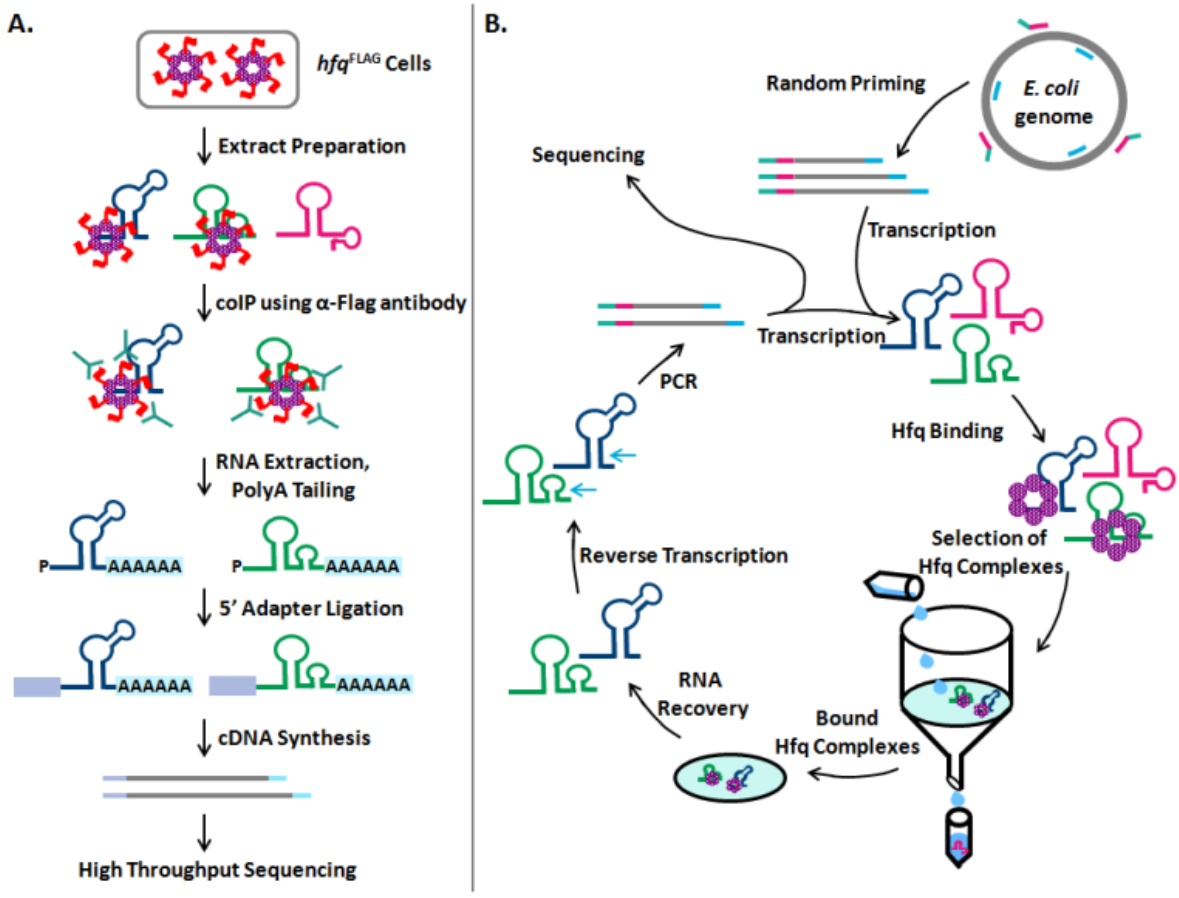


Figure 5. Discovery of Hfq Binding RNAs. A. Co-IP of Hfq bound RNAs using chromosomally FLAG tagged Hfq (adapted from Sittka *et al.*)[30]. Cellular extracts from *hfq*^{FLAG} cells are prepared and co-IP with an α-FLAG antibody is performed. RNAs are extracted and modified to incorporate a polyA tail and 5'phosphate. A 5' adapter is ligated followed by cDNA synthesis and high-throughput sequencing to identify the bound RNAs. **B.** Genomic SELEX (adapted from Lorenz *et al.*)[80]. A genomic library is created by random priming using primers that incorporate a T7 promoter and primer binding sites for reverse transcription and PCR. The library is transcribed to RNA and a binding reaction with Hfq is performed. Bound complexes are selected using filter binding. Bound RNAs are recovered from the filter followed by RT-PCR. The cycle is repeated multiple times followed by sequencing to identify the aptamers.

successfully developed antibodies in *P. aeruginosa* and *L. monocytogenes* for this purpose [64, 66]. The other option is to use an affinity or epitope-tagged Hfq which provides an excellent opportunity to perform this experiment without first preparing an antibody. Ramos *et al.* took advantage of the affinity tag method and discovered 24

novel sRNA in *B. cenocepacia [67]*. They purified a His-tagged Hfq protein that was subsequently incubated with an isolated RNA pool, followed by capture of the Hfq-RNA complexes using Ni-NTA agarose magnetic beads. An epitope-tagged Hfq system was developed by Sittka *et al*., in *Salmonella* in which they created a chromosomal FLAG-tagged Hfq protein (Figure 5A) [47]. To obtain the Hfq bound RNAs they incubated a FLAG antibody with cell lysates and separated the bound from unbound RNAs using protein-A sepharose beads. One thing to keep in mind when using an epitope tag is that its presence may perturb RNA binding and therefore bias the results. The Hfq antibody or the FLAG-tag antibody detection directly from cell lysates provides the benefit of detecting transcripts that were bound *in vivo*.

Once the Hfq bound RNAs have been isolated they can be identified by microarrays, shot gun cloning, enzymatic sequencing, or high-throughput sequencing. A pioneering study used direct detection of bound RNAs on genomic microarrays to detect 20 novel sRNAs as well as a number of mRNAs that interact with Hfq [49]. The sensitivity of this method was unparalleled at the time but required the use of an antibody specific for RNA:DNA hybrids as well as a species specific high density microarray. These features limit its use in other bacteria of interest. Co-IP has also been used in combination with enzymatic RNA sequencing and shotgun-cloning (RNomics [63]) to identify novel sRNAs in *L. monocytogenes* and *P. aeruginosa* respectively. The use of enzymatic sequencing was a success because it identified Hfq binding sRNAs in *L. monocytogenes* for the first time, but it required large amounts of RNA and time consuming sequencing gels making it unsuitable for large scale analyses. Similarly, shotgun cloning was able to identify new sRNAs on a small scale but the lengthy

cloning and use of capillary electrophoresis make it sub-optimal for high-throughput investigations. That being said these approaches are successful and make use of techniques that are readily available at relatively low cost.

The advent and recent affordability of HTS has likely made it the ideal choice for identification of Hfq bound transcripts from co-IP. This method obtains sequence information for a large number of RNAs at one time making it more feasible to identify both mRNAs and sRNA in a wide variety of growth conditions. It does not have species specific requirements so it can be used regardless of sample origin. Also, the alignment of the cDNA clusters can often determine the 3' and 5' ends of sRNAs. This method was implemented by Sittka *et. al.,* in combination with FLAG-tag Hfq co-IP to identify 1,253 mRNAs that were bound to Hfq as well as large number of sRNAs [47]. However, this method, as well as any other protocol involving cDNA synthesis, may have a bias against sRNAs because of the restricted capability of reverse transcriptase to process through highly stable structures [68].

Classification of an Hfq bound RNA as an sRNA or an mRNA is the final critical step in the discovery process. Once the transcript has been identified and its location mapped to the genome several criteria can be used to make the determination. mRNAs are often already annotated in the genomes of sequenced bacteria so assignment as an mRNA is relatively simple. If the species is not annotated, one can look for the classic characteristics of an open reading frame (ORF), including; conserved regulatory sequences, a ribosome binding site, and reduced conservation of the third nucleotide of codons. For sRNAs, there are no hard and fast rules for required features. One seemingly tried and true predictor of an sRNA is an orphan rho-independent terminator

and many searches incorporate this criterion [63, 64, 69]. Although, it should be noted that there are examples in multiple bacterial species of regulatory RNAs that also code for short peptides, recently reviewed by Vanderpool *et al. [70]*. A length component is often incorporated as well. This criterion can be implemented during gel fractionation or when scoring sequencing results and commonly enforces a general size range of ~ 50-500 nucleotides [49, 66]. Genomic location is also typically considered because, historically, sRNAs have been found to be transcribed as standalone transcripts in intergenic regions [47, 49]. This requirement is a good place to start in a novel organism, but the results will not be comprehensive. It has been observed that sRNAs can be derived from the 3' ends of transcripts and from genes coding for tRNAs [46, 69]. So, for an exhaustive search, one should not only look in intergenic regions. Conservation of sRNA candidates among closely related species can also be taken into consideration but can become difficult as the sequences rapidly become disparate. Often, sRNAs involved in metabolic processes will be well conserved among related species but sRNAs found in pathogenicity islands or in cryptic prophages are species specific [69]. Given that most of these rules apply to some but not all sRNAs it is advisable to combine them in a way that can help identify the sRNAs but will not be exclusionary to certain types.

While cross-linking has not been used to pull down Hfq associated RNAs thus far, we would be remiss to neglect the cross-linking and immunoprecipitation (CLIP) assay due to its success in identifying other RNA-protein interactions [71-74]. This method uses UV cross-linking to create covalent bonds between RNA and protein that are in direct contact with one another. Cross-linking at 254 nm occurs due to the natural

photoreactivity of nucleic acids, as well as some amino acids, at this wavelength. A typical CLIP experiment involves *in vivo* cross-linking followed by lysis and partial digestion of RNA. The RNP complexes are purified by co-IP to select for the protein of interest and the bound RNA fragments are identified using high-throughput sequencing. This technique identifies bound RNA and also provides information on the location of the interaction between the binding partners. Chi *et al.* used CLIP to identify miRNA (microRNA) and mRNA binding partners of Argonaute in the mouse brain [71]. Cross-linking provides advantages over co-IP alone. First, cross-linking directly reflects RNA-protein interactions *in vivo* because the bonds are formed in whole cells rather than lysates or purified RNA pools. This process excludes the formation of unnatural complexes due to limiting concentrations of different cellular components that could result in the detection of biologically irrelevant interactions. Second, the RNAs obtained more accurately reflect direct targets because RNAs bound to a protein that associates with the bait protein are not pulled down. Regardless, it is still necessary to validate candidates *in vitro*. A disadvantage of CLIP, especially for potential use in the Hfq system, is the low cross-linking efficiency of purine bases. This caveat may limit the identification of mRNA binding partners. The success that cross-linking has had in the miRNA field makes the CLIP assay a logical candidate for use in the discovery of Hfq associated sRNAs and mRNAs as well as identification of Hfq binding motifs.

1.2.2 TRANSCRIPTOMICS AND PROTEOMICS

Transcriptomics and proteomics provide information on the effects of Hfq on transcription and translation which can lead to the identification of Hfq binding partners as well as insights into function. These methods do not provide evidence for a direct

interaction between Hfq and RNA nor can they distinguish between primary and secondary effects, so interpretation must be performed with care. In addition, some changes in transcript/protein levels may only occur during specific conditions or may be too small to detect, so there is often the potential to miss or overlook important regulatory events.

Transcriptomics in wild type (wt) and *hfq* deletion strains can lead to detection of Hfq binding sRNAs and mRNAs. In these analyses wt and Δ*hfq* cells are often grown under various stress conditions, followed by isolation of total RNA and detection using microarrays or high-throughput sequencing. Transcriptomics identifies RNAs whose transcription or degradation is significantly affected (directly or indirectly) by Hfq [47, 75, 76]. A direct effect occurs when Hfq acts on the transcription rate or decay rate through physical contact with the gene or mRNA. An indirect effect occurs when a change in transcript level occurs due to the action of Hfq on some other DNA, RNA, or protein. Transcriptomic analysis only cannot distinguish between these mechanisms, so it is often coupled with another technique like Hfq co-immunoprecipitation [47]. The growth conditions can also be manipulated to disfavor the effects of transcriptional regulators that are known to be connected to Hfq [77]. Mapping the affected transcripts to the genome identifies the genes and their functions, if annotated, can be suggested. For example, transcriptome profiling of *S. enterica, S. maltophilia*, and *Y. pestis* found that Hfq affects the levels of genes associated with stress response and virulence [47, 75, 76]. Microarrays are also effective to detect transcript levels but they require a high density oligonucleotide array to be available for the bacterium of interest [76]. Roscetto *et al.* have taken advantage of the increased affordability of HTS, *in lieu* of microarrays,

to identify mRNAs that show changes in transcript levels due to Hfq, to predict novel sRNAs and to annotate transcription start sites in *S. maltophilia [76]*. They sequenced RNA from wt and *hfq* mutants as well as in the presence and absence of 5' phosphate dependent terminator endonuclease (TEX). Comparison of RNAs from wild type and mutant strains highlighted changes in transcript levels caused by Hfq, while the TEX treatment allowed them to annotate transcription start sites and identify potential sRNAs. Northern analysis and qRT-PCR can be used to validate observed changes in transcript abundance although it has been noted that the abundance levels measured by qRT-PCR are lower than those obtained in microarray results [75].

Proteomics can be used to characterize global control of gene expression at the post-transcriptional level by monitoring which proteins show significant expression changes in the presence versus absence of Hfq. Examining protein levels can identify targets that are regulated translationally and would have been missed by transcriptome analysis. This approach often uses 2-D gel electrophoresis to identify proteins with differential expression but the technique resolves only a fraction of total protein, so proteins with low abundance, low solubility, or that co-migrate with another species may not be detected. These studies have been done with the intent to find Hfq-sRNA targets rather than both sRNAs and mRNAs. Using MALDI-TOF mass spectrometry, Barra-Bily *et al.* were able to identify a set of 55 proteins with expression differences in an *hfq* mutant in *S. meliloti* [78]. Proteomics alone cannot distinguish between transcriptional and post-transcriptional/translational regulation, but a sample-matched procedure that combines transcriptomics and proteomics can resolve this problem. This method was used in *S.Typhimurium* using half of a culture for RNA isolation and microarray analysis

and the other half for proteomics analysis using LC-MS [24]. Ansong *et al.* compared the change in transcript levels with the change in protein content to distinguish between direct and indirect regulation. They found that the majority of effects in *hfq* mutant strains were due to post-transcriptional events. Proteins from their results were validated by western blot analysis and agreed with previously published results. Another benefit of using simultaneous transcriptomics and/or proteomics approach is that no tagging or isolation of Hfq was required.

1.2.3 GENOMIC SELEX

A significant problem that plagues all of the techniques described above is that they require the RNA to be transcribed at detectable levels under the selected condition. While the use of HTS has made it easier to obtain data from multiple growth conditions, it is unreasonable to expect a researcher to assay all possible growth or stress conditions under which an sRNA could be expressed. A complimentary approach to discover Hfq binding RNAs avoids this issue of growth dependent expression by screening a genomic library for Hfq binding RNAs in a systematic evolution of ligands by exponential enrichment (SELEX) experiment (Figure 5B) [79].

The uncoupling of RNA detection and growth conditions occurs by creating a genomic library via random priming of all endogenous DNA sequences behind a T7 promoter. Transcription of the library yields a pool of RNAs that represents the entire genome of the bacterium; therefore, all possible transcripts are present regardless of growth condition. However, the caveat is that the RNAs start and stop at random genomic positions and do not correlate with actual transcription start sites or termination sites. Once the RNA pool is created, Hfq is added and allowed to bind to its RNA

partners. The bound and unbound transcripts are separated using filter binding and the enriched transcripts are reverse transcribed and amplified. The PCR products obtained are transcribed into a new RNA pool and the selection process is repeated. In their protocol, the whole cycle was repeated 9-10 times at which point the Hfq binding RNAs were sequenced and mapped to the genome [79]. They also verified that the identified aptamers bound to Hfq in a cellular environment by employing a yeast three hybrid assay [79]. SELEX was able to recover many known Hfq binding sRNAs and mRNAs although it missed some of the most well know and prolific species. This oversight may be due to these well known RNAs having a lower affinity than the selected aptamers or to reverse transcription stops as a result of their highly structured nature. It is also possible that some transcripts were overlooked because they were misfolded or amplified in a manner that altered Hfq affinity. A notable result from this study was the large number of aptamers that corresponded to the antisense strand of protein coding genes. This observation differs from the focus on *trans*-sRNAs as Hfq binders. The location of these *cis*-antisense transcripts near start codons and intervening sequences between genes in operons suggests a potential role for Hfq in translation regulation, gene processing and expression within polycistronic messages.

1.3 FOLDING AND INTERACTIONS

Once Hfq's binding partners have been identified, one may begin to consider the nature of these interactions. A large amount of biochemical and crystallographic data are now available to support the identification of RNA binding surfaces on *E. coli* Hfq. It is generally accepted that A/U rich elements (typical of sRNAs) bind to the proximal surface and that (ARN) tracts (typical of mRNAs) bind to the distal surface [30-33, 36].

Existing evidence also supports a role for the lateral surface in binding U-rich sequences found in the body of sRNAs and for the C-terminal extension in binding longer RNA fragments [40-42]. Crystal structures in other organism including *S. aureus* and *B. subtilis* have shown that species specific Hfq-RNA binding occurs [29, 39]. With the discovery of sRNAs and Hfq in pathogenic bacteria as well as their link to virulence, the need to characterize the specificity of binding and the binding surfaces of these Hfq homologs is of particular interest. Crystallographic data provide tremendous insight into these questions but this chapter will focus on biochemical and biophysical techniques that are readily available to a wide variety of labs.

Another question to answer is where does Hfq bind on mRNAs and sRNAs? This question is more difficult because binding sites that have been characterized often have unique features based on the specific RNA studied. This heterogeneity has prevented the formation of an exact definition. A general trend seen in Hfq binding sites on sRNAs is the presence of single stranded A/U rich regions flanked on one or both sides by a stable stem loop structure [33, 80-83]. These motifs have been found in the body of the RNA as well as at the very 3'end of the RNA where it is part of the polyU stretch of the rho-independent terminator [33-35]. The importance of Hfq interactions with mRNAs did not become apparent until recently, so these sites have just started to be defined. However, several well-studied examples provide valuable insight and it has been established that, in most bacteria, the sequence of the binding site is $(ARN)_x$ and it is present in highly structured 5'UTRs of regulated messages [36, 37, 84]. All three of the validated $(ARN)_x$ sites lie to the 5' side and in close proximity to their sRNA binding

sites. When the $(ARN)_x$ motif of these messages is mutated it results in decreased ternary complex formation and dysfunctional regulation.

Duplex formation between an sRNA and mRNA is often central to the regulatory outcome desired in response to stress and environment. Hfq serves to aid in duplex formation by remodeling RNA, by increasing the local concentration of the two RNA molecules, or by increasing the rate of structural opening [18-21]. The effect that Hfq has on duplex formation is vital to understanding how a specific regulatory pair functions. Hfq is an RNA chaperone and it has been proposed to remodel RNAs into more favorable structures for duplex formation. This activity has been shown in some instances and not others; therefore, investigating this possibility in an RNA of interest can provide insight into how Hfq promotes duplex formation [33, 65, 85]. Elucidating the relative contribution of thermodynamics and kinetics to the Hfq-RNA interactions is also important in understanding how a specific regulatory outcome is achieved. In a cellular environment Hfq is abundant but its concentration remains limiting and RNAs have to compete with each other for binding [50, 51, 57, 58]. The ability of a regulatory pair to affect its regulatory outcome is dependent upon its ability to compete for Hfq. This competition is modulated by how tightly and how fast the RNAs associate and dissociate from the secondary and ternary complexes with Hfq.

1.3.1 ELECTROPHORETIC MOBILITY GEL SHIFT ASSAY (EMSA)

EMSA is a very common, easy and adaptable assay that can be used to answer a wide variety of questions regarding Hfq-RNA interactions. The technique is based on the change in migration of RNA upon binding of a protein. Use of P^{32} labeling and phosphorimaging allows for accurate quantitation. The assay can be used qualitatively

to determine whether or not an RNA binds Hfq or quantitatively to allow the determination of thermodynamic and kinetic parameters.

An EMSA should be performed on a native polyacrylamide gel poured at the percentage optimal for migration of the bound complex into the gel and for resolution of the free and bound RNA complex, which is dependent on the size of the RNA. Typical gel percentages are 4-8% and may also contain 3-5% glycerol which can improve complex resolution. The acrylamide:bisacrylamide ratio used is typically 29:1 to accommodate the large size of the Hfq-RNA complex and gels are typically run in 0.5-1X TBE buffer often at 4ºC to stabilize the complex during resolution. While the use of EDTA in the running buffer deviates from the conditions used in RNA conformational studies, we have found it is not detrimental and simplifies the experiment by eliminating the need for buffer recirculation and long running times.

It is important to obtain a homogenously folded RNA population, but due to the complex structure of some sRNA and mRNAs this can be difficult. Multiple folding conditions can be evaluated by changing monovalent salt conditions, magnesium ion concentrations and annealing conditions. Typical monovalent salt concentrations are from 100 mM to 500 mM and magnesium concentrations are from 1 mM to 10 mM. sRNAs with regions of self-complementarity have exhibited the tendency to form homodimers which must be avoided [30, 86]. This tendency can be exacerbated by high magnesium ion concentrations. The RNA should be annealed prior to binding by heating to 75-95ºC followed by a period of cooling. The temperatures and durations vary between labs but we have found that 1 minute at 90ºC in the absence of magnesium

followed by slow cooling at room temperature for 30 minutes works for many RNAs[30, 37, 84].

Binding specificity can be influenced by the salt concentration as well as addition of competitor RNA. It is well known that Hfq can interact non-specifically with RNAs mainly due to electrostatic interactions between the negatively charged RNA and the overall positive charge of Hfq. This non-specific interaction is stronger at low salt concentration. In contrast, specific RNA-protein interactions are less dependent on ionic strength, due to the added stabilization provided by the favorable free energy associated with the specific contacts made. The general outcome is that as salt concentration is increased the interaction becomes more specific and the affinity decreases. This effect has been observed with Hfq as it has been shown that high salt concentrations will decrease the affinity of some sRNAs for Hfq [87]. We have found that the salt conditions used for folding provide a good balance between specific and non-specific interactions. It is common in RNA-protein binding assays to add a competitor RNA to reduce non-specific binding. This addition should be considered carefully in the case of an Hfq-RNA binding reaction, as Hfq has been shown to specifically bind tRNA and poly(A) RNA which may inadvertently alter the measured binding constants [88].

Once the assay conditions are selected, the goal of the experiment should be chosen from several options: the presence or absence of an interaction between the RNA and Hfq can be determined, the effect of Hfq on duplex formation can be assessed or thermodynamic and kinetic parameters can be obtained. If thermodynamics is the focus, equilibrium dissociation constants (K_d) can be determined by titrating an RNA

with increasing amounts of Hfq so that a range of free and bound complexes is present. A trace concentration of ^{32}P labeled RNA that is, at least 10-100 fold below the K_d should be used. The Hfq concentrations should cover two orders of magnitude above and below the K_d and should maximize the number of data points in the binding transition region. Trace RNA conditions simplify the K_d calculation by allowing one to assume that the free protein concentration changes insignificantly. When determining a K_d the binding reaction must be incubated long enough to achieve equilibrium, typically 5-30 minutes at room temperature. For very tight binders longer incubation may be required due to slow off rates. The binding reactions are then combined with loading buffer containing glycerol or sucrose, loading dyes of choice and then resolved. It should be recognized that loading dyes may affect migration of RNP complexes and can be omitted to avoid problems. A drawback of EMSA is that the gel may need to be run for several hours to adequately resolve the complexes.

To obtain thermodynamic parameters the free and bound bands are quantified from the phosphorimage. The percent bound RNA is determined and then plotted versus the log Hfq concentration. These points are then fit by a nonlinear least-squares analysis to a cooperative binding model (for Hfq the cooperativity values typically fall between 2 and 3). Multiple binding events may occur because one RNA may bind multiple Hfq hexamers. This effect can be observed in the case of Hfq binding DsrA (Figure 6) as well as with other RNAs. This case may be dealt with by using a partition function for two sites or by simplifying the data to consider only the K_1 events in which case all shifted bands are summed to yield a “bound” state. The analysis of the gel shown in Figure 3A demonstrates the two site fitting method based on equations 1-3.

$$f_{DH\ I} = cpm_{DH\ I}/cpm_{total} = ([Hfq]/K_1)^n / Q_{DH} \quad (1)$$

$$f_{DH\ II} = cpm_{DH\ II}/cpm_{total} = ([Hfq]^2/K_1K_2)^n/Q_{DH} \quad (2)$$

$$Q_{DH} = 1 + ([Hfq]/K_1)^n + ([Hfq]^2/K_1K_2)^n \quad (3)$$

where $f_{DH\ I}$ and $f_{DH\ II}$ are the fractions of DsrA in complexes D•HI or D•HII, the Hfq concentration is of monomers, K_1 and K_2 are binding constants for the first and second site, n is the Hill coefficient and the function Q_{DH} in (3) is the sum of the terms for each bound state. To obtain a binding constant for each state one simultaneously fits equations (1-3); if only the first binding constant is desired an equation like (4) can be used instead.

$$Q_{DH} = [Hfq]^n / (K_d)^n + [Hfq]^n \quad (4)$$

The K_d determined in equation (4) should have a value similar to the K_1 value obtained from the partition functions in equations (1-3). Some labs use the dual binding fit whereas others use the single site. The choice of which fitting to use is based on perceptions of physiological relevance. The decision is not straight forward as the topic is still debated. The binding of two Hfq hexamers by one RNA may be an effect only observed *in vitro* due to the trace conditions of RNA and the large concentrations of Hfq. This condition may not exist in the cell because of competition for Hfq. A recent study by the Weichenrieder group particularly calls the biological relevance of multiple Hfq binding into question because they found strong evidence that an sRNA binds both the proximal and lateral surfaces of Hfq [41]. If this is the case it is unlikely that one sRNA could bind multiple Hfq protein except under *in vitro* conditions of trace RNA.

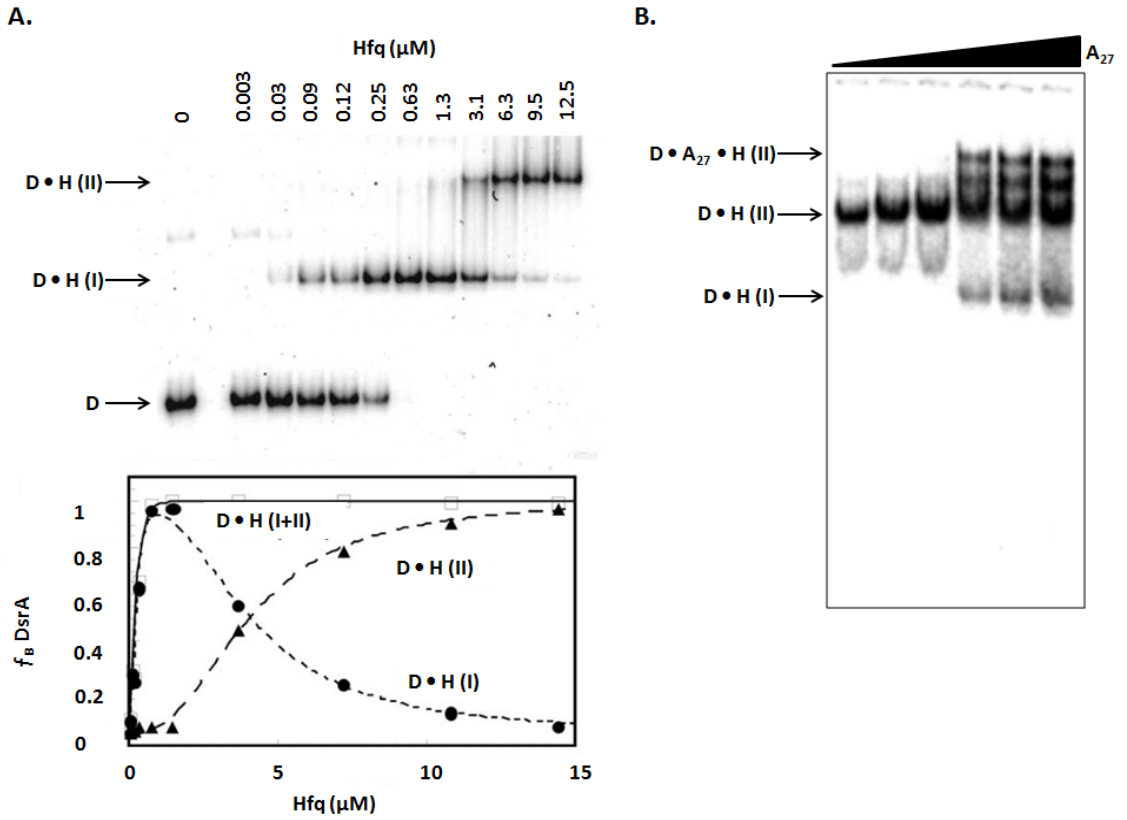


Figure 6. Electrophoretic Mobility Shift Assays. A. Two site fitting used for DsrA and Hfq (image from Lease and Woodson)[87]. ^{32}P labeled DsrA was titrated with 0-12.5 μM Hfq monomer. The D•H (I) band is DsrA bound to one Hfq hexamer and the D•H (II) band is DsrA bound to two Hfq hexamers. Plot of the fraction of DsrA bound (f_B) versus Hfq concentration showing the fitting of each site independently and combined. **B.** Competition assay of A_{27} with pre-formed D•H (II) complex (image from Mikulecky *et al.*)[31]. Titration of increasing concentrations of A_{27} (0-3 μM) leads to the formation of a ternary complex.

A variation of this technique, called a competition assay, can be used to assess which face of Hfq an RNA is binding as well as its ability to bind compared to other RNAs (Figure 6B). This method is particularly useful for determining binding of RNAs by Hfq homologs whose binding specificities have not yet been determined. This approach uses a preformed Hfq-RNA complex which is then challenged with increasing concentrations of a competitor of interest. The ability of a competitor to promote dissociation of the RNA from the preformed complex can then be assessed by

monitoring an increase of the free RNA. To use this assay to gain information about the Hfq binding face specific for an RNA of interest, the preformed complex should contain Hfq and an RNA for which the binding face is known. The ability of a competitor molecule to remove the known RNA from Hfq indicates that they bind the same face and a lack of ability to compete indicates binding on a different area of Hfq. One specific application of this technique was used by Mikulecky *et al.* to determine the RNA binding sites on *E. coli* Hfq (Figure 6B). In that experiment a DsrA-Hfq complex was preformed and unlabeled A_{27} was added at increasing concentrations. As the amount of A_{27} increased the formation of DsrA-Hfq-A_{27} occurred, indicating that Hfq binds DsrA and A_{27} on different faces and that they act independently.

EMSA can also be used in a straightforward experiment to evaluate the effect that Hfq has on duplex formation of a regulatory RNA pair. The Aiba group studied the duplex formation of SgrS and *ptsG* over time in the presence and absence of Hfq to investigate if Hfq could enhance the rate of duplex formation [89]. To explore the effect of Hfq they added Hfq to the binding reaction and then extracted Hfq with phenol before loading the reaction onto the gel. Before treatment with phenol it is advisable to first digest with proteinase K to prevent the RNA from transferring to the organic phase along with Hfq. Within one minute, a significant amount of duplex had formed in the presence of Hfq, suggesting that Hfq strongly enhances the rate of duplex formation. Rapid duplex formation in the presence of Hfq highlights a limitation of EMSA. The time it takes to prepare the samples may exceed the time it takes for the complex to form so one may not be able to quantify fast events, although quench-flow techniques can resolve this issue.

Kinetic parameters can also be determined using EMSA by following binding reactions over a time course. The fraction of each complex can be plotted against time and fit to rate equations. This application was used to demonstrate the effect of Hfq on the rate of DsrA and *rpoS* annealing [36]. In this experiment, the two RNAs were monitored over time for the formation of duplex, in the presence and absence of Hfq. Using this technique, Soper and Woodson showed that Hfq increased the rate of duplex formation ~ 30 fold [36].

The use of EMSA to evaluate the binding of Hfq to truncated RNA constructs has been used to identify the portions of RNA that are necessary for binding. This approach was used to identify the lengths of the 5'UTRs of *fhlA* and *rpoS* required for Hfq binding and sRNA-mRNA duplex formation [36, 37]. In both cases constructs with varying 5'UTRs were made and assayed for their ability to form a duplex. Salim and Feig, as well as Soper and Woodson, were able to determine the relevant 5'UTR length for optimal duplex formation using this approach[36, 37].

1.3.2 FILTER BINDING ASSAYS

Filter binding assays allow for the measurement of both thermodynamic and kinetic properties of Hfq-RNA binding [87]. Unlike EMSA, where complexes are separated in a gel matrix, filter binding employs a double filter to separate the bound from unbound RNAs. The top nitrocellulose membrane binds the RNA-protein complexes and the bottom charged membrane binds free RNA. The two membranes are seated in a dot blot apparatus and samples are drawn through by applying a vacuum. Quantitation of the RNA and RNA-protein complexes is performed using phosphorimaging. Some particular benefits to this assay are the ability run on high-

throughput 96 well plates, to manipulate the volume of the reactions to obtain optimal detection, and its high sensitivity and low cost. This method is particularly useful for determining fast kinetics of binding due to the rapid rate of complex separation [87]. One potential drawback however, is that you can no longer resolve multiple binding events; as discussed earlier these events may or may not be relevant in a given study. Equilibrium dissociation constants can be obtained by titrating the RNA with increasing amounts of Hfq and fitting the data to standard binding isotherms. Kinetic parameters can also be determined by keeping the RNA and protein concentrations constant and varying incubation time. Control experiments in the absence of protein should be performed to account for non-specific nucleic acid binding to the nitrocellulose membrane. This technique was implemented to investigate and compare the binding properties of nine different sRNAs [87]. Olejniczak found that these sRNAs had similar affinities for Hfq but varied in their ability to compete for Hfq binding. The binding properties determined using the filter binding assay agreed with those obtained using other methods under the same conditions.

1.3.3 SURFACE PLASMON RESONANCE

Surface plasmon resonance (SPR) has been used to study the thermodynamics and kinetics of Hfq binding to both sRNAs and mRNAs. SPR monitors changes in the refractive index near the surface of a sensor that occur due to binding events. One strength of this technique is the simultaneous, real time measurements of both kinetic and thermodynamic parameters. In SPR, one binding partner is immobilized on the sensor surface and the other is continuously flowed in. When a binding event occurs, the refractive index increases and when the complex dissociates, the refractive index

decreases. The results are plotted as response units versus time and are most commonly fit to a simple 1:1 Langmuir binding model to obtain k_{on} and k_{off} values. The K_d can then be calculated by dividing the k_{off} by k_{on}.

There are several steps that must be taken in order to execute a successful SPR experiment investigating an Hfq-RNA interaction. Though theoretically it shouldn't matter, it is most typical to immobilize the RNA on the surface of the sensor and flow in Hfq. The larger size of Hfq provides a greater change in response when the two molecules interact [90]. Also, the negative character of the chip surface can repel the RNA if it is chosen as the analyte [91]. For a high affinity interaction like that of Hfq with an RNA, the RNA should not be immobilized at too high a concentration or problems associated with mass transfer can arise [90]. We have found that ~ 3 fmol works well in the case of *fhlA*. To prepare the RNAs for SPR, they are biotinylated at the 5' end and purified using a spin column. It is critical that the samples are very pure as the presence of contaminants could affect the SPR signal or interact with the analyte and impact binding. A benefit of SPR is that it is a label free approach but it does require immobilization which could lead to changes in binding. Unfortunately, this technique is not suitable for high-throughput analysis as only a few samples can be analyzed at a time and each experiment requires 5-15 minutes.

This approach has been used to analyze the kinetics and thermodynamics of Hfq binding to the mRNAs *fhlA* and *ompA* and to the sRNA MicA [18, 37]. The Wagner group used SPR in addition to EMSA and filter binding to obtain K_d values, and association and dissociation rates for *ompA* and MicA [18]. In both cases the values obtained were similar between the three techniques demonstrating the value of each in

obtaining reliable data. In the study of *fhlA*-Hfq, SPR was used to demonstrate that the $(ARN)_x$ motif is important for distal site interactions and to support a wrap-around model for *fhlA* binding. This model suggests that the RNA binds to both surfaces of Hfq at once [37]. To investigate the importance of the $(ARN)_x$ site contact with Hfq, the ability of constructs with or without the site to interact with Hfq were compared. It had previously been shown that *fhlA* interacts with both Hfq surfaces, so the data was fit to a parallel binding model where *fhlA* can interact with either side of Hfq independently before forming the complex where both sites are bound. The step where both sites are bound was omitted from the fitting because the technique cannot register that type of unimolecular rearrangement. The inability of SPR to detect internal rearrangements of this type is its shortcoming. The *fhlA* construct that contained both the proximal and distal site had two low nanomolar K_d values whereas the construct with only the proximal binding site had only one, indicating that the $(ARN)_x$ site is important for distal surface binding. Salim and Feig also performed a competition experiment by pre-binding Hfq to *fhlA* and then flowing in A_{18}, DsrA, or both RNAs. All three scenarios led to faster than direct k_{off} rates (with no competitor RNA) which suggests that *fhlA* binds in a wrap-around fashion. These experiments highlight the use of SPR to obtain information beyond thermodynamic and kinetic parameters.

1.3.4 OTHER BIOPHYSICAL METHODS

Isothermal titration calorimetry (ITC) is most widely recognized in studying DNA-protein interactions and protein biophysics but has also been successfully used to obtain thermodynamic information and binding stoichiometry of an RNA and protein interaction [30]. ITC directly measures the heat released or absorbed during a chemical

reaction by monitoring the power consumption required to keep a sample cell and a control cell at the same temperature over the course of a reaction [92]. Direct measurement allows for a more accurate determination of thermodynamic data than a gel shift. Some issues that have limited its usefulness are the need for a large sample as well as the inability to deconvolute the energy parameters from multiple binding interactions or structural rearrangements.

Fluorescence anisotropy measures the change in polarized light emitted from a fluorophore in solution during a binding event [93]. This change is a result of decreased tumbling of the labeled molecule upon binding of a larger molecule. This phenomenon allows for the evaluation of a molecule's binding properties by providing a direct measure of the bound to free ligand ratio. Fluorescence is a safer option than radiolabeling but it is less sensitive and bulky which may affect binding. A benefit of this approach is that it is solution based which omits a separation step and therefore may more accurately reflect true equilibrium binding. This approach can be applied to Hfq-RNA systems by labeling the RNA molecule with 6-carboxyfluorescein, titrating it with increasing amounts of Hfq and observing the change in anisotropy [94-97]. The data are plotted as anisotropy versus time and fit by a nonlinear least-squares analysis to a two step binding model.

Fluorescence anisotropy was used to investigate the RNA binding surfaces on Hfq in a similar fashion as EMSA. Sun and Wartell assessed a variety of Hfq mutants followed by binding studies with RNA substrates [95]. In agreement with previous studies, they found that DsrA binds to the proximal surface and that A_{18} binds to the distal surface of Hfq. They also used the fluorescence anisotropy data to determine

reaction stoichiometries which led to some ambiguities regarding the binding ratio of the A_{18} Hfq interaction. Uncertainty in the amount of Hfq required to saturate binding of the labeled RNA caused an underestimation in the amount of bound RNA, leading to a discrepancy with ITC data. This incongruity was later resolved by allowing flexibility in the variable that accounts for the fraction of bound RNA [96]. Determining an accurate binding model of other than two state reactions can be challenging using fluorescence anisotropy if the anisotropy change between the two states is not well defined and/or if there is cooperative binding. EMSA is typically more suitable because of the added information provided from visualization of discrete bands that represent different complexes. These observations can provide binding stoichiometry and guide the correct selection of a binding model.

1.3.5 CHEMICAL AND ENZYMATIC RNA MODIFICATION

The use of chemical and enzymatic analysis of RNAs can be employed to determine the secondary structure of an RNA, the Hfq footprint, and structural changes upon Hfq binding. Additionally, some techniques allow structure determination and protein interaction mapping *in vivo*. One approach uses a complementary set of enzymatic and/or chemical modifications that react with the nucleotides in different ways to provide a complete assessment of each nucleotides environment. To determine Hfq binding sites on the RNAs, the probes can be used in the presence and absence of Hfq. In the presence of Hfq some nucleotides will become protected, indicative of a binding site. In addition to seeing protection, some nucleotides may become more reactive, indicative of secondary structure rearrangements. Two different methods can detect the cleavages or modifications. One route uses reverse transcription with an end labeled

primer to detect both scissions and modifications for RNAs of any length (by using multiple primers). The other technique uses end labeled RNA for direct detection but can only be used for shorter molecules, typically less than 300 nucleotides in length. The fragments obtained from these methods are then separated on denaturing polyacrylamide gels along with one or two ladders that assign the site of cleavage/modification on the RNA. Efforts to obtain a uniformly folded RNA as well as the selection of binding buffer conditions should be taken as discussed for EMSA.

One illustrative example of the use of nucleases was the determination of the effect of Hfq on the *sodB* mRNA and its regulatory partner, RyhB sRNA [98]. Geissmann *et al.* used a combination of RNaseA, which cleaves 3' to single stranded cytosines and uracils; RNase T1, which is specific for single stranded guanines; RNase I, which cleaves any single stranded residue; and RnaseV1 that is specific for double stranded regions and provides positive evidence for helical regions. This probe combination allows for sufficient coverage of the RNA to provide an accurate secondary structure. RNases are large and therefore show signs of steric hindrance and care must be taken optimize enzyme concentration and incubation time as the presence of secondary cleavage events can lead to misinterpretation of the data. The data obtained allowed for the accurate determination of secondary structures of the two RNAs as well as footprints pinpointing the Hfq binding site(s). Also, the occurrence of enhanced cleavage at certain residues in the presence of Hfq can show a loop opening event or other rearrangements such as in the case of Hfq binding to *sodB* mRNA.

Another useful probe are the Tb^{3+} or Pb^{2+} ions, which cleave single stranded RNA in a sequence independent manner. The small size of these ions avoids the steric

hindrance issues that RNases have, which allows for detection of subtle structural changes upon Hfq binding. The Masse lab used this method to detect an enhanced interaction between the regulatory pair, RyhB and *iscS*, in the presence of Hfq and the Hfq binding site on the *iscS* mRNA [99]. The concentration of the ion must be optimized to obtain conditions where less than one cleavage occurs per RNA molecule. The reaction can be quenched at the optimized time with addition of EDTA. Lead(II) has also been used to determine secondary structures *in vivo* and could potentially be used to map Hfq interaction sites *in vivo* in the future [100].

Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) is a chemical modification based technique that takes advantage of the ability of the hydroxyl selective electrophile, N-methylisatoic anhydride (NMIA), to react without sequence specificity with more flexible/accessible nucleotides [101]. The use of SHAPE to determine secondary structures and footprinting provides the advantage of only having to use one chemical modification technique to obtain necessary structural information. Modifications are revealed by reverse transcription and resolution on denaturing gels or by capillary electrophoresis. Capillary electrophoresis analysis allows for a significant increase in throughput and software is available to analyze the raw data and obtain reactivities for each nucleotide [102, 103]. Our lab has successfully used SHAPE in combination with capillary electrophoresis to determine the secondary structures and Hfq footprinting of *glmS* and *fhlA* mRNAs [37, 84].

Several considerations are important to successfully implement SHAPE to study RNA-Hfq interactions including: RNA design, RNA folding, RNA modification and primer extension conditions. To detect adduct formation reverse transcription (RT) is used. RT

can lead to a loss of information due to pausing at the 3' end during the initiation phase and at the 5'end because of an intense band equivalent to the full length extension product. To avoid this loss of information, the RNA can be inserted into a structured cassette, first described by the Weeks lab, where the RNA is flanked by highly structured hairpins and also an RT primer binding site on the 3' end [101]. While the cassette improves the read through, it may still interfere with mapping Hfq binding sites at the polyU tract of the rho-independent terminator since the 3' end is unnatural in these constructs. The stability of the hairpins ensures that the cassette structure does not interfere with folding of the RNA of interest. To facilitate analysis of many RNAs we have created a modified pUC19 vector containing the cassette behind a T7 promoter so that any RNA of interest can be cloned into the vector and transcribed. The RNA must be renatured prior to modification as described for the previous techniques.

To modify the RNA, NMIA is added at a concentration and time that must be optimized to obtain only one adduct formation per molecule. NMIA +/- reactions are run in parallel so that natural RT stops can be accounted for in the data analysis. In order to obtain footprinting data, Hfq +/- reactions can be run as well. Hfq is added to a final concentration of 1 μM hexamer and allowed to bind at room temperature for 30 minutes before reaction with NMIA. After NMIA reaction the RNA is ethanol precipitated or, in the case of Hfq + reactions, it is first digested with proteinase K and then extracted with phenol-chloroform prior to ethanol precipitation. The primer extension reaction is performed using RNA Superscript III in four separate reactions: NMIA +, NMIA -, and two sequencing ladders created by including ddNTPs into the reaction mixture. Each

reaction contains an RT primer with a unique fluorophore that allows identification of the

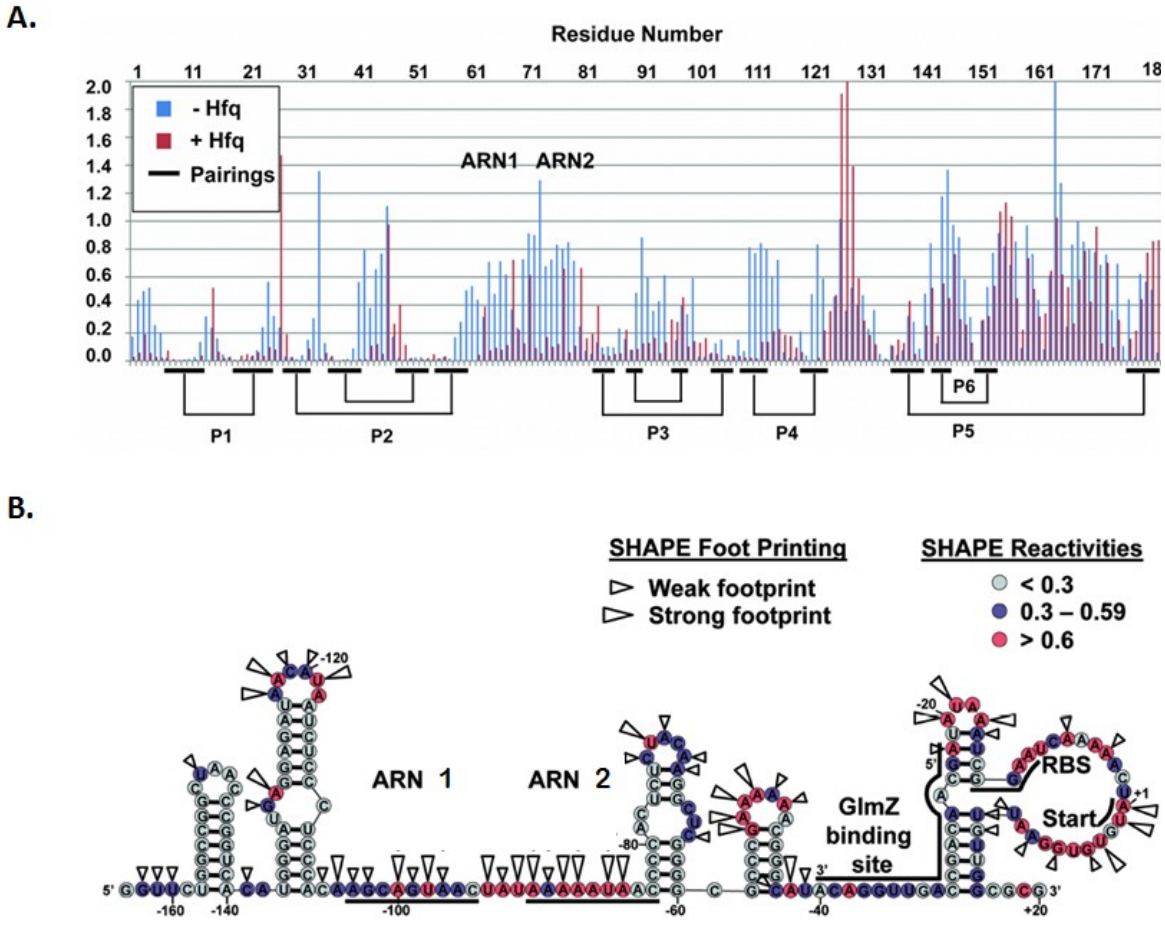


Figure 7. SHAPE Derived Structure and Hfq Footprinting of *glmS*. A. Normalized reactivities for each nucleotide in the presence and absence of Hfq (image from Salim *et al.*)[85]. Double stranded residues are indicated by P1, P2, etc….. and Hfq binding $(ARN)_x$ sites are indicated. **B.** Schematic of the SHAPE derived secondary structure with reactivities and Hfq footprint superimposed. RBS is the ribosome binding site and GlmZ binding site is the binding site of the regulatory RNA. Footprints were deemed weak if Hfq binding resulted in a reactivity change between 0.3-0.59 and strong if the change was > 0.6.

different reactions in the capillary electrophoresis readout.

The reaction is then separated by capillary electrophoresis. Reactivities for each nucleotide are determined by analyzing the raw data with ShapeFinder (Figure 7A) [104]. Data for Hfq + reactivities are obtained from a unique set of reactions that can be

run in a parallel lane. The reactivities for the Hfq +/- reactions can then be compared to determine where protection has occurred (Figure 7). The resulting data is used to determine the fold of the RNA using RNAstructure and the Hfq protection can be mapped (Figure 7B) [105]. The structures for *fhlA* and *glmS* that we determined using this approach added to the evidence for an important Hfq binding interaction at $(ARN)_x$ sites in the 5'UTRs of regulated messages [37, 84]. This method provides accurate, high-throughput structure determination and footprinting. The cost of fluorophore labeled primers is high but the use of a universal RNA cassette makes it a worthwhile one-time investment.

New developments in SHAPE that describe high-throughput analysis and *in vivo* structure mapping have recently been published [106, 107]. These techniques have not yet been applied to bacterial sRNA systems but hold promise for investigating Hfq-RNA interactions. Lucks *et al.* recently described high-throughput SHAPE analysis that is able to obtain structural information from an *in vitro* pool of RNAs that are distinguished from one another using bar-codes [106]. This method is able to obtain quantitative high resolution structure information for hundreds of RNAs in a single experiment. It is important to study RNA *in vivo* because the biologically relevant structure may exist only in the cellular environment. In addition, RNA-protein interactions are represented in the data. Chang *et al.* designed two new electrophiles, 2-methyl-3-furoic acid imidizolide (FAI) and 2-methylnicotinic acid imidizolide (NAI), which maintain the selective reactivity to hydroxyl groups but are non-toxic and have a sufficient half life in cells to modify RNAs *in vivo*. They found that NAI had a higher reactivity and chose to use it to validate their technique by probing 5S rRNA in mouse embryonic stem cells and in

yeast. When SHAPE data was overlaid with the crystal structure, they found that NAI had modified the RNA at the predicted nucleotides in as little as 1 minute. Comparison of *in vitro* and *in vivo* SHAPE structures for the 5S rRNA led to the identification of important contact sites with other RNA and proteins. One can easily imagine this technique being used to map the structures of mRNAs and sRNAs that interact with Hfq and to determine Hfq binding sites *in vivo*. Some potential complexity lies in separating the effect of protein binding and structural changes on the reactivities and declaring the identity of the protein binding partner.

1.3.6 ISOENERGETIC MICROARRAY MAPPING

Microarray mapping is a unique approach to secondary structure determination, Hfq binding site identification, and Hfq derived structure change. This technique is based on the ability of single-stranded RNA regions to hybridize with complementary oligonucleotide probes in contrast to double stranded RNAs [82]. A microarray with probes specific for the RNA of interest is created to match the probe specifications required for the particular target. The structure of the RNA is determined alone and then various complexes can be studied by comparing the hybridization intensity in the presence of other complex components. The incorporation of locked nucleic acid and modified nucleotides are incorporated where necessary to account for varying thermodynamic stabilities of the probes due to the specific sequence. This technique can be used with a broader set of conditions than with chemical and enzymatic assays that often require specific conditions for reactivity. The method is limited by the thermodynamic stability of the target molecule structure and the stability of its interaction with other biomolecules. This approach has been used to determine the

structure of DsrA in complex with Hfq and *rpoS,* and OxyS in complex with Hfq and *rpoS* or *fhlA* [108]. Fratczak *et al.* obtained structures for both sRNAs that agreed with previous data and identified previously suggested Hfq binding sites. They were also able to confirm that the DsrA secondary structure is not altered upon Hfq binding and that Hfq facilitates sRNA-mRNA duplex formation. The broad application of this technology has been minimized because of the large amount of effort that must be invested to create a unique microarray for each RNA of interest.

1.4 FUNCTIONAL CHARACTERIZATION

The number of Hfq dependent sRNAs identified in various bacterial species is large but only a small set have well defined biological functions. Bacterial sRNAs are not easily grouped into categories that indicate their functions and consequently, the function of these sRNA regulators often have to be elucidated on an individual basis. Many of the techniques discussed in Section 1 to identify Hfq binding mRNAs also give some information about function if the gene has been annotated. In addition to those techniques we will present approaches that allow for the identification of the RNA binding partner, given an Hfq associated sRNA or mRNA of interest (Table 1). Binding partner identification is often the first step after an initial discovery technique. After identifying potential RNA partners it is necessary to validate that the interaction is direct and that there is a real biological effect.

1.4.1 BIOINFORMATICS

Due to the availability of many bacterial genomes, bioinformatics approaches for the discovery and analysis of sRNAs have flourished. There are many ways that computational tools can be employed to help elucidate the functions of Hfq binding sRNAs and mRNAs, specifically by aiding in the prediction of an RNA binding partner of

Technique	Strengths	Weaknesses
Bioinformatics	• Increase efficiency by guiding lab work • Adaptable to species specific characteristics	• High error rate • Requires experimental validation • Prior knowledge required
sRNA Over Expression	• Study sRNAs that normally have low expression • High throughput screening	• Can disrupt the balance of the sRNA network • sRNAs not present in library will be missed • Does not distinguish direct vs. indirect effects
sRNA Pulse Expression	• Help discern direct vs. indirect effects	• Can disrupt the balance of the sRNA network
sRNA Knockouts	• High throughput phenotyping using bar-codes • High throughput screening for mRNA regulators	• Pleiotropic effects from gene knockout • Resolving changes in target expression

Table 1. Overview of Approaches for RNA Binding Partner Identification.

a given sRNA or mRNA. The most useful aspect of these approaches is the ability to guide lab work to obtain results in a more efficient manner. This guidance saves time and money in the lab. Computational approaches are often not sufficient on their own due to false positives and fake negative feedback and therefore must be validated experimentally. In addition, prior information about the system to be studied is necessary to create a useful tool. These tools have been successful in organisms where Hfq binding sRNAs and mRNAs are well characterized and have the potential to be

modified easily to accommodate species specific characteristics of the network. Bioinformatics can be a particularly useful tool when studying pathogens or bacteria that are hard to grow and manipulate in the lab. These studies could be facilitated by an initial computational analysis followed by experiments in a model bacterium.

One of the earliest examples of employing bioinformatics to identify a target mRNA was a simple BLASTn search to identify a 16 nucleotide region of complementarity between MicC and *ompC [109]*. These searches are useful in identifying interactions that have long regions of continuous complementarity which is unfortunately a minority of Hfq-dependent sRNAs. Despite this limitation, Jørgensen *et al.* have very recently used BLASTn to identify an mRNA target of McaS after a proteomics/transcriptomics approach failed, demonstrating its utility as a starting point for RNA binding partner identification [110]. This approach is also useful because it requires no prior knowledge to guide the search beyond the requirement of complementarity between the two RNAs. Another relatively simple bioinformatics approach is to look for the presence of a transcription factor binding motif. The transcription of some sRNAs is controlled by transcription factors [60, 111]. By identifying the transcription factor that controls expression of the sRNA, the function may be apparent based on the role of the transcription factor. For instance, Papenfort *et al.* were able to identify two σ^E-dependent sRNAs involved in *omp* mRNA regulation using this method [60].

Once a set of targets for a given sRNA have been validated, the knowledge of those interactions can guide a computational search for new mRNA targets [112]. Sharma *et al.* first defined a binding motif for GcvB based on 16 known target binding

sites using the MEME (multiple em for motif elicitation) software [113]. By providing the sequences of the know targets the program was able to identify an 8 nucleotide long motif that was present in all but 2 of the mRNAs. To identify previously unknown targets of GcvB the motif was used to search the -70 to +30 regions of all annotated *Salmonella* protein coding genes using a MAST (motif alignment and search tool) [113]. Frequently, sRNAs interact with mRNAs in this region of the 5'UTR, but this parameter should be chosen based on the known targets of the specific sRNA of interest. Widening this criterion may lead to more false positives. The annotated transcription start sites of the genes should also be taken into consideration. If the interaction was found from -60 to -70 but the RNA is transcribed starting at -50 then the putative interaction is likely irrelevant. The genes that showed a significant match to the motif were then input into *TargetRNA [114]* to identify the targets that had the strongest base-pairing with GcvB. Overall they obtained 42 potential mRNAs that passed all of the bioinformatic criteria; 4 of the 5 that they chose to validate showed regulation. This technique successfully identified known and new targets that were missed by a transcriptomics approach and demonstrated the utility of a combining bioinformatics with other experimental approaches. A drawback of the method is that a large amount of previous knowledge is needed to train the computational queries, limiting its use in finding interactions for sRNAs that have few known targets or in organisms where sRNAs are not well characterized.

In addition to designing your own unique search strategy, there are many accessible programs that have been designed to allow researchers to easily perform bioinformatics studies without designing their own algorithms. These programs and their

detailed methods were recently reviewed by Li *et al.,* so we will just provide a brief overview of a few programs here [115]. *TargetRNA* was designed to identify potential mRNA partners given the sequence of an sRNA and the genome of interest. mRNA-sRNA interactions are scored based on the hybridization between the two RNAs without considering intramolecular base pairing or pseudoknots. The omission of the secondary structure of the RNAs is a limitation because the presence of these structures can significantly affect the likelihood of an interaction. The program also provides parameters that can be specified by the user such as seed length and the location of the interaction site relative to the promoter. Overall their approach was able to identify ~ 70% of the RNAs used in the training set. *TargetRNA* was one of the first programs developed to predict sRNA targets in bacteria. It was designed using a limited amount of known information which may make it less useful than some of the newer programs. That being said it has been successfully incorporated into several recent studies [112, 116, 117].

Many other programs have become available to aid in the identification of mRNA-sRNA interactions. sRNATarget was developed by Zhao *et al.* by incorporating 35 positive (validated interactions) and 86 negative targets into its training set. Unlike TargetRNA this approach also considers the secondary structure of the RNAs [118]. They were able to obtain a greater accuracy rate for predicting the training set than *TargetRNA*. The program IntaRNA evaluates RNA-RNA interactions using a complex algorithm based on hybridization and accessibility of the target site [119]. This program is effective but is computationally demanding, whereas an alternate server called RNApredator achieves similar accuracy in less time [120]. The program sTarPicker has

been shown to outperform the above methods in target prediction and accuracy of binding site prediction. sTarPicker uses a two-step hybridization model that first picks targets based on stable seed interactions and then on extended hybridization of the entire binding site [121]. All of these available tools can aid in the discovery of Hfq associated RNA binding partners when combined with other techniques and therefore contribute to the determination of their biological functions. The current searching methods will continue to evolve as new information about the sRNA regulatory network is learned. Some insights that may improve predictions are to include requirements for Hfq binding sites in the mRNAs and to focus on known binding motifs for particular sRNAs.

The Collins group took a unique approach to define the functions of bacterial sRNAs by using network biology to take advantage of existing microarray data to elucidate the functions of sRNAs [122]. Knowledge of sRNA interactions can often lead to clues about the function of the sRNA. This is the first program to take advantage of the large body of known interactions to make functional predictions for the whole sRNA network [122, 123]. First they applied a Context Likelihood of Relatedness (CLR) algorithm to a compilation of existing microarray expression profiles that were obtained under various conditions [124]. This algorithm identifies regulatory relationships using an inference approach and identified 459 potential targets. They were then able to identify functional enrichment in seven sRNA subnetworks by assigning gene functions to the putative targets. They validated the functional implications of three of these sRNAs. This technique is useful because there are several sRNAs known to regulate multiple mRNAs who all function in a similar physiological process [112, 125]. The

identification of that process allows new targets to be inferred based on their involvement in that pathway. This approach was based on microarray data and therefore does not distinguish between direct and indirect interactions. The incorporation of proteomics data would be an improvement. In general, the more data included in the computational analysis, the greater the predictive power. This method can easily be adapted to other organisms with profiling information.

1.4.2 MANIPULATION OF sRNA EXPRESSION

A widely used approach for defining the function of an sRNA is to manipulate its expression. Many variations including over expression, pulse expression and knockouts have been used to identify the mRNA targets of an sRNA or to identify the sRNA regulator of a given mRNA or phenotype. The basic concept behind these experiments is that changing the expression of an sRNA will lead to detectable changes in transcript levels, protein levels or changes in phenotype.

Creating an sRNA over expression strain involves cloning the RNAs into a high copy plasmid behind an inducible or constitutive promoter. It is necessary to place the transcription of the sRNA under control of an alternative promoter because some sRNAs will not be highly expressed under their natural promoters even when present in a high copy number plasmid. The high copy number expression minimizes the effect of any chromosomally derived sRNAs. The sRNA should be inserted such that transcription begins at the natural transcription site which can be determined using a technique such as 5'RACE if not known [126]. This approach allows for the study of sRNAs that may be poorly expressed naturally or are toxic to the bacterium. A caveat of

sRNA over expression is the potential to cause inadvertent consequences by disrupting the balance of the natural sRNA network, which can lead to confusing results.

Given an sRNA of unknown function, a good way to begin characterizing it is to determine the identity of proteins that show changes in expression when the sRNA is over expressed. The Wagner group used this approach to identify the regulation of *ompA* by MicA [127]. They observed differences in protein expression from strains with high, normal or low MicA expression using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). Proteins that showed changes of greater than 2.5 fold between the strains were subsequently identified using MALDI-TOF. The OmpA protein showed the greatest change in abundance and was subsequently validated as a MicA target. This method was also used by Frohlich *et al.* to classify SdsR as a regulator of *ompD [111]*. In this case a significant change in OmpD expression was identified from a simple 1D-PAGE analysis due to its characteristic size, and then verified by northern and western blots. Proteome analysis does suffer from the inability to differentiate between direct or indirect effects, and mRNA stability or translational regulation as the mechanism of control.

In organisms where the majority of sRNAs have been discovered, an sRNA over expression library can be created to screen the effects of a large number of sRNAs on a given mRNA or phenotype. The utility of this approach was demonstrated by the identification of an additional sRNA that regulates *rpoS* [128]. An sRNA library with 26 Hfq binding sRNAs was created and co-transformed with an *rpoS*-lacZ fusion. The β-galactosidase output was monitored for significant increases or decreases and led to the identification of four sRNAs previously unrecognized to regulate *rpoS*. By observing

the ability of two of the putative sRNAs to act on *rpoS* in strains where the positively acting sRNAs were deleted, they were able to determine that the effects produced in the original screen where indirect. In the deletion strains down regulation of *rpoS* no longer occurred. This observation illustrates the need to be aware of effects caused by artificially titrating Hfq from natural sRNAs and target mRNAs, which can be an issue during sRNA over expression. Mandin and Gottesman went on to confirm the regulation of *rpoS* by the sRNA ArcZ. A useful feature of this approach is that once an sRNA library has been created it is easy to rapidly screen any target mRNA of interest by simply cloning it into a fusion vector.

The Gottesman lab further used the sRNA over expression library to identify sRNAs involved in cell motility by inducing the sRNAs and observing their behavior on motility plates [129]. They identified 8 sRNAs that had a recognizable effect on bacterial mobility. Once sRNAs associated with motility were identified they hypothesized which mRNAs would be logical targets based on a relationship to the phenotype studied. They reasoned that the most efficient means of regulating motility would be the genes at the top of the cascade. They tested this hypothesis by examining the effect of the sRNAs on mRNA-lacZ fusions for the genes of interest and found several legitimate regulatory pairs that they further characterized.

sRNA over expression libraries are a useful way to rapidly screen potential targets for direct interactions with sRNAs but they require prior knowledge of the majority of sRNAs in an organism. The Gottesman lab has used a multicopy plasmid library of the whole genome that negates the need to know the sRNAs in an organism [130]. They identified sRNA regulators for two genes involved in antibiotic resistance

that were found to bind Hfq. They constructed a *lacZ* fusion of their gene of interest, *dpiB*. They transformed a pBR322-based *E. coli* genomic library, into a strain carrying this *dpiB-lacZ* fusion, under the assumption that any genomic element in this library would produce white (Lac-) colonies on MacConkey Lactose plates containing arabinose is they regulate *dpiB* [131]. The observed several colonies with this phenotype and sequenced the corresponding plasmids. They found several fragments of protein coding genes as well as two known sRNAs. If the sRNAs are unknown in the organism of study than one can deduce that the fragment may be an sRNA, based on the typical characteristics of sRNAs discussed previously. The relevance of the sRNA-mRNA predicted is then further characterized and validated.

Pulse expression is a technique that makes use of an inducible promoter to briefly over express an sRNA in a strain where that sRNA of interest is deleted; this process is followed by transcriptome analysis [59, 112, 132]. By over expressing the sRNA for a short time, the differences between direct and indirect effects are more discernible. In this experiment, sRNA expression is induced and total RNA is extracted at a specific time point, usually 10-15 min, and analyzed on a microarray. High-throughput sequencing or qPCR could also be used to analyze the RNA pool. The pulse duration should be optimized to successfully distinguish between indirect and direct effects as different systems may work faster than others. For example, a direct target of RyhB was degraded in 3 minutes and the mRNA of an indirect target followed closely behind at 7 minutes; if analysis was done at 10 minutes the two effects would have been indistinguishable [59].

This technique has identified additional targets of both RyhB and GcvB [59, 112]. The GcvB study incorporated an added feature that allowed identification of mRNAs that bind to a specific motif of the sRNA [112]. Several sRNAs possess a conserved mRNA binding motif that binds to multiple targets [81, 83, 133, 134]. Performing pulse expression with a wild type strain, in addition to a strain where the binding motif is ablated, can help identify mRNAs that have expression changes due to interaction with that conserved binding motif. Because this technique monitors transcript levels it is important to recognize that only targets who show a decrease in mRNA levels will be detected. sRNAs that function primarily by translation inhibition will yield negative results.

sRNA knockout libraries can also be used to establish the targets of sRNAs and the phenotypes associated with them. One way to create the knockout strains is to disrupt the chromosomal sRNA genes by insertion of a drug resistance cassette facilitated by λ phage recombination. The mutation can be confirmed by PCR amplification using primers flanking the recombination site. One of the most significant problems facing this technique is the potential to disrupt neighboring genetic elements which could muddle the interpretation of the observed effect. To combat this problem a cassette with a transcriptional terminator to prevent read through into downstream genes can be used or homologous regions to the flanking genes can be incorporated into the cassette ends. Another way to create a library is to incorporate bar-codes into the insertion cassettes so that individual strains can be pooled for phenotype studies and the sensitivity or resistance of each strain can be identified using microarray analysis. This approach allows rapid phenotyping of a large number of strains. An issue

that one could encounter when using knockout libraries is the inability to resolve changes in target expression due to naturally low sRNA expression levels. In addition, many different growth conditions may need to be tested in order to observe a regulatory event. The success of this approach for sRNAs that also encode a peptide, like SgrS, may also be limited because it would be difficult to associate the outcome with the sRNA rather than the peptide [135].

An sRNA knockout library was successfully implemented to identify a regulator of *ompX* in *S. typhimurium* by Papenfort *et al.* [132]. They observed that *ompX* was associated with Hfq in two different studies (one in *E. coli* and one in *S. typhimurium*) which led them to suggest that it is regulated by sRNAs. Also, previous work that indicated the conservation of 35 *E. coli* sRNAs in *S. typhimurium* led them to include the homologs in their library [27, 49, 136]. Incorporation of homologs could prove useful for other bacteria where sRNAs have not yet been identified but where computational methods have predicted homologs of known sRNAs. To screen the library for sRNAs that affect the expression of OmpX, they grew the relevant knockout strains and performed western blots to compare the amount of protein present compared to the wild type strain. They observed a significant increase of expression in one knockout strain indicating a specific regulatory effect. A useful validation experiment that they performed was to complement the knockout strain with a plasmid carrying the sRNA to observe the return of normal regulation. Overall they were able to find an sRNA regulator of *ompX* in *S. typhimurium* and further characterize it in their study.

Jin *et al.* used a similar approach to identify an sRNA regulator associated with a specific phenotype [137]. They compared the ability of sRNA knockout strains to

recover from acid stress to identify sRNAs associated with the acid stress pathway. A valuable control that they performed was to only delete genes adjacent to the sRNA that showed a sensitive phenotype to verify that it was the sRNA deletion, not disruption of adjacent genes, which caused the effect. Next they sought to determine the target of the identified sRNA, GcvB. Neither of the previously know GcvB targets played a role in the phenotype but they did find evidence using an *rpoS*-lacZ fusion construct that suggested that GcvB up-regulated *rpoS*. Curiously, they were not able to identify any complementarity between the two RNAs. This result highlights the fact that interpretation of phenotype screens can be precarious as phenotypes may arise due to any number of regulatory events not necessarily canonical sRNA mediated effects.

Bar-coded deletion libraries have been used to assess protein coding genes in *E. coli*, and in yeast but Hobbs *et al.* was the first to tailor this approach to sRNAs [138]. Using this technique, they identified the Hfq-dependent sRNAs RybB and MicA as important in cell envelope stress. An advantage of this approach is the ability to phenotype a large number of deletion strains simultaneously. Homologous recombination was used to insert an antibiotic resistance gene in the place of the sRNA. Uniquely, they incorporated bar-codes distinct for each deletion so that they could be identified and quantified by microarray. They also incorporated common primer binding sites to be used for amplification before microarray analysis. To identify which genes were associated with cell envelope stress phenotypes they grew overnight cultures of all of the strains individually and then combined them for stress challenge. The genomic DNA was purified and the barcodes amplified followed by hybridization to a commercially available microarray. Signals must be corrected to account for the non-

linear relationship between actual bar-code concentration and fluorescence signal to obtain relative abundance [138]. Relative abundances that are significantly increased compared to non-stressed cells indicate a resistance to stress and decreased signals indicate sensitivity to stress. The strains that exhibited phenotypes are then assessed in a one on one stress challenge and complementation experiments are performed to verify the results for a select group.

1.4.3 VALIDATION USING REPORTER GENE FUSIONS

The techniques presented so far in this section have aimed to elucidate the functions of Hfq binding RNAs by determining their RNA binding partner in a relatively high-throughput fashion. The nature of these techniques can lead to false positives or to the identification of regulation that is occurring by a mechanism other than base-pairing with Hfq-dependent sRNAs. It is therefore necessary to perform any number of validation techniques on an individual basis to determine if there is a direct base pairing interaction between the RNAs, if regulation occurs by affecting mRNA stability or by blocking translation, and if the process is Hfq dependent. Several classic approaches can answer these questions, including, northern blot analysis, toe-printing, structure mapping and mutational analysis [110-112, 127, 128, 130, 132]. Reporter gene fusions have become a popular way to validate sRNA-mRNA interactions as well as biological significance and will be the focus of our discussion below.

Fusions of mRNAs with *lacZ* and *gfp* allow monitoring of direct effects of an sRNA on the target regardless of the natural transcription level of the mRNA in a given condition. The constructs usually include the 5' UTR, starting from the annotated transcription start site through approximately 10 codons. This region is incorporated in

frame with the reporter gene and is under the control of an inducible promoter [139]. It should be noted that interactions further into the coding sequence have been observed. If a previously seen regulatory event does not occur the length of coding sequence included in the fusion may need to be extended [111]. An inducible promoter is preferable to the natural gene promoter so that the effect on translation instead of transcription does not have to be verified in additional experiments. Any other non-sRNA related regulatory regions of the mRNA should be removed or otherwise accounted for in order to draw clear conclusions about the regulatory outcome caused by the sRNA. The fusion can be created chromosomally or in a plasmid. The chromosomal fusion more accurately reflects natural gene expression but is more time consuming. A low copy plasmid fusion can provide a similar effect and is a simpler and less time consuming strategy.

To assess the effect of an sRNA on the reporter gene a plasmid containing the sRNA of interest is transformed into to the strain harboring the fusion. Transcription of the sRNA can be under control of a constitutive promoter or an inducible promoter but it is important that both mRNA and sRNA are expressed at the same time. Uncoordinated transcription of regulatory partners can disrupt regulation by sequestering Hfq. Translation output of *lacZ* constructs can be quantified using a β-galactosidase assay to determine the activity of the enzyme. This assay is somewhat more time consuming than measuring fluorescence in the GFP assays. GFP expression can be monitored by colony fluorescence, Western blots with an α-GFP antibody, cell lysates, and in whole cell liquid cultures. Measuring colony fluorescence is easy but it is less sensitive and not as quantitative. Measuring from whole cell cultures omits a lysis step and can save time

but measuring from a lysate can increase reproducibility and doesn't require a flow cytometer. The Western blot is the most time consuming but can detect fusion protein levels even if fluorescence signal is low.

The *lacZ* reporter is the most traditional option and it has been successfully used to monitor Hfq-dependent sRNA regulation in many cases [127, 130, 140]. In the studies discussed below the constructs were used to verify regulation observed in a large screen, to definitively show a direct interaction between the two RNAs and to validate the binding site of the RNAs. Mandin *et al.* created a chromosomal *dpiB-lacZ* fusion to reproduce regulation by RybC that was observed in a genome wide screen that they performed [130]. They went on to verify the computationally predicted RNA-RNA binding site by performing mutational analysis. Three point mutations were incorporated into the predicted RNA binding site of the plasmid borne sRNA and its ability to affect β-galactosidase activity of the mRNA fusion was monitored. They observed that the fusion was no longer regulated and they were able to restore regulation by introducing compensatory mutations into the mRNA fusion. This assay unequivocally demonstrates that a direct interaction between the two RNAs is responsible for regulation and defines important residues involved. Udekwu *et al.* performed a similar compensatory mutational analysis but ablated six nucleotides thought to take place in RNA binding [127]. Depending on the RNA pair the number of nucleotide mutations necessary to destroy regulation may differ but care should be taken to ensure that the mutations do not cause significant secondary structure changes that could contribute to the observed outcome.

The use of GFP fusions has become popular more recently to study the regulatory outcomes of Hfq dependent sRNA-mRNA pairs. The Vogel lab established this approach by studying several regulatory pairs in a uniform manner using the two plasmid system and was able to reproduce all previously observed regulation [139]. Since then it has been used in several applications. Papenfort *et al.* performed a compensatory mutation analysis to pin point the interaction site between *ompX* and CyaR [132]. The GFP fusion assay has also been used to assess the ability of a series of mRNA constructs of different lengths to be regulated by an sRNA partner [111].This experiment allows one to obtain preliminary information about the location of the RNA-RNA binding site. Nine potential targets obtained in a pulse expression experiment were confirmed in a more high-throughput manner using GFP fusions as well [112].

Our lab has also incorporated this technique into the detailed characterization of *glmS* regulation (Figure 5) [84]. After characterizing the interaction of *glmS* and Hfq using EMSA and determining the *glmS* secondary structure using SHAPE, the GFP assay was employed to further investigate the importance of the Hfq binding site on *glmS*. We incorporated point mutations into the predicted Hfq binding site on *glmS* and monitored the ability of Hfq and GlmY/GlmZ to regulate the message. Based on GFP expression, we showed that the Hfq binding site was critical for regulation of *glmS* [84]. The assay that we used involved co-expression of the *glmS-gfp* fusion and the sRNA plasmids upon addition of arabinose. Arabinose conditions were optimized to obtain GFP expression levels necessary for observing regulatory changes. The experiment was performed in triplicate using independently grown overnight cultures that were diluted on the following day to start the assay. Cells were harvested in early stationary

phase and lysed using TritonX-100 in the presence of protease inhibitor and lysozyme. The lysate was centrifuged and the supernatant was assayed for GFP signal using a multi-label plate reader. The fluorescence signal was normalized to an identical culture where GFP was not induced.

1.5 CONCLUSIONS

In this introductory chapter, we have discussed various methods to discover Hfq binding RNAs, to characterize their interactions and to investigate their functions, with the goal of serving as a guide to select the best suited techniques for individual systems and questions. New techniques, such as high-throughput sequencing, CLIP, and *in vivo* and high-throughput RNA structure probing promise new discoveries on the horizon. The existing knowledge from model systems can help pave the way to investigation of sRNAs and Hfq in pathogens that could serve as potential therapeutic targets. Whether you are starting from square one in an organism where Hfq and sRNAs have not yet been characterized or you are interested in a specific regulatory pair in a well-known system that you wish to understand better there are many tools to guide your query. The significance of the roles that Hfq-associated sRNAs play in coordinating gene regulation has never been more obvious and there is no doubt that we will be greeted with even more surprising features and roles as we continue to study these fascinating systems.

Many of the techniques described in this chapter have been used in our lab and will appear in the rest of this thesis. This introduction serves to introduce the reader to the plethora of techniques that were available to us, why we chose the specific approach that was taken and the experimental conditions used. The following chapters

describe the characterization of Hfq binding sites in known target mRNAs and the use of those features to predict novel targets. Known targets were investigated using mfold, SHAPE, and EMSA[101, 141]. We developed a bioinformatic approach to identify novel mRNA targets based on the presence of an $(ARN)_x$ site. This approach incorporated existing bioinformatic tools, mfold and IntaRNA, as well as a genome wide sequence searching tool custom made by our lab (Swett and Feig, unpublished data)[119]. We validated our predictions using SHAPE, EMSA, and GFP fusion constructs[139]. By understanding the techniques employed to investigate RNA-Hfq interactions the reader will be able to clearly understand the logic and utility of the work described throughout the rest of this document.

CHAPTER TWO: PREVALENCE AND CHARACTERISTICS OF mRNA-HFQ BINDING SITES IN *E. coli*

2.2 INTRODUCTION

Developments in the field of bacterial *trans*-encoded sRNAs have progressed in a lopsided direction towards discovery and characterization of the regulatory sRNAs; the mRNA targets have taken a back seat. The discovery of a high affinity Hfq binding site within the 5' UTR of several target mRNAs, which is required for regulation to occur, has enforced the notion that both RNAs in this regulatory equation have equal, albeit different, importance. It is imperative to study the interactions of mRNAs with Hfq in order to better understand how the regulatory network functions. A result of this historical disparity is that the rate of target mRNA discovery has lagged behind that of sRNAs, leading to significant under identification. Contributing factors for this imbalance include incomplete knowledge about base pairing rules, location of sRNA binding sites, and what types of conditions lead to specific occurrences of different sRNA-mRNA interactions [142]. Identification of target mRNAs leads to characterization of sRNA functions, often linking them to other previously well defined regulatory pathways, and is crucial to the understanding of the overall sRNA regulatory network. Previous approaches used to identify target mRNAs include microarray, translational gene fusions, co-immunoprecipitation, and bioinformatics, all of which have contributed in important ways to the discovery process but leave room for improvement [142]. The characterization of a specific Hfq binding site in combination with existing techniques may aid the discovery of target mRNAs by adding another parameter with which to search.

The distal binding site of Hfq was first described by Mikulecky *et al*. using mutational analysis [30]. Mutations on the distal surface led to a decrease in binding to poly(A) RNA. Years later, Soper and Woodson identified an AAYAA sequence motif that was required for Hfq binding in the 5' UTR of the *rpoS* mRNA (Figure 8) *[36]*. This sequence was broadened to $(ARN)_x$ when a crystal structure of *E. coli* Hfq bound to poly(A) RNA was solved by Link *et al.* [31]. This motif was also identified as an Hfq binding sequence in a genomic SELEX experiment [79]. Two more examples of $(ARN)_x$ motifs in mRNAs have since been characterized and it has become clear that Hfq binding motifs play an important role in facilitating regulation [37, 84]. The requirement for $(ARN)_x$ sites in these well illustrated examples demonstrates the need for

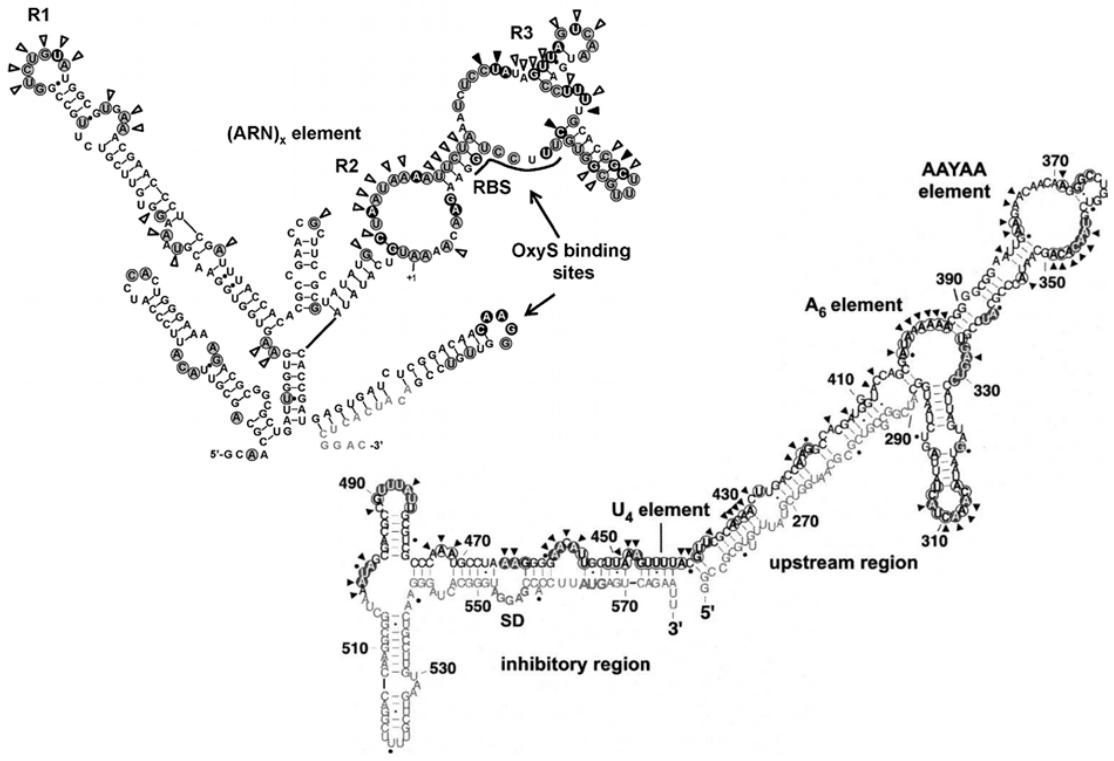


Figure 8. $(ARN)_x$ Motifs in Target mRNAs. Motifs that have been characterized in the 5'UTRs of *fhlA* (above)[38], and *rpoS* (below)[37] are shown. The motif in *fhlA* is denoted as $(ARN)_x$ and in *rpoS* as AAYAA. Both structures were determined using SHAPE and have Hfq foot printing data superimposed.

researchers to determine the appropriate UTR length to use in model systems. In the past only the sRNA interaction site was considered necessary, leading scientists to use incomplete constructs and obtain misleading results [65, 86]. The *glmS, rpoS*, and *fhlA* mRNAs all contain single stranded $(ARN)_x$ sites in their highly structured 5'UTRs (Figure 8). In addition, there may be a connection between the sRNA binding site and the Hfq binding site. Hfq binding sites have been found anywhere from less than 20 to 80 nucleotides from the sRNAs binding site [36, 37, 84, 98]. Panja and Woodson investigated the idea of proximity between the sites using model RNAs and found that the most effective Hfq binding sites were located to the 3' side of the sRNA site and within 20 nucleotides [143]. The nucleotide distance could be overcome when structure brought the two sites spatially closer. For efficient regulation of Spot42 targets, the Storz lab found that the Hfq and sRNA binding sites could not be overlapping [144]. Further studies of known targets are necessary to determine a specific requirement for proximity, if any.

The ability of Hfq to bind to mRNAs, sRNAs and other proteins leads to the reality that even though Hfq is abundant in the cell it is a limiting factor. This effect can be observed as Hfq sequestration in the presence of over-expressed RNAs or mis-matched sRNA/mRNA partners that lead to disruption of the sRNA network [50]. It is of great interest how Hfq is able to facilitate such a rapid, 1-2 minute, response to stress in the complex cellular milieu. The concentrations and binding affinities of different RNAs for Hfq may provide a tuning mechanism for the network by allowing one response to dominate over others when necessary. The presence and strength of $(ARN)_x$ binding sites in mRNAs likely play an important role in this dynamic; potentially one where Hfq

binds to mRNAs to mark them for regulation if and when the cognate sRNA is transcribed, resulting in a rapid and specific response to stress.

The discovery of sRNAs in *E. coli* is nearly complete but the process in other organisms is in its infancy [145, 146]. Homology searching is one approach that has been used to characterize sRNAs beyond *E. coli* [126, 147, 148]. This technique works well for core *trans*- sRNAs in closely related bacterium. These sRNAs tend to be involved in the regulation of processes central to cellular homeostasis and are therefore relevant for many species. On the other hand, variable sRNAs are often involved in virulence and can be located in pathogenicity islands, which makes the identification of these genes by homology search unsuccessful [149, 150]. Even the more conserved core sRNAs are only maintained throughout a single class of bacteria; for example, GcvB is found throughout γ-Proteobacteria [151]. The target interaction regions of sRNAs exhibit a higher degree of conservation than the rest of the molecule [152, 153]. This phenomenon is most obvious in sRNAs with multiple targets. sRNAs with a single target are less constrained and can co-evolve with their target, which is sometimes evident as compensatory changes in the sequence of the interaction regions. Interestingly, mRNAs do not demonstrate any significant degree of conservation at their interactions sites [153]. It is possible that regulation is conserved as well as the sequence of the interaction site but the location can change. Both sRNAs and mRNA targets show conservation of the accessibility of the interaction sites [152, 153]. The limited conservation of *trans*-sRNAs indicates that they are a rapidly evolving class of gene regulators, which makes it difficult to determine how they initially evolved and also how they will continue to change in the future. The lack of sequence and structural

homology also makes it difficult to use known sRNAs in *E. coli* to discover new sRNAs in more distant species such as *clostridia* and *pseudomonads.* Functionally homologous *trans*-sRNAs that have little to no sequence conservation but facilitate similar responses and act on similar targets have been identified. For example, the sRNA FsrA in *B. subtilis*, was shown to down-regulate similar targets as RyhB in response to iron starvation as has been observed in *E. coli* [154]. So while the conservation of sRNAs is limited evidence suggests that the regulation of some target genes may persist even if is it by a different sRNA. If the target mRNAs can be more thoroughly identified in *E. coli,* then they could be used as starting points to search for targets and sRNAs in more distant organisms.

The discovery of an important Hfq binding motif in target mRNAs has shifted the attention of the field, which was once dominated by the study of the sRNAs, to their targets. In addition to the role that $(ARN)_x$ motifs play in the dynamics of the sRNA network we envision their use as a bioinformatic tool. While the majority of sRNAs in *E. coli* have been discovered, the number of known targets is well below the predicted total. By determining the characteristics common to $(ARN)_x$ motifs in known mRNA targets they can then be used as search criteria to identify new targets. Not only will this analysis identify targets in *E. coli* but also in other organisms. This approach can be modified to account for species specific differences in Hfq binding to identify target mRNAs. It is also likely that targets identified in *E. coli* are also targets in other bacterial species and can be used as a starting point for target and sRNA identification in those organisms. Using bioinformatics as an initial tool in the discovery process can guide laboratory experiments in a productive and efficient manner. Previous computational

approaches for target discovery have focused on the features of sRNA-mRNA binding [115]. This method has been honed and works fairly well in *E. coli* where the abundance of sRNAs are known but it is no longer useful when trying to apply it to other species were that knowledge does not currently exist. The use of an Hfq binding site as an identifier of mRNA targets has never been used before and mitigates the need to know sRNAs. Our novel approach would allow researchers to identify targets in a wide variety of organisms *in silico* and then test the predictions *in vivo* in their organism of interest or in *E. coli* as a model system.

2.3 RESULTS

To better understand the common features of $(ARN)_x$ motifs we examined mRNAs that are known to be regulated by sRNAs in an Hfq-dependent manner. Known sRNA binding sites and $(ARN)_x$ sequences were mapped onto computationally predicted secondary structures for a set of mRNA 5' UTRs. The results are presented schematically (Figure 9) and show that most of the 5' UTRs fold into highly paired structures and have multiple single stranded $(ARN)_x$ sites that could be accessible for Hfq binding. The role that multiple sites in an mRNA play is unknown but one can imagine that they may increase Hfq binding thereby giving it priority over other targets. Alternatively, in targets that are regulated by multiple sRNAs, a unique $(ARN)_x$ for each sRNA might be required. Another possibility is that they bind Hfq in order to recruit other proteins to the RNA. Many of the sRNA binding sites are located near the start codon. This position facilitates the role of many of them in modulating expression by interacting with ribosome binding sites.

A total of 45 mRNAs known to be regulated by sRNAs in an Hfq-dependent

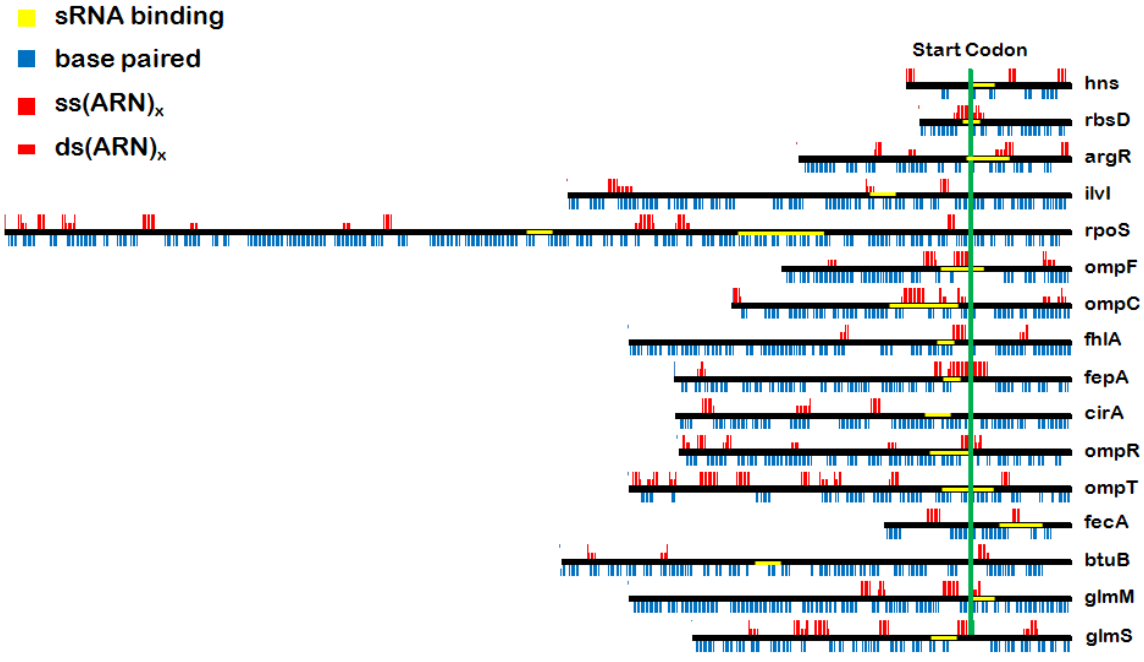


Figure 9. Context of (ARN)x Elements in 5'UTRs of Regulated mRNAs. Structures of known messages were predicted by mfold. Positions of the known sRNA binding sites are shown relative to the start codon.

manner are annotated in the TarBase database [155]. In order to expand the characterizations of $(ARN)_x$ sites from above, we analyzed these RNAs in a systematic way to note their collective features (Figure 10). Our approach starts by noting the presence of the sequence $(ARN)_x$ where x is 2 or more within the annotated transcription start site (or -200 if not annotated) and +60 relative to the start codon. Next we determined if the sequence was single stranded and what the specific structural context was, as determined by computational folding. We focused on the structure of the sequence for two reasons; one, the motif must be single stranded in order to be accessible for Hfq binding; and two, specific context may increase the specificity of the sequence. The probability of an $(ARN)_2$ sequence existing is once every 64 nucleotides therefore an additional structural requirement may be necessary to selectively target the

correct mRNAs. We found that all but one of the mRNAs had a single stranded $(ARN)_x$ sequence and that the majority were located in regions of complex structure like hairpin loops or multi-branch junctions (Figure 11). The one mRNA, *fecD*, is transcribed as part of an operon and therefore does not have its own, annotated, transcription start site therefore we used -200 when determining the computational fold. This may not represent the actual 5' UTR and could have negatively impacted the accuracy of the predicted structure. We hypothesize that the actual 5' UTR of this message has a single stranded $(ARN)_x$, but it was missed due to a folding error. The presence of a single stranded $(ARN)_x$ motif in the 5' UTRs of almost all targets known to be regulated suggests that it is a common feature for target mRNAs.

While investigating known target mRNA we observed that one of the mRNAs analyzed contained a discontinuous $(ARN)_x$ in a hairpin loop. This is an $(ARN)_x$ where

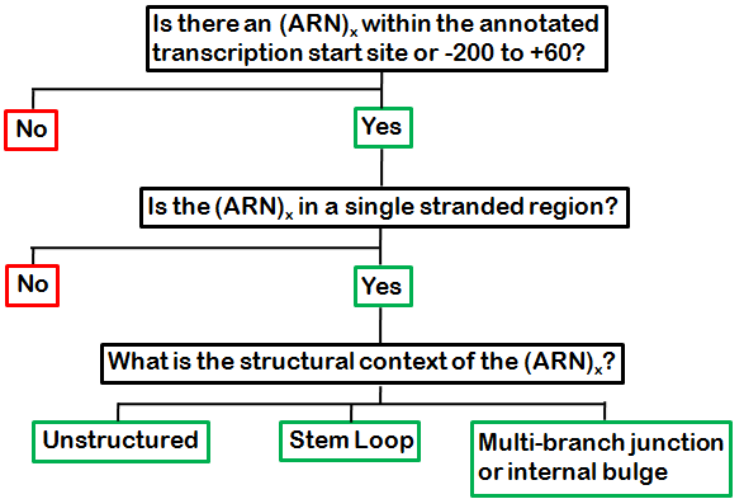


Figure 10. Bioinformatic Approach to Analyze $(ARN)_x$ Motifs. The presence of the sequence $(ARN)_x$ (where X=2 or more) within the transcription start site (or -200) to +60 of the 5'UTR of the mRNAs was noted. These mRNAs were further analyzed to determine if the sequence was single stranded and, if so, its specific structural context. Structural observations were made using computationally predicted folds.

the repeats are interrupted by one or two nucleotides. A discontinuous $(ARN)_x$ may be able to bind Hfq because an extra nucleotide added 3' to the N site does not negatively

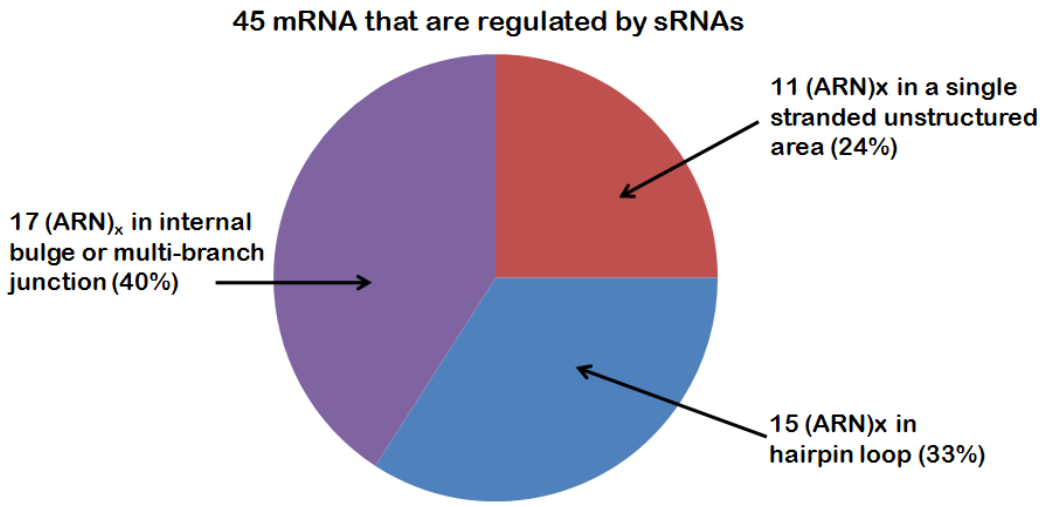


Figure 11 Bioinformatic Analysis of Known mRNA Targets. mRNAs known to be regulated (as annotated in TarBase) were analyzed according to the schematic in Figure 10. A discontinuous site is an $(ARN)_x$ where the ARN repeats are interrupted by one or two nucleotides.

affect binding affinity [31]. Also, because of the repeating nature of the binding motif on the distal face of Hfq, if one motif is skipped it seems likely that the next could still be bound by an ARN. To investigate this possibility, a native gel mobility shift experiment was carried out using an *in vitro* transcribed model of the hairpin loop observed in the *dps* mRNA (Figure 12). Three models were constructed; one with a discontinuous repeat as observed in the natural RNA, one that contained G nucleotides instead of A in repeats 1 and 3 and one that contained G nucleotides instead of A in all 4 repeats. Gels shifts were performed with the three constructs in addition to A_{18}, which is a known distal face binder (Figure 13). All of the constructs bound with approximately 10-fold less affinity than did A_{18}, indicating that discontinuous $(ARN)_x$ sequences do not bind Hfq well. Therefore, these types of sequences should not be included in the search for potential mRNA targets.

The R and N sites of the repeat decrease the specificity of the sequence and increase the likelihood of obtaining false positives when used in predictions. This phenomenon led us to determine if there was any observable nucleotide preference at those sites in the pool of known mRNA targets. A frequency logo [156] was created from the $(ARN)_x$ sequences found in the 45 mRNAs. The logo shows that the R position is more often A than G and that the N position is more often an A but this propensity is not large enough to imply conservation at the N site (Figure 14a). We also created a frequency logo that included 3 nucleotides flanking both sides of the $(ARN)_2$ site but this did not show any consensus in those positions (Figure 14b).

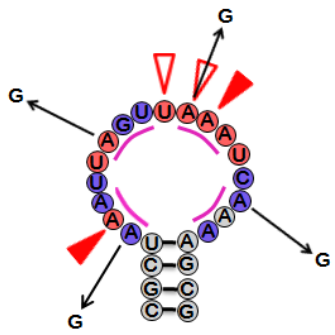


Figure 12. Model Hairpin for Discontinuous $(ARN)_x$. The model is designed based on the hairpin observed in the SHAPE structure of *dps*. ARN repeats are underlined in magenta. Arrows denote point mutations inserted to test the binding affinity of the sequence.

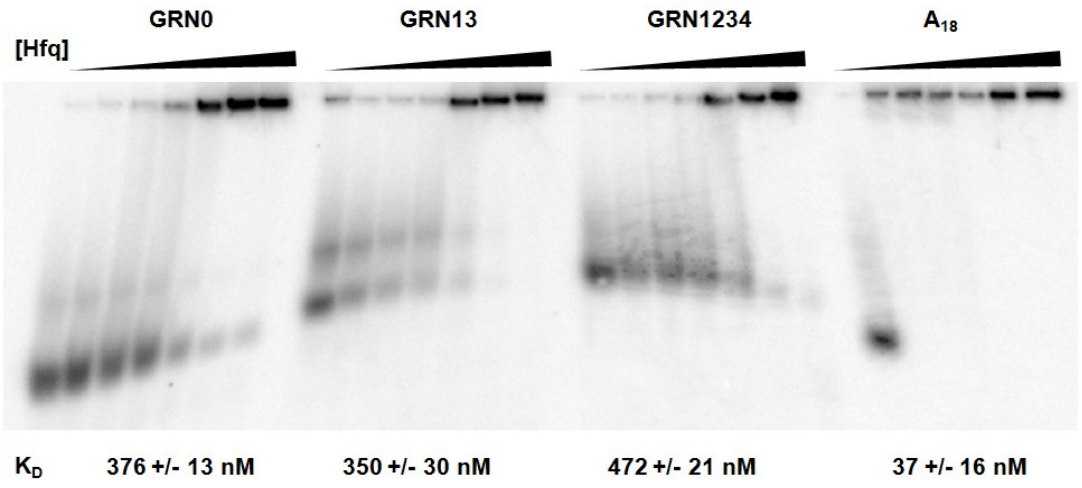


Figure 13. Ability of Hfq to Bind Discontinuous $(ARN)_x$ Sites. Native gel shifts of a discontinuous site hairpin construct and mutant constructs. GRN0 is the unaltered site, GRN13 contains G residues in place of A at the 1 and 3 repeat, and GRN1234 has a G in place of the A in all four repeats. Binding with A_{18} is shown as a positive control for distal site binding.

Computationally predicted secondary structures do not necessarily represent the actual structure of the RNA. To more accurately determine the secondary structure of

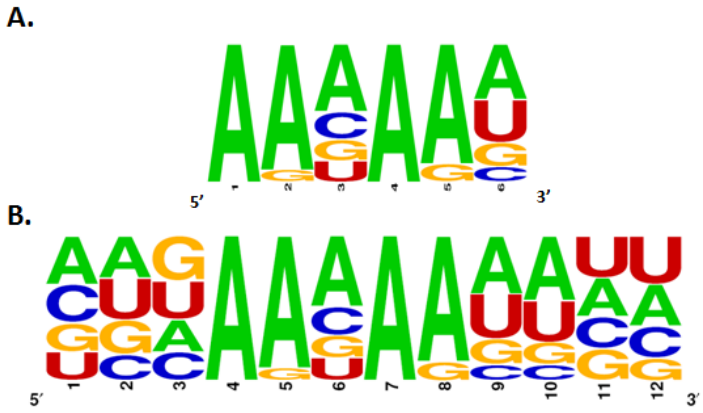


Figure 14 Frequency Logo of $(ARN)_2$ in TarBase mRNAs. A. Consensus sequence of $(ARN)_2$ in TarBase mRNAs was created using the $(ARN)_2$ site from each mRNA that fit the most criteria of the hypothesized Hfq binding site. **B.** Same $(ARN)_2$ sites but with 3 flanking nucleotides on either side of the site.

the 5' UTRs of mRNAs known to be regulated by sRNAs, we chose a set to perform Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) [101]. This technique, developed by the Weeks lab, allows for the experimental determination of RNA secondary structures with single nucleotide resolution by chemically modifying the RNA with N-methylisatoic anhydride (NMIA) (Figure 15) [101]. It is important to use experimental methods to determine these structures rather than relying solely on a computationally based fold like mfold because these folds have only 40-70% accuracy where the accuracy decreases as the size of the RNA increases [157]. Using computational methods in combination with SHAPE data leads to an RNA structure accuracy of 96-100%[157]. NMIA will react with the 2' OH of the ribose of nucleotides that are not constrained by interactions with other nucleotides [101]. The modified nucleotides are then observed as terminations in a reverse transcription (RT) reaction that uses a fluorescently labeled reverse transcription primer for quantification [104].

Three other reactions are run in tandem, each with a different colored RT primer [104]. One reaction with an unmodified RNA provides a signal that can be

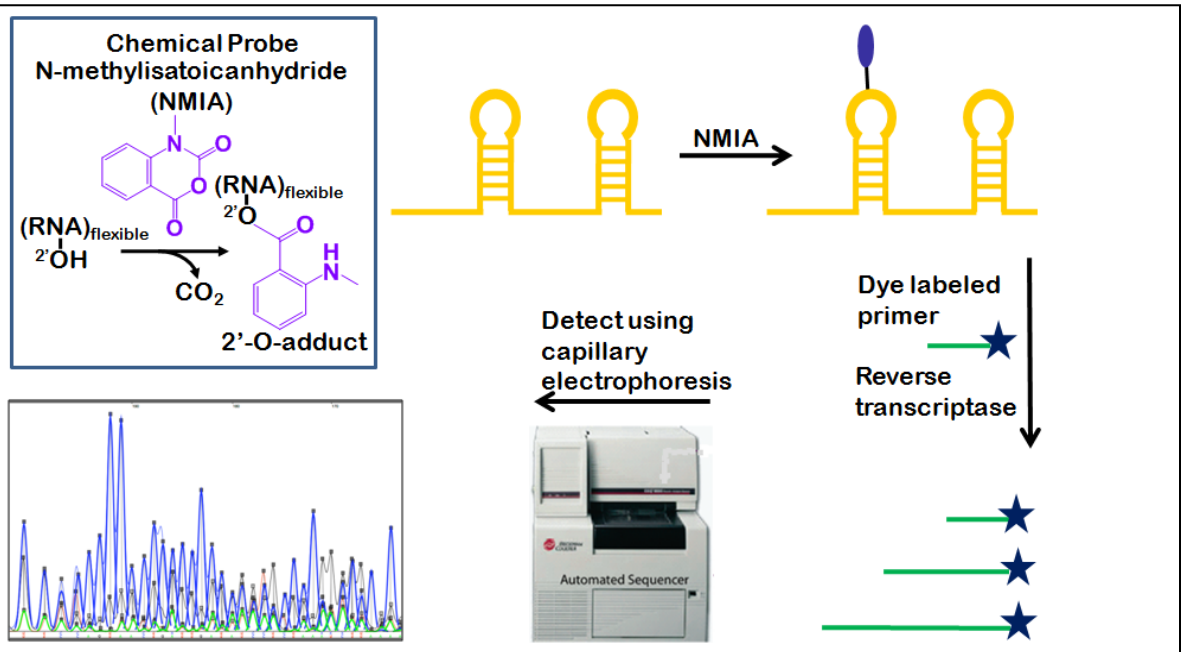


Figure 15. Overview of a SHAPE Experiment. Nucleotides that are flexible react with NMIA to make a nucleotide with a bulky adduct. Nucleotides in single stranded regions are more flexible and therefore react with NMIA. Adducts are detected as stops in an RT reaction. An NMIA+, NMIA-, and 2 sequencing reactions labeled by different colored fluorescent dyes are combined into one and fragments are resolved by capillary electrophoresis. When the data is analyzed there is an output of SHAPE reactivities for each nucleotide[103].

subtracted from the modified signal to account for any natural RT terminations [104]. The other two reactions are sequencing reactions that use ddATP or ddGTP to provide the sequence to which the NMIA + and NMIA- reactions can be aligned [104]. The 4 reactions are combined and the DNA fragments are separated by capillary electrophoresis using a DNA sequencer [104]. Values corresponding to the reactivity of each nucleotide with NMIA are obtained from the data using the ShapeFinder software [104]. This software adjusts the baseline, corrects signal decay, aligns the reaction peaks with sequencing peaks, and integrates the peaks as well as other necessary

processing [104]. The reactivity data are then used as constraints to provide an accurate fold of the molecule using the RNAstructure program [105].

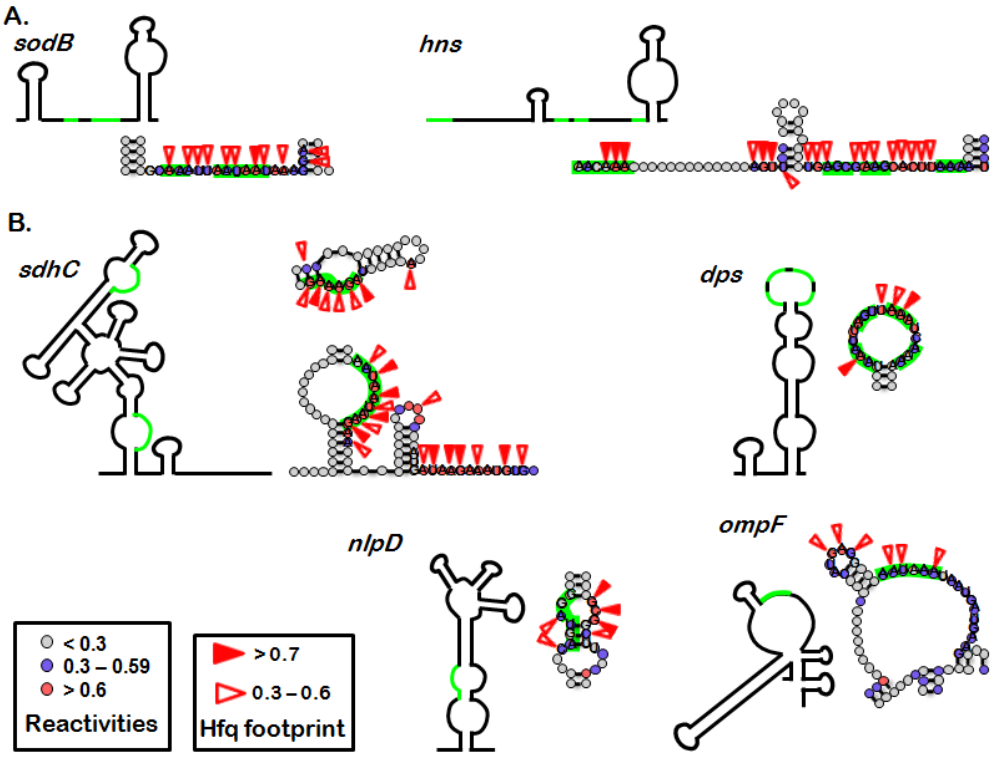


Figure 16. Structures of mRNAs with Hfq Foot Prints as Determined by SHAPE. A schematic of the overall structure as well as a detailed view of the $(ARN)_x$ site. $(ARN)_x$ sites are highlighted in green. Reactivities are shown by the colored circles and Hfq foot-prints are shown by the red triangles. **A.** $(ARN)_x$ sites flanked by hairpin loops. **B.** $(ARN)_x$ sites located within structural features.

The obtained SHAPE structures fell into two main classes. One class is made of mRNAs that have $(ARN)_x$ sites in unstructured single stranded regions that are flanked by structure (Figure 16a) and the other where the $(ARN)_x$ sites are located within regions of complex structure (Figure 16b). We also found that the structures determined using SHAPE were largely in agreement with the computationally predicted structures, suggesting that the use of predicted structures is sufficient and that it is not necessary to perform SHAPE on every mRNA of interest. This feature is important for increasing the throughput nature of our bioinformatic identification approach. In addition to

secondary structure prediction, SHAPE is a useful tool for footprinting. We compared the reactivities of the nucleotides in the presence and absence of Hfq to determine if

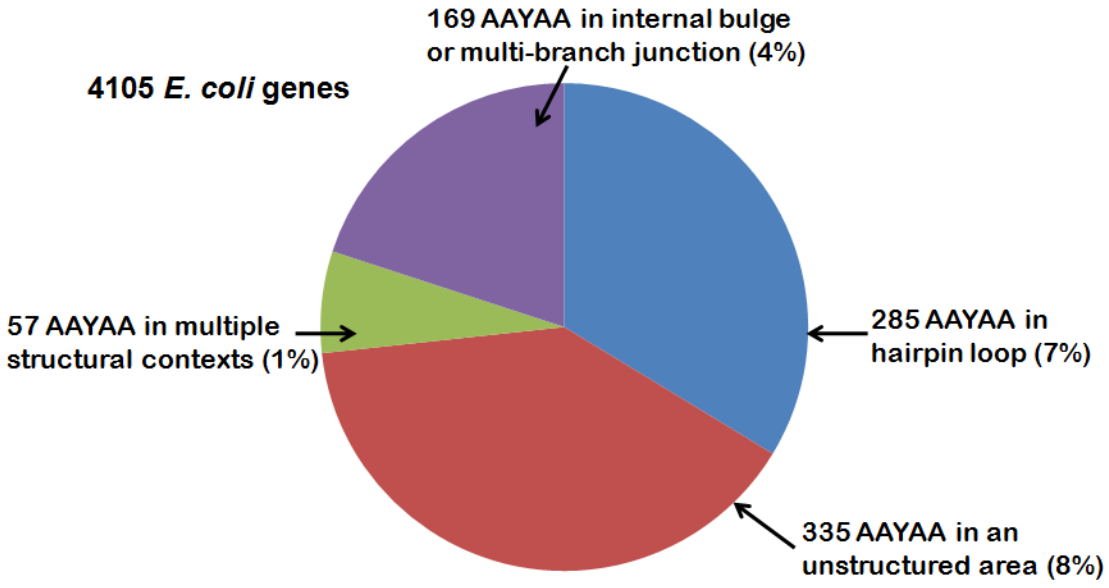


Figure 17. Genome Wide Characterization of AAYAA Sites in the 5' UTRs of mRNAs in *E. coli*. The 5'UTRs of all *E. coli* mRNA were searched for the sequence AAYAA. Computationally predicted structures were analyzed to determine the specific structural contexts of the sites.

Hfq binds $(ARN)_x$ sites *in vitro*. We found that the $(ARN)_x$ sites of the mRNAs that we investigated did in fact show significant protection from NMIA in the presence of Hfq (Figure 16).

The use of bioinformatic and experimental tools to investigate the characteristics of Hfq binding to $(ARN)_x$ motifs were then used to predict novel Hfq binding mRNAs. We carried out a genome wide computational search of *E. coli* to determine which and how many mRNAs have such a motif. We used an approach identical to the one outlined in Figure 10 with the exception that the sequence searched for was AAYAA. Historically, this was the first definition of the sequence within an mRNA to bind Hfq and is more stringent than $(ARN)_x$. Searching for this sequence may reduce the number of false

positives obtained in the search. The process of searching for this sequence in all 4,105 protein coding genes of *E. coli* was automated using an algorithm designed by Rebecca Swett (Swett and Feig, unpublished data). We found that 49% of the 4105 genes in *E. coli* contain an AAYAA sequence in their 5' UTR and 21% have a single stranded AAYAA within or flanked by regions of structure (Figure 17). Based on this search we suggest that 21% of *E. coli* mRNAs bind to Hfq and therefore may be targets of sRNA regulation. In the next chapter, we will show examples of predicted mRNAs that bind to Hfq *in vitro* and that 63% of a set chosen for validation demonstrated regulation *in vivo*.

2.4 DISCUSSION

Knowledge regarding the interactions between Hfq and targets mRNAs has not been developed to the same extent as sRNAs. The recent identification of an important Hfq binding motif in mRNAs has led us and others to shift our focus to this less understood area and investigate the importance of $(ARN)_x$ motifs in the sRNA regulatory network. We observed many important characteristics by investigating the $(ARN)_x$ motifs in 5' UTRs of mRNAs known to be regulated by sRNAs and Hfq. These messages are often highly structured and contain multiple $(ARN)_x$ sequences. The relatively high probability for the sequence $(ARN)_x$ to appear in the *E. coli* genome suggests that specific structural contexts may increase the specificity of the motif. The role of multiple $(ARN)_x$ sites is currently unknown but several possibilities exist. Hfq may be bound to 5' UTRs with multiple $(ARN)_x$ motifs, more often resulting in a greater chance of ternary complex formation, and therefore, regulation of that message. Targets that are regulated by more than one sRNA may have a specific $(ARN)_x$ to be used for regulation by each sRNA. Alternatively, only one of the multiple sites is functional as observed in

the only well studied example of this characteristic, *glmS* (Chapter 4) [84]. The motifs are single stranded and tend to lie within or surrounded by highly structured portions of the 5'UTR. Hfq binds single stranded RNAs; therefore, in order for the motif to be accessible for Hfq binding it must be single stranded. Two general types of structural contexts have been observed in extensively characterized examples thus far. In the cases of *fhlA* and *rpoS,* the motifs exist in internal bulges and, in the case of *glmS* (see Chapter 4), the $(ARN)_x$ is located in an unstructured single stranded stretch flanked by hairpin loops [36, 37, 84]. These cases correlate nicely with what we observed from the SHAPE structures of known target mRNAs. More $(ARN)_x$ motifs will have to be characterized in depth to determine a specific requirement for structural context. The use of SHAPE confirmed these conclusions and we predict that they are key features of $(ARN)_x$ motifs. A frequency logo created from $(ARN)_x$ motifs in known targets revealed that the R site has a preference for A but the N site shows no significant nucleotide bias. This preference justifies searching for an A at this site to reduce false positives, though we recognize that it may lead to false negatives. As more $(ARN)_x$ sites are validated the characteristics of this motif will become more clear but with the observations that we have made, this motif can be used as a search tool for novel target mRNAs.

We have taken advantage of the discovery of an Hfq binding motif in target mRNAs to develop a new approach for bioinformatic driven identification of regulated messages. Target identification has lagged behind sRNA discovery but is critical to understanding sRNA function and the dynamics of the regulatory network. All of the predictive models to date have focused on the sRNA-mRNA interaction to predict targets but the imperfect complementarity shared between the two has complicated the

effort and limits the applicability to organisms where the sRNAs have already been discovered. Effective computational target prediction is time saving and less labor intensive than traditional approaches using microarrays, HTS, or case by case analysis. It can facilitate the study of sRNA regulation in model systems rather than in dangerous or difficult to grow organisms. This searching approach can easily be modified to incorporate species specific features of Hfq binding. Predictions are not without error so the goal of developing such an approach is not perfection but rather to be able to identify high quality potential targets to be validated experimentally. We used criteria based on the characterized Hfq binding motif to search the 5' UTRs of mRNAs in the entire *E. coli* genome and found that 21% of mRNAs have a single stranded AAYAA located in complex structural regions. This result agrees with a study performed in *Salmonella* where 20% of mRNAs were shown to be bound to Hfq *in vivo* [47]. We predict that the mRNAs identified in our search bind to Hfq and may be regulated by sRNAs. This prediction expands the set of regulated known mRNAs from about 50 to upwards of 800. This degree of regulation helps explain the pleiotropic effects observed in the absence of Hfq in *E. coli* [26]. We should note however that the presence of a suitable $(ARN)_x$ motif may mean that a message binds Hfq but does not mean that it is absolutely regulated by sRNAs. In fact in the next chapter, we demonstrate that only about 63% of positive bioinformatics hits that were tested demonstrated regulation by sRNAs. Never-the-less, this data set represents a large increase in the number of mRNAs potentially undergoing regulation by *trans*-sRNAs.

2.5 MATERIALS AND METHODS

2.5.1 Bioinformatic Search of known sRNA targets

Sequences of the mRNAs known to be regulated by trans-sRNAs in an Hfq dependent manner were obtained from the TarBase database[155]. The sequences were obtained from the ecogene.org database for the region -200 to +60 or from the transcription start site to +60[158]. Annotated transcription start sites were obtained from the biocyc.org database[159]. The sequence were input to mfold and the folds were then analyzed for the presence of $(ARN)_x$ sequences and its position and structural context noted.

2.5.2 Bioinformatic Search of the *E. coli* genome

A list of all *E. coli* gene start positions and sense were obtained from the EcoGene database and formatted as a .csv file [158]. The genes were sorted by sense, and the start positions for both forward and reverse sense genes were output to separate files. A search was performed across the *E. coli* K-12 genome, wherein the region from -200 to +60 was searched for the sequence AATAA or AACAA, setting zero as each gene start position iteratively. The 260 nucleotide range and start position were output into a .csv file by line for all lines containing either the AATAA or AACAA sequence. This process was repeated for all negative sense genes using the *E. coli* K-12 genome complement strand sequence. Start position was matched back to gene name for functional analysis and the extracted 260 nucleotide region was submitted to mfold for structural analysis. Annotated transcription start sites for the biocyc database were used to discard any mRNA that contained and AAYAA in the region -200 to +60 but within the start site [159].

2.5.3 Construction of SHAPE Plasmid and Preparation of SHAPE RNAs

pMM110003 was constructed to serve as a parent plasmid for all further SHAPE experiments and contains a sequence cassette inserted into pUC19 with restriction sites inside the cassette for inserting any RNA of interest. The sequence was of the cassette was obtained from reference [101] and the cassette was created by primer extension of two complementary oligonucleotides from IDT.com. The sequence of the insert is

5'GGACACGAATTCCTATAATACGACTCACTATAGGCGACGGCCTTCGGGCCAAGG
TACCTCAGCGCTTCCTTAAGTCGATCCGGTTCGCCGGATCCCAAATCGGGCTTCG
GTCCGGTTCACGACCTGCAGGTCTACAAGCTTCCGAGC 3'

The restriction site to insert the cassette are *EcoRI* and *HindIII*, the restriction sites to clone an RNA of interest into the SHAPE vector are *KpnI* and *AflII*. For synthesis of RNA, the plasmids were linearized with *PstI* and run off transcription was performed. RNAs were purified by denaturing PAGE electrophoresis.

2.5.4 Expression and Purification of Hfq

Hfq was expressed and purified as previously described [30].

2.5.5 Chemical SHAPE analysis

SHAPE, as described previously [37, 101, 103], was performed to determine the secondary structure of mRNAs of interest with the following changes. To fold the RNA, 1 pmol was heated to 95 °C in a buffer containing 10 mM Tris-HCl pH 7.5 and 60 mM KCl. The (NMIA+/-) reactions were incubated with NMIA at 37 °C for 40-60 minutes, depending on the length of the RNA. Primer extension was carried out using 3 µl, 0.4 µM WellRED D4 primer for the NMIA+ reaction and 3 µl, 0.6 µM WellRED D3 primer for

the NMIA- reaction. For the 2 sequencing reactions, 3 µl, 2.2 µM WellRED D2 and 1 µl, 2.0 mM ddCTP were added to one sample and 3 µl, 2.2 µM WellRED D2 and 1 µl, 1.5 mM ddGTP to the other. The following parameters were used for separation on the Beckman CEQ 8000 DNA sequencer: capillary temp: 60 °C; denature temp: 90 °C; time 150 seconds, injection voltage; 5 kV, time 20 seconds; separation voltage 3 kV and separation time 100 minutes.

2.5.6 Chemical SHAPE footprinting

Footprinting using SHAPE was performed as described previously [37] with the following changes. The RNA was folded in the buffer described above. RNA was incubated in the presence and absence of 0.5 µM Hfq with NMIA at 37 °C for 40-60 minutes, depending on length. The NMIA reaction was quenched by adding 1 volume of 250 mM DTT.

2.5.7 Frequency Logo Creation

The frequency logo was created by submitting $(ARN)_2$ sequences observed in known mRNA targets to the website **http://weblogo.berkeley.edu/logo.cgi** and selecting the frequency plot option.

2.5.8 Milligan Transcription of Model Hairpins

The hairpin was created based on the discontinuous $(ARN)_x$ observed in the structure for *dps*. A common top strand with the T7 promoter was created to anneal to unique bottom strands containing the desired hairpin sequence.

Top strand: TAATACGACTCACTATA

GRN0: GCGCTTTTGATTTAACTAATTTAGCGCTATAGTGAGTCGTATTA

GRN13: GCGCTTTTGATTCAACTAATTCAGCGCTATAGTGAGTCGTATTA

GRN 1234: GCGCTTTCGATTCAACCAATTCAGCGCTATAGTGAGTCGTATTA

Transcription conditions were as follows; 100 nM top strand, 100 nM bottom strand, 40 mM Tris-HCl pH 8, 5 mM DTT, 1 mM spermidine, 0.01% triton X-100, 4 mM each NTP, 20 mM $MgCl_2$, 20 U/µl T7 RNA Polymerase. Reaction was incubated for 5 hours at 37ºC followed by DNase1 treatment. Reactions were purified by denaturing PAGE.

2.5.9 Gel Shift Analysis of Model Hairpins

RNAs were ^{32}P labeled by first dephosphorylating with calf intestinal alkaline phosphatase and then phosphorylating with T4 polynucleotide kinase in the presence of ATP gamma ^{32}P. RNA was gel purified. In preparation for binding the RNAs (amount determined to provide 15,000 CPM per lane) were heated to 95 ºC for 3 minutes in 50 mM Tris-HCl pH 8, and 100 mM KCl followed by cooling at room temperature for 15 minutes. Then 10 mM $MgCl_2$ was added followed by Hfq and the mixture was incubated at room temperature for 30 minutes. Hfq was added in varying amounts to achieve concentrations from 0 to 2.1 µM. Data were using a cooperative binding model using the equation: $Q_{bound} = [Hfq]^n/(K_d)^n + [Hfq]^n$

CHAPTER THREE: VALIDATION OF PREDICTED mRNA TARGETS

3.1 INTRODUCTION

We developed a computational approach to predict novel Hfq-binding mRNAs. The ability of an mRNA to bind Hfq suggests that the message may be regulated by *trans*-sRNAs although other reasons for Hfq binding cannot be ruled out. We chose a computational approach due to new information available about Hfq-mRNA binding, its high-throughput nature, as well as the ability to adapt the technique to other bacteria. Bioinformatics is an extremely useful tool that can be used to guide laboratory experiments but often has a degree of error. Therefore, it is necessary to validate computational predictions and to determine their biological significance. For the method to be useful, there must be a relatively high throughput way to validate predictions and it should have a significant positive discovery rate.

Sources of error in our approach include the use of computational folding, using AAYAA instead of $(ARN)_x$, and an incomplete knowledge of Hfq binding site requirements. The use of computational folds to determine the structural context of $(ARN)_x$ sites is important to our approach because it significantly improves the throughput of the technique as compared to lab based structure determination. It does introduce a degree of error into the technique; for example, a computationally based fold like mfold has an accuracy of 40-70%, reaching the greatest amount of error as the size of the RNA increases [157]. To determine an accurate secondary structure, enzymatic and chemical probing experiments can be performed (discussed in detail in Chapter 1) but are time consuming and must be done on a case-by-case basis. We used one such technique, SHAPE, to validate the secondary structure predictions for

known target mRNAs made by mfold in Chapter 2 and found them to closely resemble the predicted structure. These observations indicate that computational folds are sufficient for most mRNAs to correctly predict whether an $(ARN)_x$ motif is single stranded. We continued to use SHAPE to validate a subset of predicted targets which will be described in the following section. We chose to use the sequence AAYAA in our genome wide search (rather than $(ARN)_x$) which may have resulted in some true targets being missed. The more stringent criterion, AAYAA, describes the sequence of the first Hfq binding site identified; it wasn't until the crystal structure of Hfq bound to polyA RNA was determined that the specificity was widened [31, 36]. The more ambiguous nature of $(ARN)_x$ led us to use AAYAA in order to reduce a potentially large number of false positives with the sacrifice that some false negatives might arise. Limited examples of validated $(ARN)_x$ sites may have affected the ability of our approach to accurately predict targets. The rules that we determined for an Hfq binding $(ARN)_x$ motif were based upon the extensive characterization of two $(ARN)_x$ motifs and the nature of $(ARN)_x$ sites in other known mRNA targets. We made the assumption that $(ARN)_x$ sequences in the 5' UTRs of known target mRNAs that resemble the motifs in *glmS, rpoS*, and *fhlA* are also necessary for Hfq binding and regulation by sRNAs. While this scenario seems likely, more motifs will have to be validated to know for sure. There is also the question of the function of multiple $(ARN)_x$ sequences and proximity to the start codon and sRNA interaction site that still remain unanswered. As more information about these motifs becomes available our approach can be modified to improve the accuracy of the predictions.

All bioinformatics approaches require validation for a variety of reasons, including the reasons discussed above. There are a variety of ways that predictions regarding RNA-RNA interactions, RNA-Hfq interactions, and whether or not a target is being regulated, can be validated. The technicalities of many of these are presented in Chapter 1. RNA-RNA and RNA-Hfq interactions can be investigated by a variety of *in vitro* methods including: EMSA (Section 1.4.1), ITC (Section 1.4.4), SPR (Section 1.4.3), and footprinting (Section 1.4.5). All of these techniques are time consuming and can only be performed on an individual basis. It can also be difficult to execute *in vitro* experiments in a way that mimics cellular conditions to be able to make biologically relevant conclusions. A particular problem in this regard for RNA-Hfq systems is the actual concentrations of the molecules and the competition between RNAs for Hfq in a cellular environment. Being aware of these pitfalls when performing the experiments and analyzing and interpreting data can allow one to glean relevant information. The most useful *in vivo* techniques to validate interactions and regulation are gene fusion constructs (Section 1.5.3). Plasmid born systems can be performed in a relatively high throughput manner and can uncouple the expression of the components from the genome. These assays alleviate the unnatural conditions of *in vitro* experiments, for the most part, but they can fall victim to the unintended consequences of sRNA over expression. A fusion assay can be used to obtain a high standard of proof for both a direct RNA-RNA interaction that is biologically relevant and specific regulation by an sRNA. Observing changes in fluorescence when the fusion is expressed with a cognate and non-cognate sRNA can verify regulation. Proof for a direct RNA-RNA interaction can be cemented by introducing compensatory mutation into the RNAs to ablate and

restore regulation of the mRNA-fusion. In the following section, we present data using a combination of these techniques to validate Hfq binding and regulation of target mRNAs predicted using our bioinformatic approach.

3.2 RESULTS

To investigate the characteristics of the $(ARN)_x$ site and the ability of predicted targets to bind Hfq we selected the mRNAs *nhaA* and *mak,* at random to analyze by SHAPE and EMSA. The rationale for using SHAPE is similar to its use with known target mRNAs and it helps to deal with the potential of error due to reliance on

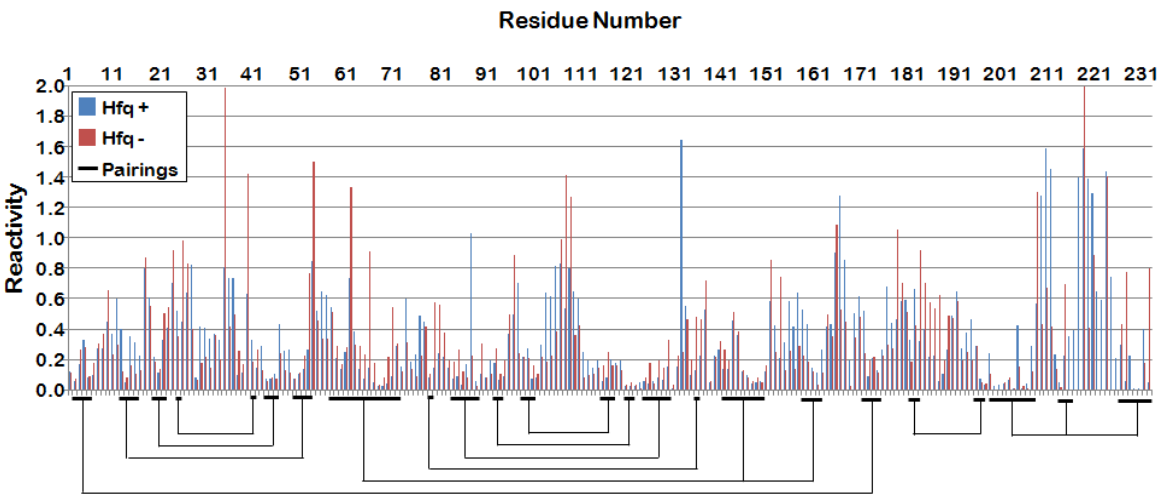


Figure 18. SHAPE Reactivities for *nhaA*. Raw intensities obtained after analysis using SHAPEfinder were normalized as described in Material and Methods to obtain reactivities for each nucleotide of the *nhaA* RNA. Data obtained from experiment in the presence and absence of Hfq are shown. Pairings that were present in the structure created by combining reactivities with the computational parameters of RNAstructure are shown.

computational folds as discussed above. SHAPE uses data obtained in the lab, in combination with computational parameters to increase the accuracy of the structure to 96-100% [157]. We used this technique to verify that the computational folds were allowing us to determine the structural context of $(ARN)_x$ sites accurately for most

mRNAs so that we could rely on the predicted folds only and increase the through put of the approach. In addition, we sought to characterize some of the predicted targets in more detail. SHAPE was performed on the targets as described in Chapter 2. Figure 18 shows a histogram of the reactivities that were determined for each nucleotide both in the presence and absence of Hfq. Regions that were predicted to base pair by RNAstructure are also indicated and correlate to regions of low reactivity. Nucleotides that were protected from reacting with NMIA by Hfq are evident by significant decreases in reactivity when compared to the no Hfq data. Structures were recreated schematically from the predicted fold based on SHAPE reactivity data and computational folding parameters as determined by RNAstructure (Figure 19). Reactivity data and Hfq footprinting is superimposed on the structure according to the key in Figure 19. Both of the selected mRNAs, *mak* (data not shown) and *nhaA* contained an $(ARN)_x$ in a highly structured, single stranded region of the 5'UTR (Figure 19). The $(ARN)_x$ motifs were located in close proximity to the translation start site as is often seen between an sRNA and its target (Figure 8). We also used SHAPE to obtain Hfq footprints for the two examples, both of which demonstrated protection at the $(ARN)_x$ site in the presence of Hfq. The presence of an $(ARN)_x$ site in these two mRNAs correlated to Hfq binding.

In the case of *mak* and *nhaA,* the presence of an $(ARN)_x$ led to a positive prediction that they would bind Hfq. The ability to bind Hfq does not guarantee that an mRNA is regulated by sRNAs *in vivo*. To investigate the possibility of *mak* and *nhaA* regulation, we predicted a likely sRNA partner for regulation using the program IntaRNA [119]. This program determines the favorability of hybridization between two RNAs and importantly it takes into consideration the accessibility of the interaction regions as well

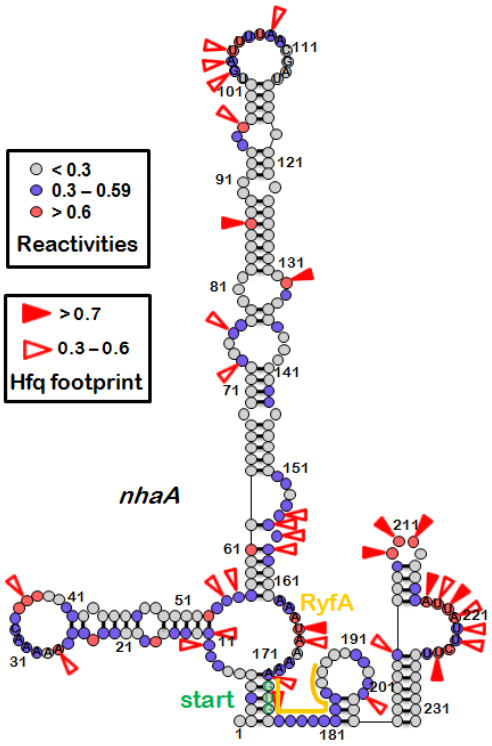


Figure 19. SHAPE Derived Secondary Structure and Footprinting of *nhaA*. SHAPE was performed on the 5' UTR of *nhaA*. The translation start site is in green and the predicted sRNA binding site is in yellow.

as a definable seed region. Accessibility is the probability of a region to be unpaired, therefore, single stranded areas are more accessible. This is an important feature of sRNA and mRNA interaction sites [152, 153]. Seed regions are nucleotide stretches that form perfectly complementary base pairing with their RNA partner and are important in the initiation and stability of an interaction [160]. Seed lengths observed in validated interactions range from nine to thirteen nucleotides [153]. IntaRNA was used to predict which of the known sRNAs was most likely to interact with *nhaA* and *mak*. One mRNA sequence was input along with a list of all of the trans-sRNAs known in *E. coli*. The program calculates the hybridization energy for each potential interaction and displays a list of up to ten pairings ranked from most favorable to least favorable. A

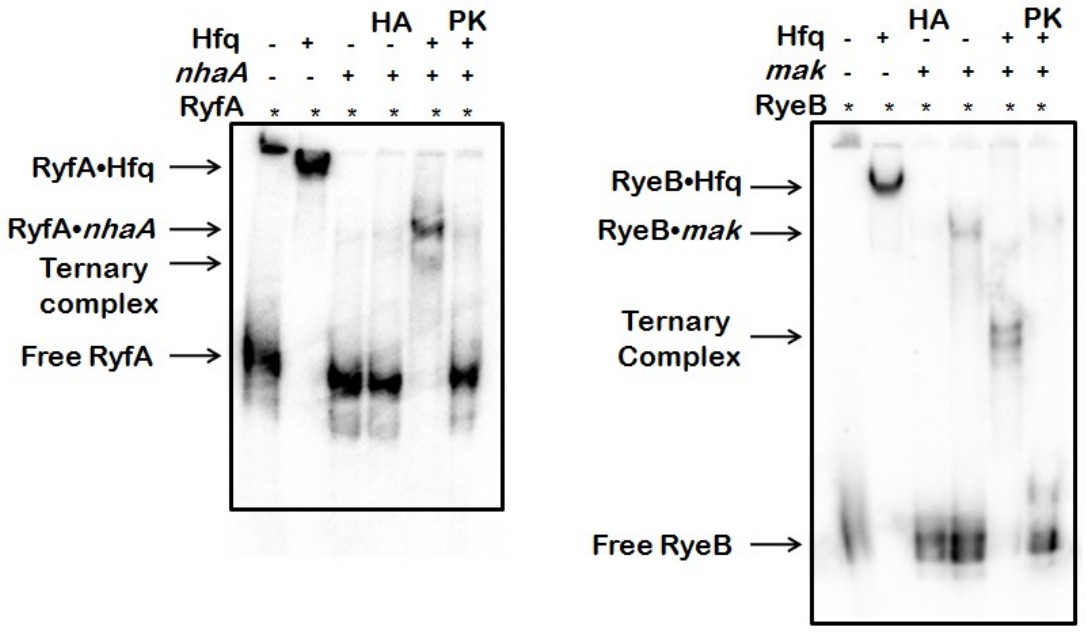


Figure 20. Secondary and Ternary Complex Formation. Native gel shift was performed. RNA-RNA duplex formation was tested by combining P^{32} labeled RyfA with 0.2 μM unlabeled *nhaA* or P^{32} labeled RyeB with 0.2 μM unlabeled *mak* . HA – heat anneal, PK - proteinaseK

graphical representation of the base paring is also provided for each. The interactions with the most favorable hybridization energies were with RyfA and RyeB, respectively. The RNAs were predicted to base pair at the translation start site of the mRNAs and within twenty nucleotides of the $(ARN)_x$ (Figure 19). Evidence suggests that a distance of more than twenty weakens the effect of Hfq binding on RNA annealing [143]. A gel shift was performed to determine if the two RNAs interact with their sRNA partners on their own or in the presence of Hfq *in vitro* (Figure 20). In the absence of Hfq, the two RNAs were able to form a duplex structure but only under heat annealing conditions. In the presence of Hfq, we observed both duplex and ternary complex formation. These complexes were Hfq dependent, as addition of proteinase K resulted in dissociation. In order to be sure of a direct interaction between the two RNAs, compensatory mutational analysis would have to be performed, but that level of investigation is more suitable for an in depth study of a specific pair of interest rather than in an initial validation stage.

Evidence of Hfq binding and RNA interaction from SHAPE and EMSA warrant continuing on to see if regulation occurs *in vivo*. Overall, we were able to predict two mRNA-sRNA pairs and demonstrate that they interact in the presence of Hfq *in vitro*.

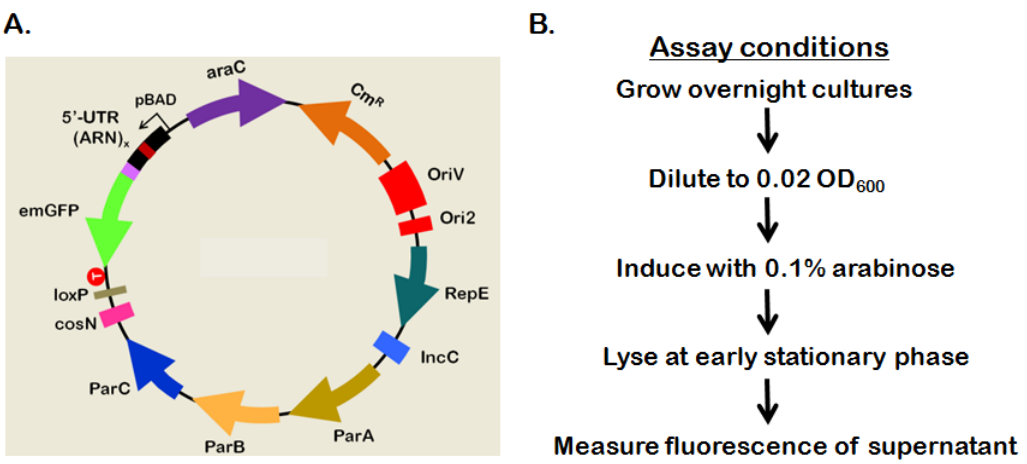


Figure 21. GFP Fusion Assay. A. Schematic of the GFP fusion plasmid[85]. The 5'UTR of the mRNA of interest is inserted in frame with emGFP and behind a pBad promoter. **B.** Basic flow of a GFP assay. More details available in Material and Methods.

The positive results observed *in vitro* led us to investigate sRNA dependent regulation *in vivo*. The ability of Hfq to bind RNAs *in vitro* does not guarantee that the same binding will occur *in vivo*, or if it does, that a biologically relevant regulation will occur. Therefore, it was critical to develop an assay to assess *in vivo* biological relevance. We chose to implement a GFP fusion assay based on its demonstrated success in studying mRNA-sRNA pairs and its straight forward/ low cost implementation (Figure 21) [84, 139]. The assay employs a two-plasmid system; a single copy plasmid containing the mRNA 5' UTR of interest inserted in frame with GFP and a high copy plasmid bearing the sRNA to be over-expressed. The sequence of the 5' UTR included the annotated transcription start site (or -200) through +60 to maintain consistency between the bioinformatics and the validation. The fusion transcript is controlled by a

P_{BAD} promoter rather than its natural promoter so that the changes in fluorescence observed are a result of translational regulation only. The sRNA is inserted into a high copy pBad24-derived plasmid that was provided by the Gottesman lab [140]. The high copy nature of this plasmid allows the study of sRNAs in typical culture conditions that are normally only expressed under specific conditions (that are often unknown). Using this plasmid, we made a library of all of the *trans*-sRNAs in *E. coli* that can be used to screen any mRNA-fusion of interest. The plasmids are both inducible with arabinose which insures that partners are co-transcribed as to not disrupt the sRNA network. The fluorescence out-put of the fusion plasmid can be monitored in 96-well plate format using a multi-well format plate reader. By measuring the fluorescence output of the fusion in the presence and absence of a potential sRNA partner we can determine if regulation occurs. While fluorescence can be measured from whole cells, we chose to use cell lysates to increase the sensitivity and precision of the assay. The relative ease of this assay allowed us to screen a large number of mRNAs at one time. We randomly selected a group of 18 mRNAs that included *mak* and *nhaA* from the *in vitro* experiments discussed above.

There is potential that some of the fusions may be regulated under normal growth conditions by endogenous sRNAs. We investigated this possibility by monitoring fluorescence levels in wild type and Hfq^- strains (Figure 22). Any regulation by natural sRNAs requires Hfq and would be lost in its absence resulting in a difference in fluorescence between fusions expressed in the two strains. To perform the assay, overnight cultures were diluted and grown for three hours followed by induction with arabinose. Cells were harvested after another three hours of growth and an

approximate OD_{600} of 1.5. Fluorescence was measured and normalized by OD_{600} at the time of harvest to account for any differences in growth. The relative expression levels were determined by dividing the fluorescence of an induced culture by that of an

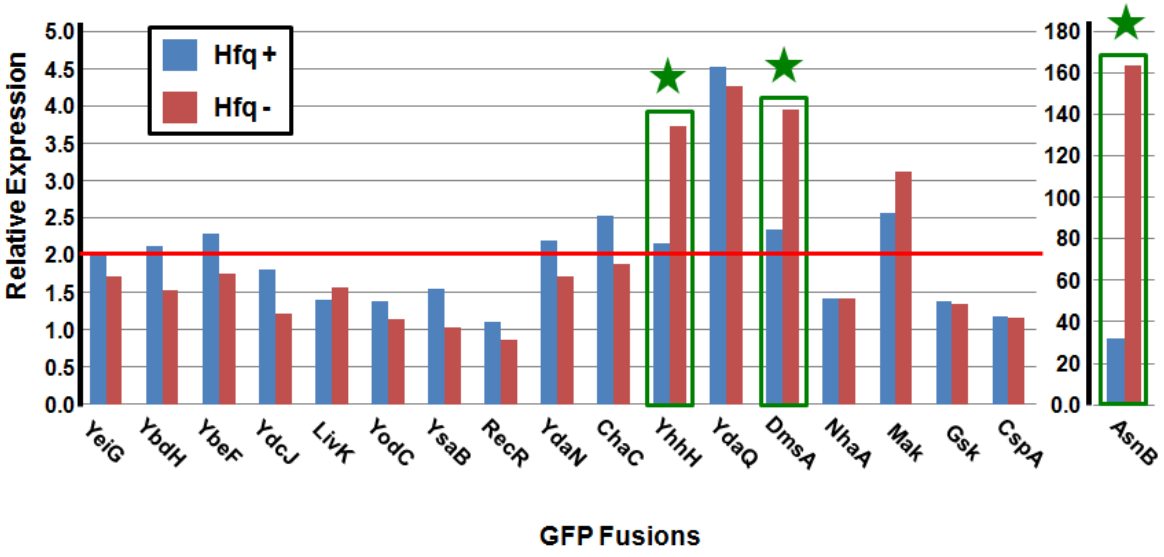


Figure 22. GFP Fusion Expression Levels. Fold change in fluorescence upon arabinose induction in wild type and Hfq- strains. Fold change was determined by dividing fluorescence values for induced cultures by uninduced. Constructs with green boxes and stars demonstrate regulation by endogenous sRNAs. The red line indicates the cutoff for which constructs have enough fluorescence induction to proceed with further experiments.

uninduced culture. We found that *yhhH*, *dmsA*, and *asnB* fusions had significant increases in fluorescence in the Hfq^{-} strains, suggesting that these constructs are down-regulated by endogenous sRNAs under the conditions of our assay. We also found that many of the constructs were exhibited fluorescence induction upon addition of arabinose; even after extensive assay optimization, only 8 constructs demonstrated an average of 2-fold or greater induction. Only the constructs with a fluorescence signal of 2-fold or greater can be tested for sRNA regulation using this assay.

We went on to test the 8 constructs that had fluorescence levels high enough to detect regulation with sRNAs. We determined the most likely sRNA partners for these mRNAs using IntaRNA and performed the GFP assay in the presence of the 2-4 sRNAs

ranked most favorably for an interaction. Initially, a large scale screen was performed to identify sRNA-mRNA partners that demonstrate a two-fold difference in fluorescence compared to the mRNA alone. The purpose of this initial screen was to quickly identify pairings that are potentially involved in regulation. Candidates that showed promise

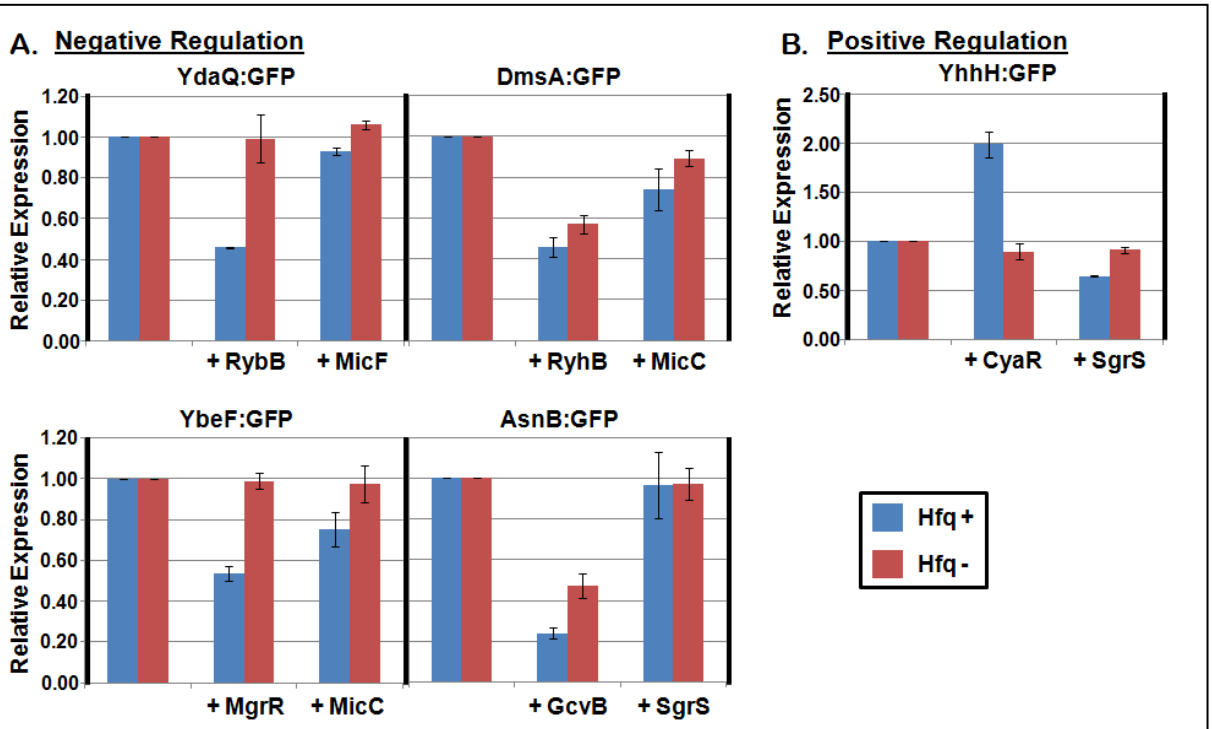


Figure 23. GFP Fusions Exhibiting Regulation by sRNAs. A. Constructs that were negatively regulated. **B.** Constructs that were positively regulated. Normalized fold change in fluorescence of GFP fusion was measured in the presence of the predicted sRNA partner (middle columns) and a non-cognate sRNA (third columns). Assay was performed in Hfq+ and Hfq- strains to demonstrate dependence on Hfq.

were further analyzed using a more rigorous assay that was performed in triplicate and with a control, non-cognate sRNA to demonstrate specific regulation by the predicted cognate sRNA.

Fluorescence levels are measured and normalized by their OD_{600} at harvest. In order to compare assays between different mRNAs (that were measured with different instrument gain), the relative expression values were normalized so that level of

induction of the mRNA alone is one. Error values were calculated as the standard deviation between the three trials. Fusions that demonstrated changes in fluorescence greater than that of the error when expressed with their cognate partner but not the control sRNA were considered positive for regulation. Overall, 5 out of 8 mRNA yielded positive results. Down regulation was exhibited by 4 (Figure 23a) and up regulation by 1 (Figure 23b). The three mRNAs (*yhhH, asnB, dmsA*) identified in the initial Hfq- strain were examined again in this assay and their specific regulatory sRNAs were discovered. Interestingly, results from the Hfq$^-$ screen suggested that *yhhH* would be down-regulated, but we only identified an up-regulating sRNA. It is likely that *yhhH* is up-regulated and down-regulated by different sRNAs, with the identification of the latter remaining elusive. While we were able to observe down-regulation and up-regulation of *yhhH* both, we were only able to identify the specific sRNA responsible for one. It may be that the other sRNA was missed due to an error in partner prediction by IntaRNA or that the sRNA has not yet been discovered and therefore wasn't included in our list of potential partners. We were able to demonstrate a 63% percent positive prediction rate for the constructs that were testable.

We were not able to identify sRNA regulation for the mRNA that we validated as Hfq binding *in vitro* (*mak*), which suggests that some mRNAs from the bioinformatics search may in fact bind Hfq but are not regulated by sRNAs. Also, if the goal is to identify mRNAs that are regulated by sRNAs, and not that just bind to Hfq, EMSA is not necessary because complex formation *in vitro* does not always correlate to regulation *in vivo*. There is also the possibility that we were unable to predict the correct partner

either due to an error in the predictive method or because of the unlikely event that the sRNA that regulates that message remains undiscovered.

We also performed assays for the positive targets in Hfq$^-$ strains to determine the dependence of regulation on Hfq. If regulation is dependent on Hfq, the observed change in fluorescence in the presence of the cognate partner sRNA will no longer occur in the Hfq knockout strain. In three cases the regulation of the construct was completely lost in the absence of Hfq, therefore the regulation of *ydaQ, yhhH*, and *ybeF* are Hfq-dependent events. In two cases, *asnB* and *dmsA*, only a modest or no loss at all was observed. This phenomenon is not inexplicable as sufficient levels of some sRNAs can bypass the need for Hfq [20].

3.3 DISCUSSION

The effective use of bioinformatics requires an efficient means for validation of predictions. We predicted that mRNAs containing $(ARN)_x$ sites would bind to Hfq and we used SHAPE and EMSA to validate this hypothesis. The two mRNAs investigated, *mak* and *nhaA*, contain $(ARN)_x$ sites similar to those present in known Hfq binding mRNAs and have demonstrated Hfq binding. We were also able to observe complex formation between the mRNAs, their predicted sRNA partners and Hfq *in vitro*. When these mRNAs were assayed for sRNA dependent regulation using an *in vivo* reporter assay, no regulation was observed. One mRNA-GFP fusion (*nhaA*) had a fluorescence level too low to be able to detect an event (discussed below) and the other (*mak*) either binds Hfq but is not regulated by sRNAs or the correct cognate sRNA was not identified. These results suggest that some of the mRNAs identified in our bioinformatics search may bind Hfq for other, currently unknown reasons and that Hfq may play roles other

than sRNA-dependent regulation in the cell. In fact, Hfq is able to stabilize the *ompA* mRNA by binding the 5' UTR and protecting the message from RNaseE cleavage [161]. We also may not have been able to identify the correct cognate sRNA using IntaRNA. A different predictive algorithm or a combination of several could be used to make a more accurate prediction in the future. It is possible that not all trans-sRNAs have been discovered, although extensive searches have been conducted for sRNAs in *E. coli*. Also, using *in vitro* binding as a step in the validation process may not be necessary if the goal of the query is to identify sRNA targets rather than all Hfq-binding mRNAs.

The regulation of our predicted mRNA targets was investigated using a GFP fusion reporter construct. This assay is easy and inexpensive. It is plasmid based which makes it less cloning intensive than creating chromosomal fusions and, because the GFP plasmid is single copy, it mimics a natural gene. An sRNA plasmid library must be created once and can then be conveniently used to investigate an endless number of mRNAs of interest. The only modestly costly component of the assay is the ability to take fluorescence measurements but instruments with this capability are commonly found in most departments. An obvious drawback is the fact that only 8/18 of the constructs made exhibited fluorescence levels significant enough to test for regulation. There are several possibilities to explain these low signals. The message may not be transcribed due to an alternative endogenous regulatory pathway. The mRNA may be misfolded or contain a decay signal and is rapidly degraded. The reporter protein may not be translated as a result of an unknown Hfq-independent regulatory pathway. The resultant protein may not be stable or may be misfolded, leading to degradation, sequestration or low fluorescence. Some mRNAs require specific processing events for

regulation to occur and plasmid construction may disrupt this route [162, 163]. In order to test the low fluorescence constructs, we could first perform qPCR or Northern blots using RNA extracted from cells expressing the fusion plasmid to determine if the transcript is being made in detectable amounts. If it is, then the same experiment could be performed in the presence of a partner sRNA and control sRNA to see if the partner is regulating the fusion by affecting the stability of the transcript. If no regulation is observed at that level a Western blot with an anti-GFP antibody could be performed to determine if the fusion is regulated at the translational level.

By implementing an initial screen of the mRNA constructs in wild type and Hfq$^-$ strains, we were immediately able to identify three Hfq-dependent mRNAs that are presumably regulated by *trans*-sRNAs. Not only was this screen a simple means to identify targets, but it also allowed identification of mRNAs that are regulated by endogenous sRNAs. It is important to identify regulation by endogenous sRNAs because targets regulated in this manner could appear as just having low fluorescence signal and be discarded. Of course not all targets can be identified this way due to expression specific conditions of many sRNAs nor does this approach identify the specific sRNA responsible for regulation.

We were able to identify regulation and the specific sRNAs responsible for 5, or 63%, of the mRNA constructs. This success rate speaks to the efficacy of using $(ARN)_x$ motifs to predict target mRNAs. The three negative results suggest that these mRNAs either bind Hfq but are not regulated or we were not able to identify the correct sRNA, as discussed above. When we monitored the ability of regulation to occur in the absence of Hfq we found that three regulatory events required Hfq and two did not.

Regulation by MgrR, RybB, and CyaR most likely require Hfq to facilitate base pairing or to stabilize the sRNA and protect it from degradation [139]. The events that did not require Hfq were most likely due to the sRNA over-expression conditions [20].

Our novel bioinformatic approach led to the discovery of 5 new and interesting mRNA-sRNA pairs. YdaQ is a putative Rac-prophage excisionase [164]. Temperate phages infect bacteria and integrate their genetic material in the hosts, becoming a prophage, where they can lie dormant or become lytic [165]. They control these stages using an integration/excision system encoded in its genes [166]. Throughout evolution, some of these phages lose their ability to form plaques, produce phage particles, or induce host lysis and therefore become trapped in the host genome as 'cryptic' prophages [167]. The host is then in control of the phage genes and, through mutagenesis and decay, it inactivates detrimental genes and maintains beneficial ones [167, 168]. Prophage genes are under strict regulation in bacteria and play important roles in antibiotic resistance, stress responses, and biofilm formation [169]. One of the first discovered and evolutionarily oldest cryptic prophages is the Rac-prophage [170]. The Rac-prophage mRNA, *ydaQ*, levels are increased during biofilm formation in *E. coli* as detected by microarray analysis [171]. In *E. coli*, increased excision of a different prophage, CP4-57, is beneficial for biofilm production [172]. This evidence, along with the sRNA-dependent regulation of *ydaQ* that we observed, suggests a role for *ydaQ* in biofilm production that is under the control of the sRNA RybB.

Another target that we found to be down-regulated was *dmsA*, which codes for the catalytic subunit of the protein dimethyl sulfoxide reductase (*dmsABC*)[173]. This protein is a member of the complex iron-sulfur molybdoenzyme family that allows *E. coli*

to be a facultative anaerobe by facilitating the use of a variety of respiratory substrates[173]. Specifically, dimethyl sulfoxide reductase acts to couple dimethyl sulfoxide reduction to menaquinol oxidation. This electron carrier function is made possible by the presence of an iron-sulfur complex located within the DmsA subunit [173]. The sRNA that we found regulates this message, RyhB, is a key regulator of iron homeostasis [174]. In situations of iron starvation, RyhB acts by preventing the synthesis of non-essential iron containing proteins [174]. Our results suggest that RyhB down-regulates *dmsA* in the presence of oxygen, when there is no need for DMSO reduction, thus sparing the use of iron for critical functions.

We observed the down regulation of *ybeF* by MgrR using the GFP assay. The gene *ybeF* encoded a putative DNA binding transcriptional regulator of the LysR family [175]. LysR-type transcriptional regulators are the largest group of transcriptional regulators with over 100 members identified in diverse bacterial species [176]. These are global regulators that can up or down-regulate single genes or operons and are involved a broad range of cellular physiology including metabolism, quorum sensing, and virulence [177-180]. The large, diverse nature of this transcriptional regulator family makes it difficult to speculate about the role of *ybeF* but we can conclude that it represents yet another transcriptional regulator under the control of an sRNA. MgrR has one experimentally verified target mRNA, *eptB*, that encodes phosphoethanolamine transferase, an enzyme that modifies lipopolysaccharides (LPS) on the bacterial cell surface [181]. The LPS is modified in a highly regulated fashion to enable cell survival and pathogenesis in the host [182]. MgrR is a component of this complex system, therefore the role of MgrR-ybeF regulation that we observed may be in LPS

modification. More investigations are required to determine the exact role of this novel regulon.

We also identified *asnB* and *yhhH* as sRNA targets. AsnB was chosen for further study and will be discussed in Chapter 4. The gene *yhhH* codes for a protein of unknown function and the *E. coli* genome contains a paralog to this gene named *ybbC*, also of unknown function[183]. *yhhH* has appeared in two systemic phenotype studies. In one study, Tenorio *et al.* observed the effect of over expression of a complete set of ORFs on biofilm formation and found that over expression of *yhhH* caused abnormal biofilm architecture [184]. Murata *et al.* identified *yhhH* as vital for survival at critically high temperature in a knock-out screen and chip assay [185]. The sRNA, CyaR, has three confirmed targets (*nadE, ompX, luxS*) that participate in seemingly unrelated tasks such as NAD synthesis, outer membrane stress, and quorum sensing[174]. All of these targets are down-regulated by CyaR, therefore the observed up regulation of *yhhH* is the first of its kind for CyaR. There may be a link between the role of *yhhH* in biofilm formation and CyaR in quorum sensing but that hypothesis requires more investigation to make any solid conclusions. Also, we hypothesize that *yhhH* is down-regulated by an as of yet unknown sRNA due to the increased fluorescence exhibited in the absence of Hfq.

In conclusion, we were able to identify 5 new mRNA-sRNA regulatory pairs using our novel bioinformatics approach for an overall positive identification rate of 63%. This technique is easy to use and adaptable to other bacteria of interest. As more information about the $(ARN)_x$ motif is learned, these details can be incorporated into the search to make it an even more valuable tool. The results of our validation suggest that

many more target mRNAs exist in *E. coli* than have been identified. If 63% of the mRNAs that we predicted in our genome wide search are targets, than a total of more than 500 target mRNAs in *E. coli* are likely. This number helps explain the pleiotropic effects observed in Hfq deletion strains as well as the fiercely competitive environment for Hfq binding. Understanding the number of targets and their identities contributes to our knowledge of the dynamics of how this network functions and the physiological processes that it coordinates. Defining the targets in *E. coli* could also lead to the identification of sRNAs and target mRNAs in other organisms through sequence, structural and/or functional homology.

3.4 MATERIALS AND METHODS

3.4.1 Bacterial Strains, Media and Growth Conditions

The *E. coli* strain TOP10 (Invitrogen) was used for all cloning and GFP assays. The Hfq knockout was also created in TOP10 cells by inserting a kanamycin gene in place of Hfq using the Quick and Easy Conditional Knockout Kit (Gene Bridges) as described elsewhere[88]. Growth conditions were in Luria-Bertani (LB) broth or plates at 37ºC. The antibiotics ampicillin (100 µg/ml) and chloramphenicol (34 µg/ml) were used as required.

3.4.2 Fusion Plasmid Construction

Fusion plasmids were created using the parent plasmid pBacEmGH which was provided by the Cunningham Lab[186]. Primers were used to amplify the mRNA of interest from the annotated transcription start site, as noted in the Biocyc database, to +60 nucleotides into the ORF. If the transcription site was not available -200 was used.

Primers included the restrictions sites for cloning, forward primers contained *NheI* and reverse primers contained *NotI* sites, respectively. Newly constructed plasmids were verified by sequencing.

3.4.3 sRNA Plasmid Construction

sRNA plasmids were created using the parent plasmid pNM12[140]. Primers were used to amplify the sRNAs from *E. coli* TOP10 cells and included the restriction sites MscI and EcoRI. Newly constructed plasmids were verified by sequencing.

3.4.4 Fluorescence Data Collection

Overnight cultures were started by inoculating 5 mL LB containing the appropriate antibiotic/s with a single colony. Cultures were grown overnight with shaking. The following morning, cultures were diluted to OD_{600} 0.2, or 0.5 for Hfq knockout strains, to create two cultures for each strain. Diluted cultures were grown for 3 hours and then one of each strain was induced with 0.005% arabinose. Cultures were grown another 3 hours to early stationary phase and 3 mL were pelleted. The remaining culture was used to obtain OD_{600} values for all samples. Pellets were suspended in 200 µL lysis buffer, 50 mM Tris–HCl pH 7.5, 25 mM NaCl, 2 mM EDTA. Cell suspensions were lysed by adding 15 µL lysozyme (20 mg/mL, Fisher), 30 µL protease inhibitor solution (one tablet of complete EDTA-free protease inhibitor (Roche) dissolved in 8 mL) and 30 µL 1% TritonX-100 for 30 minutes at 37 °C while shaking. Cell debris was pelleted and 200 µL of the supernatant was loaded into a 96-well flat bottom black plate (Corning®). Fluorescence was measured by a Tecan GENios Plus multi-label plate reader with an excitation wavelength of 485 nm and emission of 525 nm. The

instrument gain was optimized for each mRNA construct and the numbers of flashes was set at 50.

Data was analyzed by determining the fluorescence/OD values for each sample to account for variations in growth rates. The fold change in fluorescence was calculated by dividing the induced value by the uninduced value. In order to compare regulation among different mRNA-sRNA constructs, data was then normalized so that the fold change of mRNA fluorescence upon induction was 1.

3.4.5 Duplex and Ternary Complex Formation

Gel shifts were performed using 6% native polyacrylamide gels cast and run in 1XTBE. The sRNAs were treated with calf intestinal phosphatase followed by phenol:chloroform extraction and precipitation. The RNA was then labeled with ^{32}P using T4 polynucleotide kinase with γ-^{32}P labeled ATP, followed by gel purification. Immediately before performing each experiment, the RNAs were re-folded by heating to 95 °C in 50 mM Tris-HCl and 100 mM KCl, followed by cooling at room temperature for 15 minutes and the addition of 10 mM $MgCl_2$. RNA only lanes contained ~ 15,000 CPMs of ^{32}P. For duplex formation, 0.2 μM unlabeled mRNA was added and allowed to bind for 25 minutes at room temperature. To heat anneal the mRNA and sRNAs partners to promote duplex formation the two RNAs were mixed followed by heating to 95 °C for 3 minutes and cooling/binding at room temperature for 25 minutes. Ternary complex was formed by adding 0.5 μM Hfq hexamer to pre-incubated sRNA-mRNA mixtures and incubating for 25 minutes at room temperature. Ternary complexes were treated with 5 μL proteinase K (20 mg/mL) at 37 °C for 30 minutes.

CHAPTER FOUR: CHARACTERIZATION OF *asnB* REGULATION

4.1 INTRODUCTION

Chapters 2 and 3 described the methodology that we used to predict and validate 5 new mRNA targets. The target that demonstrated the most extreme regulation was *asnB*. The AsnB:GFP fusion was strongly down-regulated under normal growth conditions as evidenced by the large increase in fluorescence in an Hfq$^-$ strain (Figure 22). We then showed that the GcvB sRNA was able to down-regulate AsnB:GFP when over-expressed from a plasmid (Figure 24). We therefore chose to further investigate AsnB and the interactions between Hfq, *asnB,* and GcvB. AsnB is an asparagine synthetase that can catalyze the synthesis of asparagine with ammonia or glutamine as

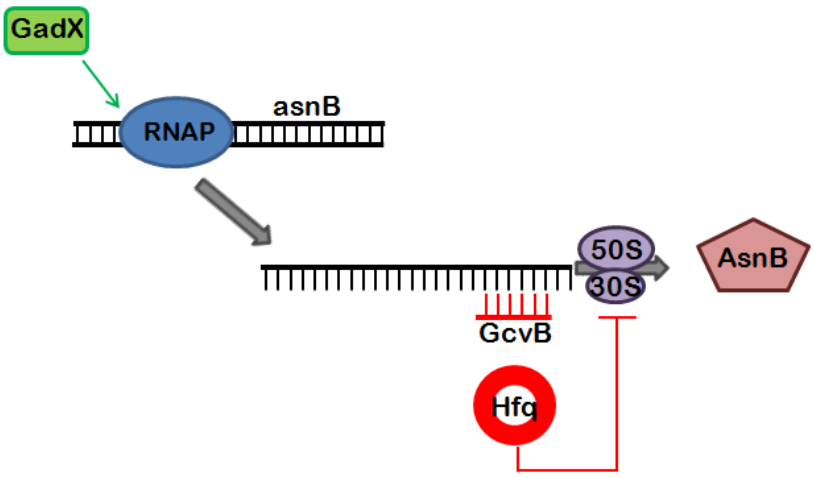


Figure 24. Modes of AsnB regulation. The transcriptional regulator GadX can up-regulate asnB transcription[191]. The sRNA, GcvB with the help of Hfq down-regulates translation of *asnB*.

a nitrogen source (glutamine is preferred) [187]. In *E. coli* an additional asparagine synthetase is encoded by an unlinked gene, *asnA* [188]. This synthetase prefers ammonia as a substrate. Both asnA and asnB must be knocked out to create an asparagine auxotroph [188]. AsnB catalyzes the transformation of aspartic acid to asparagine in three steps: aspartate is activated by the addition of AMP, glutamate and

ammonia are produced by glutamine hydrolysis, and nucleophillic attack by the ammonia leads to breakdown of the β-aspartyl-AMP intermediate to form asparagine (Figure 25a) [189]. Glutamine hydrolysis and β-aspartyl-AMP intermediate formation

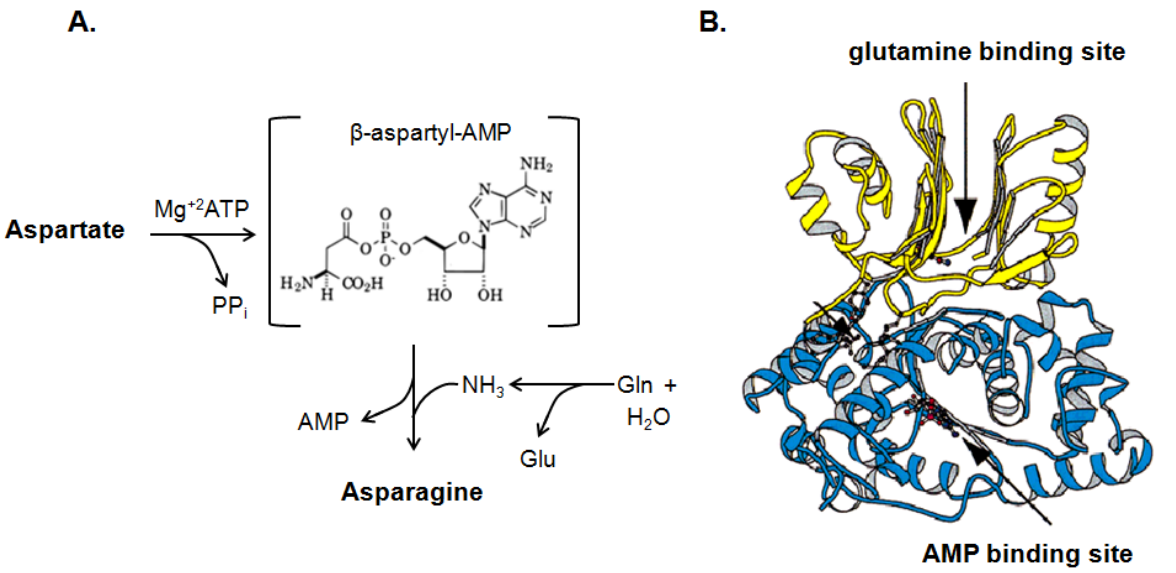


Figure 25. Asparagine Synthase, AsnB A. Reaction catalyzed by AsnB to synthesize asparagine. **B.** Crystal structure (1CT9) of AsnB showing the two distinct active used in catalysis[190]. Reprinted with permission from [190]. Copyright (1999) American Chemical Society.

occur in two distinct active sites of the enzyme as shown in the crystal structure (Figure 25b) [189]. The two active sites are separated by 19 Å and are connected by a molecular tunnel formed by hydrophobic and non-polar side chains [189].

The *asnB* transcript is only detectable when cells are grown in the absence of asparagine [187]. This fact correlates nicely with our observation that AsnB:GFP is strongly repressed under normal growth conditions. Levels of *asnB* are up-regulated by the transcriptional regulator GadX in response to acid stress (Figure 24) [190]. Under this stress condition, GadX is also responsible for increasing levels of glutamate decarboxylases [190]. Glutamate is a product of asparagine synthesis by AsnB and a required substrate of glutamate decarboxylases [190]. AsnB has also been shown to be

involved in resistance to oxidative stress and virulence in the plant pathogen *Xanthomonas oryzae* [191]. These functions link asnB to stress and virulence conditions which make it an interesting target to study.

We found that the GcvB sRNA down-regulates AsnB. Expression of GcvB is controlled by the transcription factor GcvA in response to glycine levels [192]. Under normal growth conditions in LB, GcvA is up-regulated and in turn increases levels of GcvB. GcvA is encoded divergently from GcvB, a relationship that is maintained throughout a diverse range of bacterial species [151]. GcvB is responsible for regulating 21 genes in *Salmonella* which makes it the largest regulon observed for any one sRNA [112]. GcvB represses its target mRNAs, all of which are involved in amino acid uptake and synthesis [112]. AsnB fits in well with this regulon. GcvB is one of the most well conserved sRNAs studied to date [151]. Two factors likely contribute to its conservation are 1) the large number of targets constrains mutations in the binding sites and 2) its central role in amino acid transport and metabolism which are important pathways in all bacteria. GcvB is unique from other sRNAs because it has three potential mRNA interaction sites. The R1 binding site is responsible for most of the regulatory events and the contribution of R2 and R3 are less well defined (Figure 26c) [112]. For some of the mRNAs, mutation of either R1 or R2 alone does not cause a loss of regulation but when both are disrupted an effect is observed, suggesting a degree of redundancy between the two sites [193, 194]. Only one instance of regulation requiring R3 has been observed but some of the previous studies focused strictly on R1 and R2 therefore potentially missing R3 binders [112, 195].

4.2 RESULTS

The regulation of *asnB* that we observed was predicted using a computational structure in our bioinformatics approach that was described in Chapter 2. To further

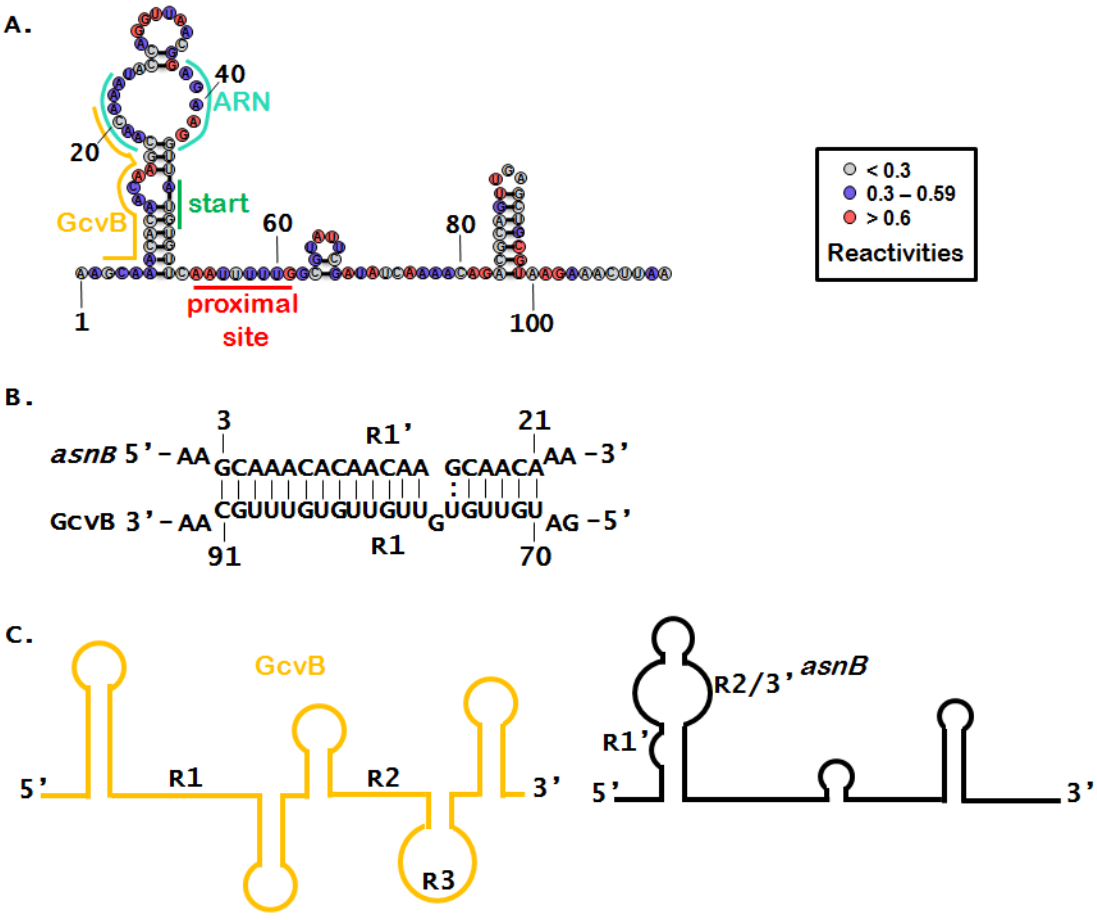


Figure 26. *asnB* Structure and Interaction with GcvB. A. SHAPE Derived Secondary Structure of *asnB*. SHAPE was performed on the 5'UTR of *asnB*. Reactivities are superimposed on the fold. The translation start site is in green and the predicted sRNA binding site is in yellow. $(ARN)_x$ motifs are marked in teal. A predicted proximal binding site is labeled in red. **B. Detailed view of the predicted hybridization between *asnB* and GcvB in *E. coli*.** Hybridization was predicted using IntaRNA. **C. Schematic of the GcvB and *asnB* Structures.** RNA interaction sites that are predicted in *E. coli* and homologs from other organisms are labeled[196].

investigate the structure of the 5'UTR of *asnB* we performed SHAPE (Figure 26). The short, 45 nucleotide 5' UTR contains several $(ARN)_x$ sequences, two of which are single stranded and located across from each other in an internal loop. Both are in close proximity to the start codon and the predicted GcvB interaction site, which lies at the 5'

end of the UTR. The interaction site contains the canonical CACAaCAY binding motif for the R1 binding site of GcvB [112]. Given its close proximity to the translational start codon, GcvB most likely disrupts ribosome binding to down-regulate translation of *asnB*. The $(ARN)_x$ sites adhere to the observation that the most effective Hfq binding sites are located 3' and within 20 nucleotides of the sRNA interaction site [143]. One $(ARN)_x$ partially overlaps with the GcvB binding site. This type of regulatory structure, where the Hfq binding motif overlaps with the sRNA recognition site, has been previously shown to have negative effects on the ability of an sRNA (SPOT42 in this case) to regulate its targets [144]. Thus, the first $(ARN)_x$ in *asnB* may not be functional. On the other hand the first site does not overlap the seed of the interaction (-3 to -15) so one could speculate that Hfq may bind that site to facilitate the initial annealing step and then dissociate to allow extended interactions to occur. Future experiments will reveal the role of the two $(ARN)_x$ motifs. There is also an AU rich single stranded stretch located 3' to the start codon that is most likely an Hfq proximal binding site [32, 33]. The *asnB* transcript may bind to both faces of Hfq a type of wrap around interaction that has been observed for other mRNAs [37, 84].

To investigate the binding properties of this newly discovered *asnB*/GcvB regulatory pair, we performed native EMSA to separately determine the K_D values for *asnB* and GcvB binding to Hfq (Figure 27). RNAs were radiolabeled and bound to increasing amounts of Hfq. Binding was allowed to reach equilibrium before performing electrophoresis. Results were quantified after phosphorimaging and fit by a nonlinear least-squares analysis to a cooperative binding model (Section 1.3.1, equation 4). Due to the trace amounts of RNA and excess Hfq, supershifts that represent one RNA

bound to multiple Hfq hexamers were observed. These complexes are a result of *in vitro* conditions and are most likely biologically irrelevant, therefore we treated all bound species as one when calculating affinities. Both RNAs bind Hfq with high affinity (*asnB* K_D = 1.0 ± 0.1 nM, GcvB K_D = 5 ± 2 nM) (Figure 27). We also observed the ability of the two RNAs to form a duplex structure in the absence of Hfq. Labeled GcvB was titrated with increasing amounts of unlabeled *asnB* and allowed to bind. The K_D for duplex formation was determined to be 42 ± 2 nM (Figure 28). The tight binding demonstrated by these two RNAs may contribute to the ability of GcvB to regulate *asnB in vivo* even in the absence of Hfq. We also performed gel shift analysis to assess the formation of stable ternary complexes. Pre-formed GcvB*•Hfq complex was titrated with increasing amounts of unlabeled *asnB* (Figure 29, * indicates the presence of a radiolabel). The pre-formed complex migrates similarly to the GcvB • Hfq multimeric species observed in Figure 27. Addition of increasing amounts of *asnB* led to the formation of a higher molecular weight complex whose intensity grew as more *asnB* was added. This high molecular weight complex most likely represents an *asnB*•GcvB•$(Hfq)_n$ ternary complex. The exact ratios of the species cannot be determined from this experiment alone but the three players are clearly interacting.

We also analyzed the ability of *asnB* to bind to two Hfq mutants. One had a mutation that disrupts binding to the proximal face of Hfq (K56A) and the other had a mutation that disrupts binding to the distal face (Y25A) [30]. The presence of an $(ARN)_x$ site as well as an AU rich stretch in the *asnB* RNA suggests that it binds to the distal site and the proximal site of the protein and therefore we predicted that we would not

detect a significant defect in binding either mutant. Gel shifts were performed as described above and we found the

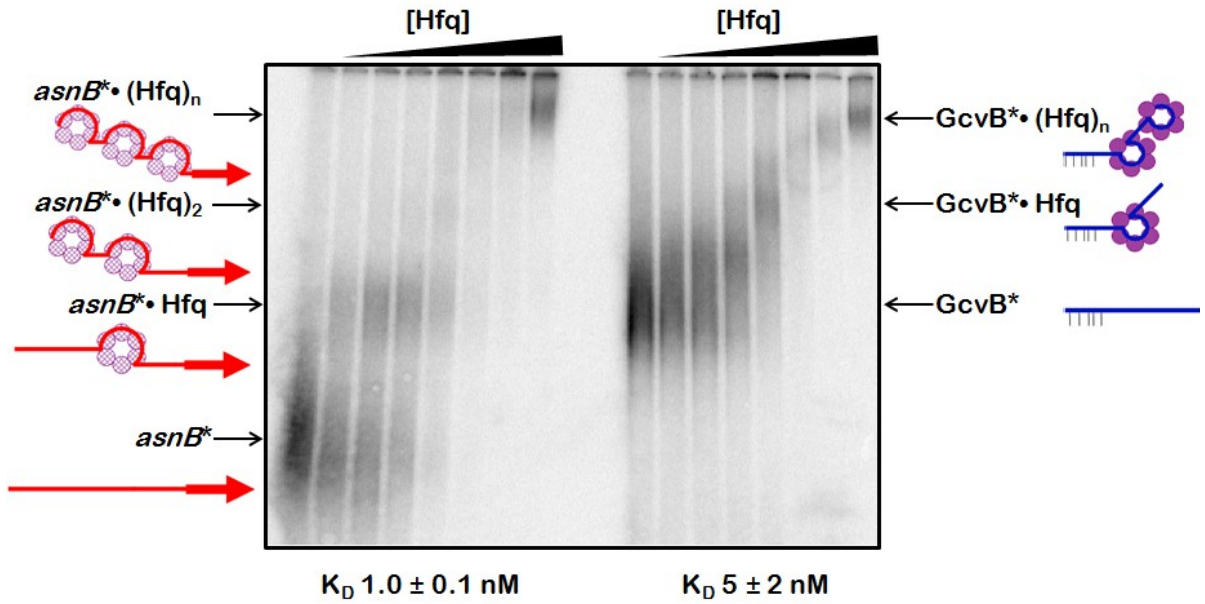


Figure 27. Binding Affinities of Hfq with *asnB* and GcvB. Native gel shifts were performed by titrating labeled *asnB* or GcvB with increasing concentrations of Hfq (0 – 1.3 μM hexamer). Data was fit to a cooperative binding model and the K_D's were determined to be 1.0 ± 0.1 nM for *asnB* and, for 5 ± 2 nM GcvB.

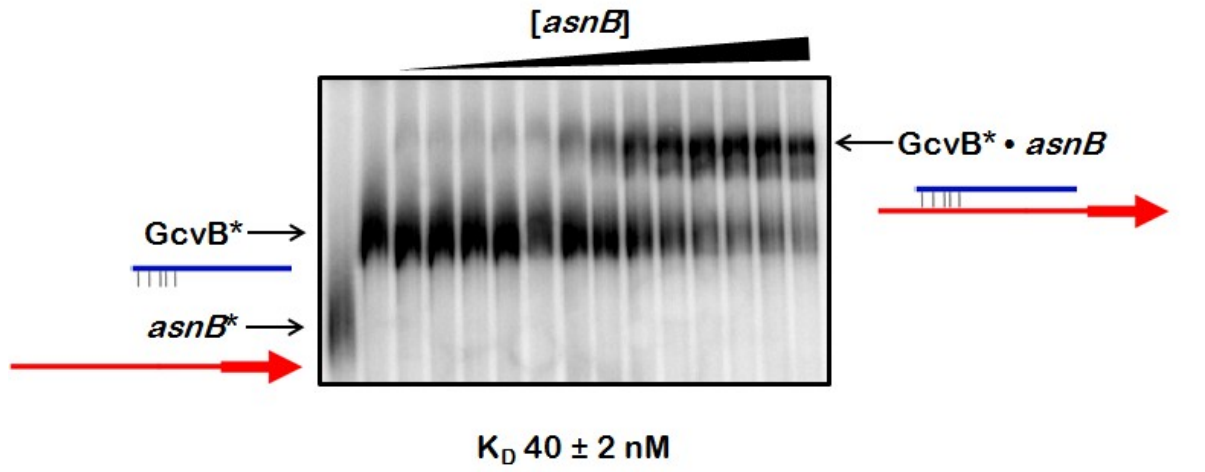


Figure 28. *asnB* • GcvB Duplex Formation. Native gel shift was performed by titrating GcvB* with increasing concentrations of unlabeled *asnB* (0 – 7.2 μM). Data was fit to determine a K_D of 40. ± 4 nM.

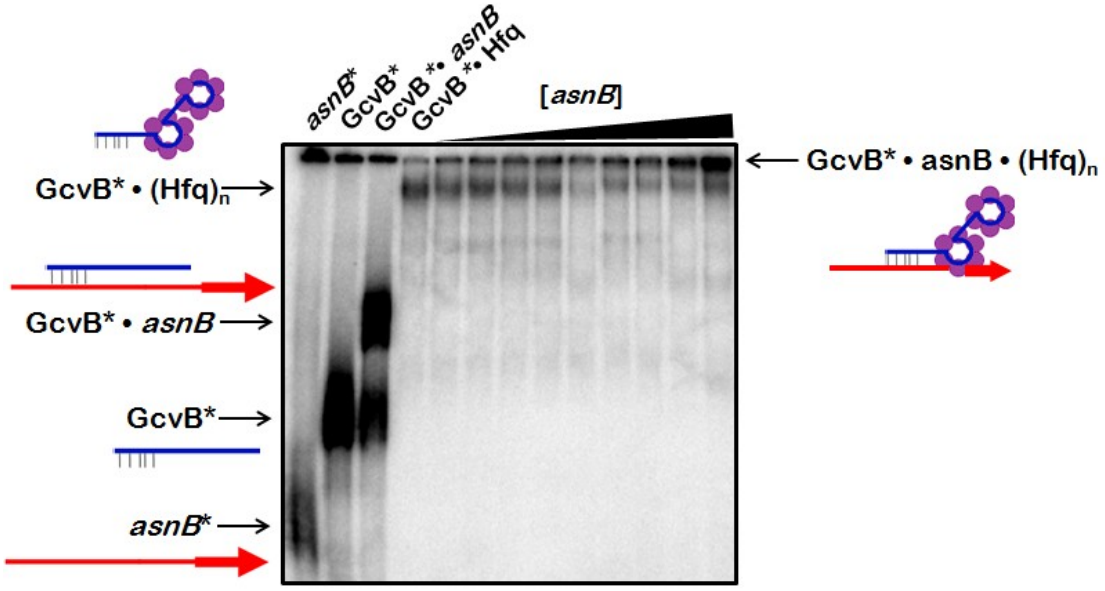


Figure 29. Ternary Complex Formation. EMSA was performed to observe the ability of *asnB*, GcvB and Hfq to form a ternary complex. Labeled GcvB was prebound to 1 μM Hfq. This complex was titrated with increasing amounts of unlabeled *asnB*.

K_D for K56A to be 11 ± 3 nM and for Y25A Hfq to be 6 ± 2 nM (Figure 30). Compared to wild type Hfq both mutants resulted in a similar decrease in affinity. This suggests that *asnB* interacts with Hfq through both the proximal and distal faces using the wrap around model [37, 84]. In this model the mRNA is able to bind both faces of Hfq and therefore a mutation of only one of the binding faces would not significantly hinder binding.

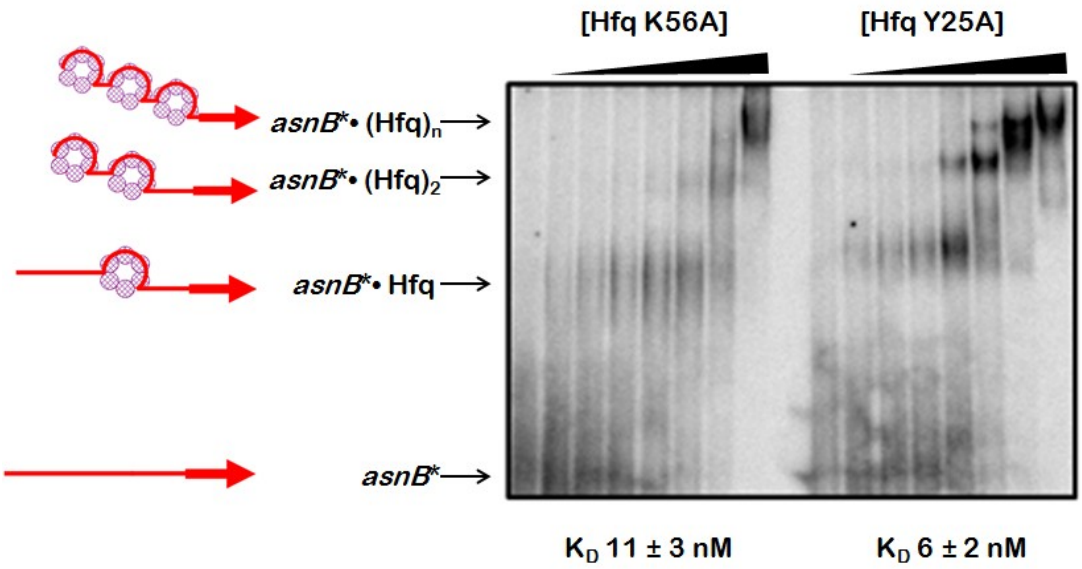
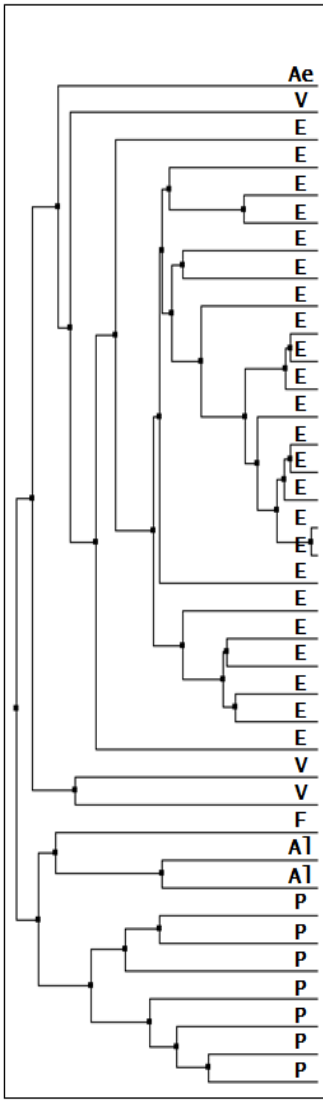


Figure 30. Ability of asnB to Bind Hfq Mutants. Native gel shifts were performed by titrating labeled *asnB* or GcvB with increasing concentrations of Hfq (0 – 1.3 μM hexamer). The K_D*'s* were determined to be 6 ± 2 nM for Y25A Hfq and, 11 ± 3 nM for K56A Hfq.

A goal of our computational approach was not only to annotate targets in *E. coli* but to be able to extend the technique to other bacteria of interest. We envision broadening our approach by *de novo* identification of targets in the organism based on the presence of an Hfq binding site. Alternatively, we hypothesize that targets identified in *E. coli* are likely to be regulated in other bacteria, especially genes involved in key metabolic or homeostatic processes. If that idea is correct then *E. coli* targets could be directly considered as potential targets in other organisms. After targets are identified, they need to be validated and an sRNA that regulates them needs to be identified. The

process of identifying sRNAs in more exotic bacteria is in its infancy therefore the knowledge of existing sRNAs may not be available. We hypothesize that a viable means of identifying these sRNAs may be to use the 5'UTR of targets to search for complementary sequences in the genome that display characteristics of an sRNA (location in an intergenic region, rho-independent terminator, absence of an ORF).

To explore this possibility we searched for homologs of GcvB and *asnB* in other bacteria to assess the extent of conservation of the RNA interaction sites and the $(ARN)_x$ sites. GcvB is one of the most conserved sRNAs and is always encoded divergently from GcvA making it relatively easy to find in distant organisms[151]. A stone stepping approach, described below, can be used were homologs are identified directly from a BLAST of the *E. coli* GcvB sequence. One then takes the most divergent of those sequences to perform a new search and so on until no more homologs are apparent. The searches often output only portions of the homolog at which point rho-independent terminators and transcription initiation sequences can be used to identify the full length transcript. The conserved synteny of GcvB and GcvA also provides another approach. By locating the GcvA sequence, which is well conserved, one can then look adjacent to it for the marks of a non-coding RNA sequence. A combination of these approaches allowed us to identify GcvB in a diverse group of bacteria that are represented on a phylogenetic tree created from the GcvB sequences (Figure 31). All but one homolog identified belonged to the γ-Proteobacteria class in the orders Pasteurellales, Vibrionales, Alteromonadales, and Enterobacteriales. The other homolog was identified in *Candidatus arthromitus* which is a Firmicute of the order Clostridia. All species also



		AsnB	GcvB BS	*asnB* BS	Hybridization energy (kcal/mol)
Ae	Aeromonas.hydrophila	✓	1	UTR	3
V	Vibrio.harveyi	✓	-	-	-
E	Sodalis.glossinidius	✓	1	ORF	5
E	Erwinia.pyrifoliae	✓	1	UTR	9
E	Pectobacterium.carotovorum	✓	1	ORF	7
E	Dickeya.dadantii	✓	-	-	-
E	Edwardsiella.ictaluri	✓	1	UTR	5
E	Serratia.marcescens	✓	1	UTR	26
E	Rahnella.aquatilis	✓	3	UTR	5
E	Raoultella.ornithinolytica	✓	1	UTR	17
E	Enterobacter.aerogenes	✓	2/3	UTR	20
E	Klebsiella.oxytoca	✓	2/3	UTR	18
E	Cronobacter.turicensis	✓	1	UTR	23
E	Enterobacteriaceae.bacterium	✓	1	UTR	17
E	Salmonella.enterica	✓	1	UTR	18
E	Citrobacter.rodentium	✓	1	UTR	8
E	Escherichia.coli	✓	1	UTR	22
E	Shigella.flexneri	✓	1	UTR	22
E	Pantoea.ananatis	✓	2/3	UTR	8
E	Yersinia.pestis	✓	1	UTR	15
E	Xenorhabdus.bovienii	✓	1	ORF	3
E	Providencia.stuarti	-	-	-	-
E	Morganella.morganii	-	-	-	-
E	Proteus.mirabilis	-	-	-	-
E	Arsenophonus.nasoniae	-	-	-	-
V	Aliivibrio.salmonicida	✓	2/3	UTR	1
V	Photobacterium.profundum	✓	2/3	UTR	6
F	Candidatus.arthromitus	✓	2	ORF	7
Al	Pseudoalteromonas	✓	2/3	UTR	8
Al	Idiomarina.loihiensis	✓	1	ORF	8
P	Pasteurella.multocida	-	-	-	-
P	Aggregatibacter.aphrophilus	-	-	-	-
P	Gallibacterium.anatis	✓	1	UTR	3
P	Bibersteinia.trehalosi	-	-	-	-
P	Mannheimia.haemolytica	-	-	-	-
P	Actinobacillus.pleuropneumoniae	-	-	-	-
P	Haemophilus.parasuis	-	-	-	-

Figure 31. Phylogenetic Tree Based on the Sequence of GcvB. The tree was created by determining the percent similarity of the sequences. Single letters to the left of the species names represent the order. Ae = Aeromonadales V = Vibrionales E = Enterbacteriales F = Firmicute, Clostridia Al = Alteromonas P = Pasteurellales. Species with a check mark have an AsnB homolog. The GcvB binding sites (BS), R1, R2, and R3 that bind to the *asnB* homolog in that species are noted as 1,2, or 3. The *asnB* binding sites (BS) that bind to the GcvB homolog in that species are annotated as being in the untranslated region (UTR) or the open reading frame (ORF). The hybridization energies listed were determined for the interactions between species specific asnB and GcvB homologs using IntaRNA.

had an Hfq homolog, except for *Candidatus arthromitus,* even though Hfq is found in other *Candidatus* bacteria. The main binding sites R1 and R2 are well conserved with

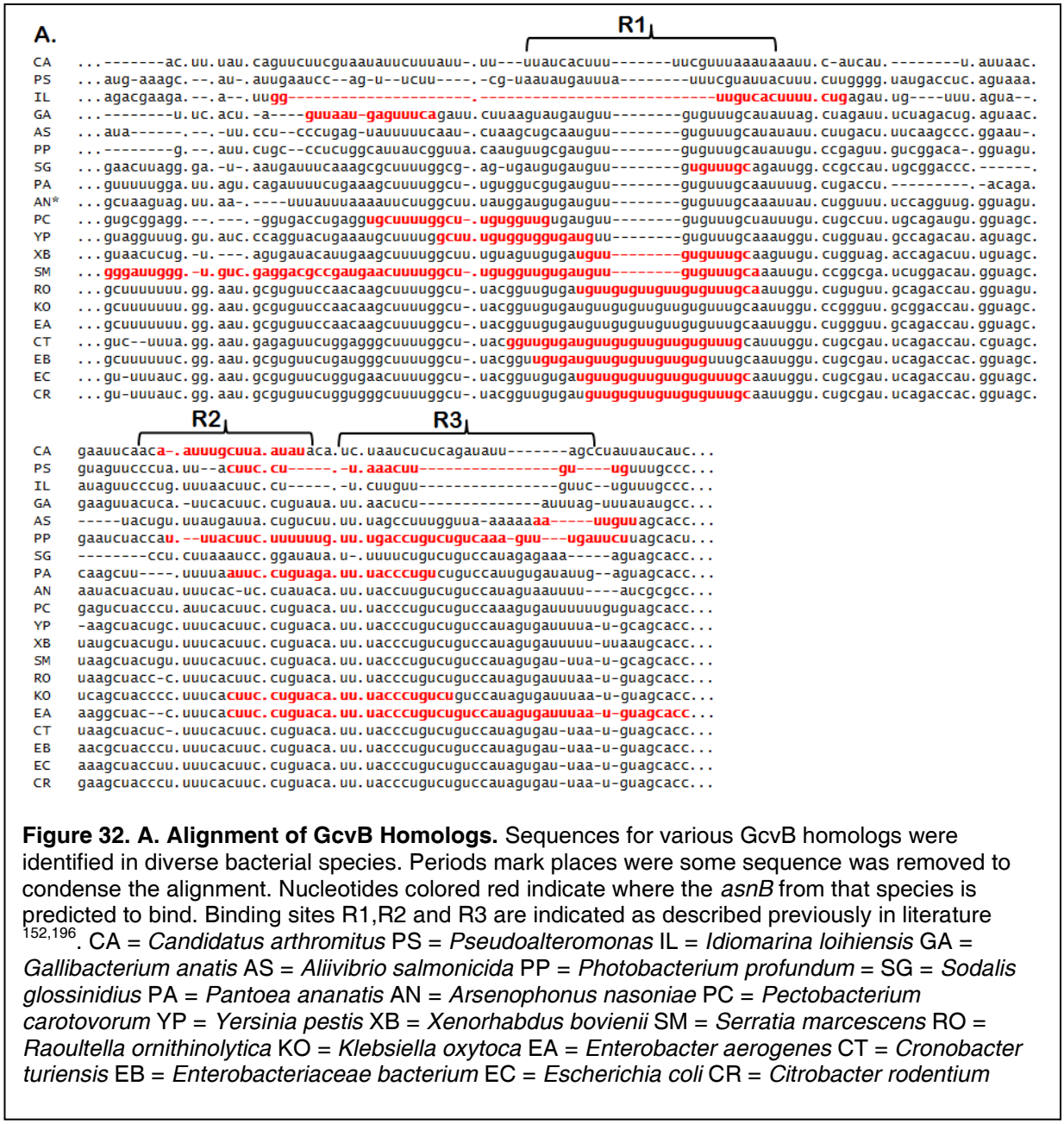


Figure 32. A. Alignment of GcvB Homologs. Sequences for various GcvB homologs were identified in diverse bacterial species. Periods mark places were some sequence was removed to condense the alignment. Nucleotides colored red indicate where the *asnB* from that species is predicted to bind. Binding sites R1,R2 and R3 are indicated as described previously in literature [152,196]. CA = *Candidatus arthromitus* PS = *Pseudoalteromonas* IL = *Idiomarina loihiensis* GA = *Gallibacterium anatis* AS = *Aliivibrio salmonicida* PP = *Photobacterium profundum* = SG = *Sodalis glossinidius* PA = *Pantoea ananatis* AN = *Arsenophonus nasoniae* PC = *Pectobacterium carotovorum* YP = *Yersinia pestis* XB = *Xenorhabdus bovienii* SM = *Serratia marcescens* RO = *Raoultella ornithinolytica* KO = *Klebsiella oxytoca* EA = *Enterobacter aerogenes* CT = *Cronobacter turiensis* EB = *Enterobacteriaceae bacterium* EC = *Escherichia coli* CR = *Citrobacter rodentium*

R2 demonstrating slightly better conservation as determined by a structurally based alignment of the GcvB sequences (Figure 26c and Figure 32).

Having identified many GcvB homologs we then sought to determine if AsnB is also present in those same organisms. If both GcvB and AsnB are present, then the

sRNA regulation pathway may be conserved. We used BLASTp to perform a homology search for AsnB and the mRNA sequence from -45 to +60 was extracted for further investigation. In general, these sequences were easier to identify than GcvB due the higher degree of conservation in the ORF. Interestingly, not all organisms that had GcvB also had AsnB (Figure 31). Most of the organisms missing the gene belonged to the Pasteurellales order. These organisms may be dependent on AsnA or a tRNA-dependent transamidation pathway for asparagine synthesis [196]. GcvB must be conserved in these bacteria to regulate other targets in its large regulon.

To investigate the conservation of the interaction between GcvB and *asnB,* multiple sequence alignments were created and the predicted regions of base pairing were annotated (Figure 26c, Figure 32 and Figure 33). If the GcvB interaction sites located in the *asnB* mRNA maintain recognizable features that would suggest that the site may be useful in identifying a complementary sRNA. To determine the interaction site we again made use of the IntaRNA program to predict hybridization energies for the interactions between the *asnB* and GcvB RNAs from different bacteria [119]. Interactions with strong hybridization energies and that are most likely conserved are restricted to Enterobacteriaceae (Figure 31). Interactions with weaker energies may represent reduced strength of regulation or, for energies below five Kcal/mol, no regulation at all. Within the group that have strong interactions (>15 Kcal/mol) most interact at R1 as seen in *E. coli* but a couple are predicted to make use of R2/R3 (Figure 26c, Figure 31and Figure 32). These two mRNAs from *Klebsiella oxytoca* and *Enterobacter aerogenes* have lost the core interaction motif for R1, CACAaCAY, but do have a GA rich section

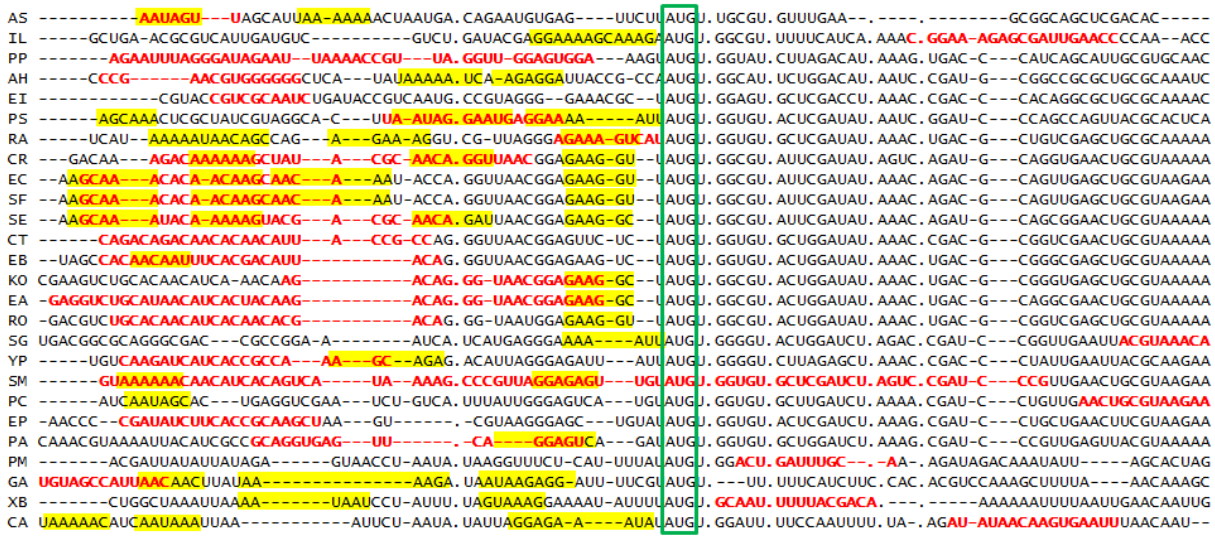


Figure 33. Alignment of *asnB* Homologs. Sequences for various *asnB* homologs were identified in diverse bacterial species. Periods mark places were some sequence was removed to condense the alignment. Red nucleotides indicate where the GcvB from that species is predicted to bind. The start codon is outlined by the green box. $(ARN)_x$ sites are highlighted in yellow. Species abbreviations are as defined in Figure 32.

slightly down stream capable of interacting with R2/3 (Figure 33). All of the strong interactions also remain in the UTR, therefore interactions predicted in the ORF may not be biologically relevant. $(ARN)_x$ motifs were present in all strong interactions except for one, *Cronobacter turiensis* (Figure 31 and Figure 33). The *asnB*-GcvB interaction predicted in this species was the second strongest observed and therefore may have a less stringent requirement for Hfq. The strongest predicted interaction was in *Serratia marcescens* due to a region of complementarity that extended the entire length of the 5'UTR and well into the coding region (Figure 33). An extended complementarity between SgrS and its mRNA target was also observed in this species and may be a hallmark of pairing specific to the organism [197].

4.3 DISCUSSION

The duplex predicted to form between *asnB* and GcvB reflects the canonical interaction between the sRNA and its other known targets. The interaction takes place between the R1 site of GcvB and a CA rich motif at the start of the 5'UTR of *asnB*. Most of the interactions between GcvB and its other targets are at the R1 site [112]. A strong hybridization energy of 22 Kcal/mol between *asnB* and GcvB is predicted, and accounts for the tight K_D of duplex formation. The duplex formation site lies within the 30S ribosomal binding region of translation initiation [198]. The mechanism of down regulation is most likely inhibition of translation.

The SHAPE derived secondary structure of *asnB* revealed two potentially important $(ARN)_x$ motifs. The sites are located across from each other in an internal loop of the 5'UTR. The most 5' site overlaps with the predicted GcvB binding site, a potentially undesirable feature for complex formation [144]. In this case though, the seed of the interaction does not overlap with the site so one could imagine that Hfq binds to facilitate an initial annealing step after which it dissociates to allow extended complementarity to form. The role of the two $(ARN)_x$ sites will become more clear with further investigation.

We investigated the binding properties of the newly discovered regulatory pair *asnB*-GcvB. Using EMSA we determined that *asnB* and GcvB bind to Hfq with high affinity, further validating the ability of our bioinformatic approach to find Hfq binding RNAs. The strong binding observed is similar to other known Hfq binding RNAs [37, 84, 87, 98]. The two RNAs alone form a duplex structure with a K_D of 40±2 nM. The requirement for Hfq to facilitate duplex formation varies among RNA partners with some

not requiring Hfq when sufficient RNA concentrations are reached [20]. The strong binding between *asnB* and GcvB may explain why a significant amount of regulation occurs even in the absence of Hfq. We also observed the formation of a ternary complex containing the two RNAs and Hfq, indicative of a regulatory complex. *asnB* most likely binds Hfq according to the wrap around model, as evident by its ability to bind Hfq mutants that are defective in binding at either its distal or proximal face. The presence of a single stranded AU rich stretch, just after the start codon in the SHAPE structure is characteristic of a proximal binding site and supports this wrap around model hypothesis. In sum, the *asnB* mRNA exhibits properties consistent with previously characterized target mRNAs.

An important goal of our bioinformatic approach is to be able to find targets in other bacterial species. One way to do this is to tailor the search to account for species specific differences in Hfq binding. For example, the Hfq homolog in *Bacillus subtilis* binds RNAs with an (RL) motif at its distal site [39]. Alternatively, we could use the targets discovered in *E. coli* as starting points in the identification of targets and sRNA in other bacteria. The basis of this approach is that the ORFs of proteins show a greater amount of conservation than sRNAs do across diverse species. One idea then is that species that have AsnB, for example, may also have a GcvB homolog. Identification of a putative GcvB site in the *asnB* mRNA could then be used to search for GcvB. There is some question of how conserved the hybridization sites in mRNA are. One study on this subject suggests that in general, they are not well conserved [153]. However, accessibility at these sites remains largely conserved. Also, the sequence of the interaction might be conserved but the actual location may move around. If the

interaction sites are not conserved enough to aid in the identification of sRNAs by sequence the mRNA targets may still be useful to identify functional homologs. Functional homologs of RyhB that have no sequence similarity to the *E. coli* sRNA but regulate similar targets in distant bacterial species have been identified [154]. This relatedness illustrates the idea that targets found in *E. coli* will also be regulated by sRNAs in other species, even if regulation is not mediated by the same sRNA.

To explore the idea of using the conservation of mRNAs to predict targets in other bacteria, we investigated the preservation of GcvB and *asnB*. GcvB is an ideal starting point due to its high degree of conservation [151]. Therefore, we didn't actually need AsnB to find a diverse group of homologs, which allowed us to observe the conservation of the hybridization site in a significant number of bacteria. Strong predicted interactions between homologs of GcvB and *asnB* throughout *Enterobacteriales* were identified. These species demonstrate a significant level of divergence in 16S rRNA sequences, yet have likely maintained the interaction between *asnB* and GcvB. Weaker apparent interactions were identified in species from other orders of bacteria (*Alteromonadales* and *Vibrionales*). The redundancy of the three binding sites of GcvB also added an interesting feature to the study. While most binding partners maintained the interaction at the R1 site some switched to the R2/3 site. The switch was due to a loss of the CA rich motif but the presence of a GA rich site that can bind to R2/3 just downstream. Interactions at this site had both strong and mild predicted interactions. Overall, a CA or GA binding motif was present in the 5'UTR of *asnB* from a diverse range of species although the location of the site did shift. Nevertheless, the hybridization site could be used to search the genomes of other

bacterial species to help identify GcvB homologs. Besides a simple genomic BLAST search the candidates could first be narrowed down using a program that identifies potential sRNAs based on key features[199]. These candidates could then be mined for sequences that complement the mRNA in that species.

While the degree of sRNA conservation is limited due to rapid evolution many of the mRNA targets remain conserved. For example, the sRNA RyhB regulates iron homeostasis and is widely conserved throughout enteric bacteria[200-203]. In more distant bacteria, such as *B. subtilis* and *P. aeruginosa*, RyhB homologs have not been found but functionally homologous sRNAs have been characterized [154, 204]. Another interesting example is the regulation of *glmS*, an mRNA that codes for the protein glucosamine-6-phosphate (GlcN6P) synthase, which is discussed in detail in Chapter 5 [139]. This mRNA is regulated by a self-cleaving ribozyme in gram-positive bacteria and by a trans-sRNA in gram-negative bacteria [139, 205]. So there are many examples where an mRNA is regulated by RNA across the bacterial kingdom even though the mechanism of that regulation or the identity of that riboregulator changes. We were able to show that the interaction between *asnB* and GcvB is conserved even though the position of the site shifts. This result suggests that using the mRNA targets to predict sRNAs may be useful but future experiments with other targets will need to be performed to fully assess this tool. It is also important to continue studying the features of $(ARN)_x$ sites as well as Hfq binding in other species to create a more robust approach.

4.4 MATERIALS AND METHODS

4.4.1 Preparation of asnB for SHAPE

DNA corresponding to the 5'UTR of *asnB* from -45 to +60 was PCR amplified and inserted into the SHAPE parent plasmid pMM110003 as described previously in section 2.5.3. Cloning was verified by sequencing. Transcription, purification and SHAPE analysis was performed as described previously in section 2.5.5.

4.4.2 EMSA

Native gels were cast and run as described previously in section 3.4.5. RNAs were labeled and folded as described previously. EMSA for RNA binding to Hfq and mutant Hfq was performed by titrating trace labeled RNA with increasing amounts of Hfq from 0 to 1300 nM. Binding was allowed to occur at room temperature for 25 minutes before resolving on the gel. Gels were dried and phosphorimaged. Images were quantified and data was fit to a cooperative binding model to determine the K_D. For duplex formation, a trace amount of labeled GcvB was titrated with unlabeled *asnB* from 0 to 7.2 μM. RNAs were allowed to bind at room temperature for 30 minutes before loading onto the gel. Ternary complexes were formed by first pre-binding trace amounts of labeled GcvB with 1 μM Hfq. Pre-bound complex was then titrated with increasing amounts of unlabeled *asnB* from 0 to 7.2 μM. Complexes were allowed to form at room temperature for 30 minutes before running on the gel.

4.4.3 Homology and Phylogenetic Analysis

GcvB homologs were identified by performing BLASTn searches using the GcvB sequence from *E. coli* and other organism or by performing a BLASTp search for GcvA.

When only portions of the RNA were found, the full length sequence was determined by identifying the rho-independent terminator and transcription initiation sequences. *asnB* mRNA homologs were identified by performing BLASTp search for AsnB. Sequence alignments were created using CLUSTAL omega. Alignments were edited and phylogenetic trees created using Jalview [206].

CHAPTER FIVE: CHARACTERIZATION OF *glmS*-HFQ INTERACTIONS[‡]

5.1 INTRODUCTION

The discovery of the $(ARN)_x$ motif occurred when the RNA-Hfq interactions with *rpoS* and *fhlA* were characterized along with the determination of a crystal structure showing A_{18} bound to Hfq [31, 36, 37]. Chapter 2 described the bioinformatic characterization of $(ARN)_x$ sites in the 5'UTRs of messages known to be regulated by Hfq, where we found them to be ubiquitous. It is essential to thoroughly investigate $(ARN)_x$ motifs in additional mRNAs to better understand their role in Hfq mediated regulation. Therefore, our lab set out to investigate the $(ARN)_x$ motif of another target mRNA. This study was headed by Nilshad Salim, with me contributing in a significant way.

We chose to study the mRNA *glmS* that codes for the protein glucosamine-6-phosphate (GlcN6P) synthase, a key enzyme in cell wall synthesis [139]. This protein is regulated in both gram-positive and gram-negative bacteria, albeit by two different forms of riboregulation. Gram-positive bacteria employ a ribozyme in the 5'UTR of *glmS* that promotes self cleavage upon GlcN6P binding [205]. In gram-negative bacteria, the translation of this mRNA is regulated by two sRNAs, Hfq, and an additional protein called RapZ (Figure 34) [207]. In *E. coli*, *glmS* is transcribed as part of the dicistron *glmUS.* Transcription is followed by RNaseE processing in the stop codon of *glmU.* The

[‡] Portions of this work were previously published in Nucleic Acids Research. Salim, N. N.; Faner, M. A.; Feig, A. L.; Requirement of upstream Hfq-Binding (ARN)x elements in glmS and the Hfq C-terminal region for GlmS upregulation by sRNAs GlmZ and GlmY. *Nucleic Acids Res.* 2012, *40*, 8021-8032.

separation of the two transcripts facilitates *glmS* specific regulation [163]. When levels of GlcN6P are high both transcripts are susceptible to degradation after processing occurs. In addition to a high turnover rate, the translation of *glmS* is naturally repressed due to an inhibitory structure that masks the RBS. The regulatory sRNA, GlmZ, stabilizes *glmS* and can directly up-regulate translation by base pairing with *glmS* to release the RBS; this process is dependent upon Hfq. Levels of GlmZ are controlled by RNaseE mediated decay (as part of the degradosome) [207]. The decay process requires a second protein, RapZ, to bind GlmZ and recruit RNaseE through protein-protein interactions [207]. When cellular levels of glucosamine-6-phosphate are low, expression of a second sRNA, GlmY is increased. GlmY shares structural features with GlmZ but does not have a *glmS* interaction site [163]. GlmY takes advantage of this structural similarity to sequester RapZ and protect GlmZ from decay [163]. Therefore, GlmY can indirectly activate *glmS* by recruiting the degradation machinery away

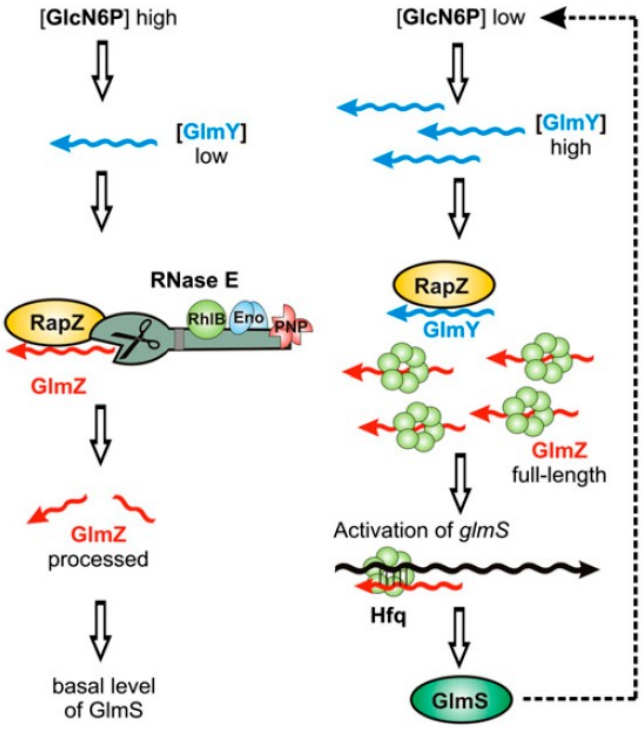


Figure 34. Regulation of GlmS. When levels of GlcN6P are high GlmZ is targeted for degradation by RapZ who recruits RNaseE. When GlmZ is processed *glmS* is repressed due to rapid RNA turnover and an inhibitory structure that blocks the RBS. When levels of GlcN6P are low GlmY is transcribed and sequesters RapZ away from binding GlmZ. GlmZ can then derepress *glmS* by binding to altering the structure of the 5'UTR. Image used from[208].

from GlmZ.

Another unanswered question in the field is the role of the C-terminal domain (CTD) of Hfq. Hfq contains two RNA binding motifs, SM1 and SM2, which are highly conserved (Figure 36). In contrast, the C-terminal extension varies in length from nine to thirty-seven amino acids and is not well conserved. In particular, Hfq homologs from gram-positive organisms have short, variable CTDs. Results from experiments that tested the ability of C-terminal truncations of the *E. coli* Hfq are contradictory. One report suggests that truncations are proficient in facilitating regulation, while another suggests that the CTD may bind to longer RNA molecules and/or increase interaction specificity by recognizing additional motifs within an RNA but that is not involved in regulation [40, 208]. There is evidence that Hfq binds to a number of proteins, many of which are part of the degradosome, but no protein interactions have been reported with the SM cores, suggesting a potential role for the C-terminal domain in protein-protein interactions [55, 56].

5.2 RESULTS

To determine the structural features of the 5'UTR of *glmS* we performed SHAPE. The reactivities that we calculated are graphed in Figure 35a, and correspond nicely with the structure determined using RNAstructure with high reactivity nucleotides mapping to areas that are unpaired. We found that there are two potential $(ARN)_x$ motifs that are located in a single stranded region between two stem loops (Figure 35b). This arrangement reflects the structural context of other $(ARN)_x$ motifs in regulated mRNAs. The $(ARN)_x$ motifs are approximately forty nucleotides away from the GlmZ interaction site therefore there may be some tertiary interaction that brings the motifs closer in

proximity to make ternary complex formation more favorable [143]. Hfq footprinting by SHAPE was also performed to see if Hfq binds to the predicted $(ARN)_x$ motifs (Figure 35). Protection of the nucleotides at both sites indicate that Hfq does bind the $(ARN)_x$ sites. We tested the binding of this mRNA to wild type Hfq and Hfq defective in distal (Y25A) versus proximal (K56A) binding to determine which Hfq face was involved in mRNA binding (Figure 35b). We saw an increased sensitivity to NMIA when the distal

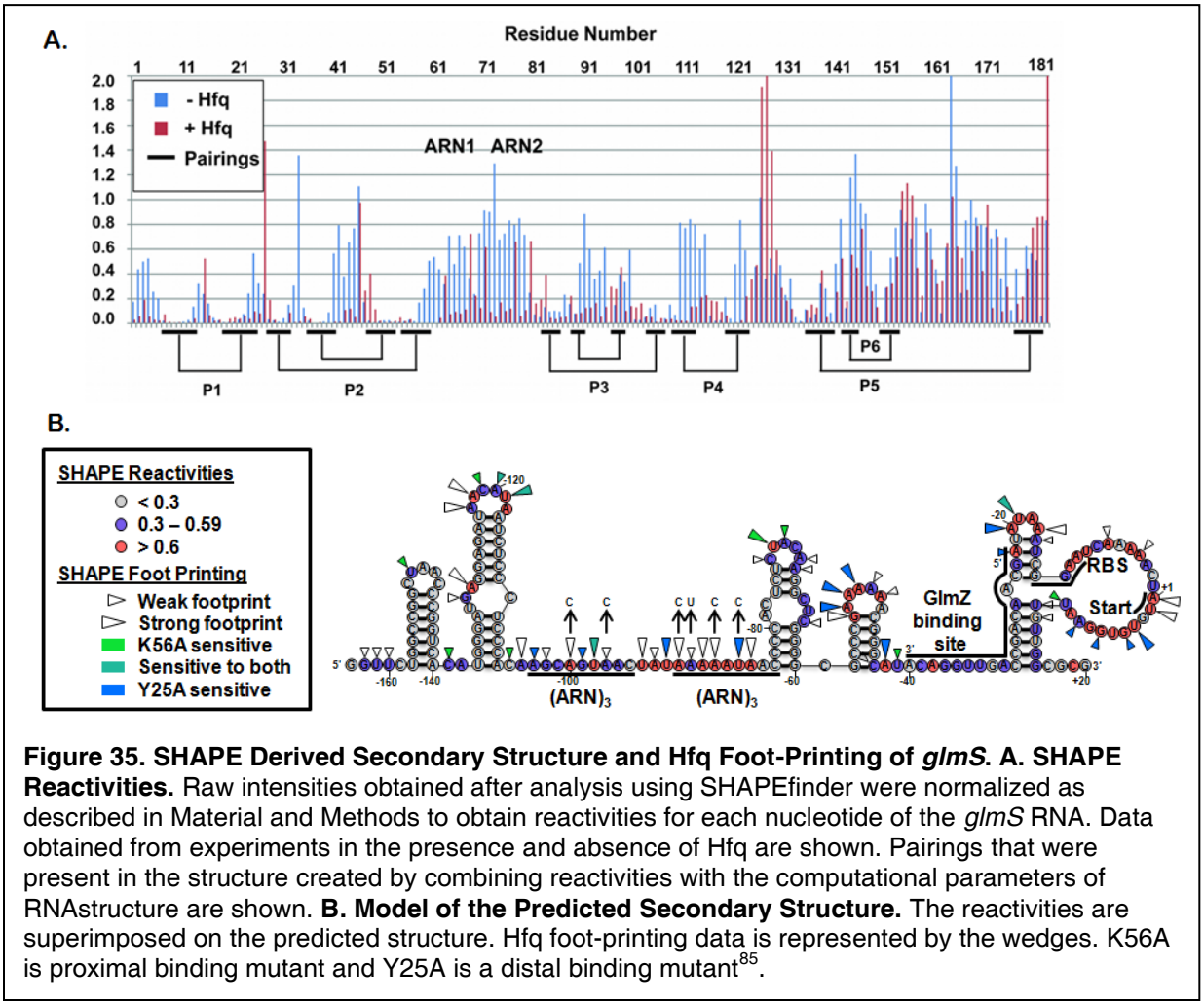


Figure 35. SHAPE Derived Secondary Structure and Hfq Foot-Printing of *glmS*. A. SHAPE Reactivities. Raw intensities obtained after analysis using SHAPEfinder were normalized as described in Material and Methods to obtain reactivities for each nucleotide of the *glmS* RNA. Data obtained from experiments in the presence and absence of Hfq are shown. Pairings that were present in the structure created by combining reactivities with the computational parameters of RNAstructure are shown. **B. Model of the Predicted Secondary Structure.** The reactivities are superimposed on the predicted structure. Hfq foot-printing data is represented by the wedges. K56A is proximal binding mutant and Y25A is a distal binding mutant[85].

mutant was bound, suggesting that the distal face of Hfq binds to the $(ARN)_x$ sites.

Nilshad Salim then went on to use a GFP reporter assay, similar to the one described in Chapter 3 to study the regulation of *glmS* by the sRNAs GlmZ and

GlmY[84]. He showed that the GlmS-GFP fusion was up-regulated by both GlmZ and GlmY in agreement with previous studies [139, 163]. The dependence of regulation on Hfq was demonstrated by performing the assay in the absence of Hfq or with Hfq mutants deficient in RNA binding. Efficient regulation was not observed in those cases confirming that Hfq is required. The same assay was then employed to determine the necessity of the two potential $(ARN)_x$ motifs. Fusion constructs were created where either site was mutated of both were mutated at the same time. Results of this experiment indicated that first $(ARN)_x$ motif is necessary for *glmS* regulation by GlmZ.

The role of the Hfq C-terminal extension was also investigated in this study. Hfq is present in about half of all bacteria and the nucleic acid binding motifs are widely conserved (Figure 36). The C-terminal portion of the protein, on the other hand, is variable, with *E. coli* having a relatively long extension and *Clostridium perfringens* and *Clostridium difficile* a short one. The role of the C-terminal extension has been widely debated, therefore we investigated the potential role of the domain in the regulation of *glmS*. Using the GFP assay, Nilshad Salim found that truncated versions of *E. coli* Hfq, full length *C. perfringens* Hfq, and full length *C. difficile* Hfq were unable to facilitate regulation of *glmS* by GlmY in *E. coli* [84]. Based on this observation, we were curious to see if there was a difference in the ability of GlmY and GlmZ to bind to *E. coli* Hfq versus *C. perfringens* Hfq. EMSA was performed with each of the two sRNAs and the two different Hfqs (Figure 37a, GlmY only shown). We found that GlmY and GlmZ both bind *E. coli* Hfq with similar affinities, but GlmY had a significant defect when binding *C. perfringens* Hfq, which suggests that the way GlmZ and GlmY bind Hfq is different and

that the CTD may play a role in facilitating regulation by non-traditional sRNAs like GlmY (Figure 37b).

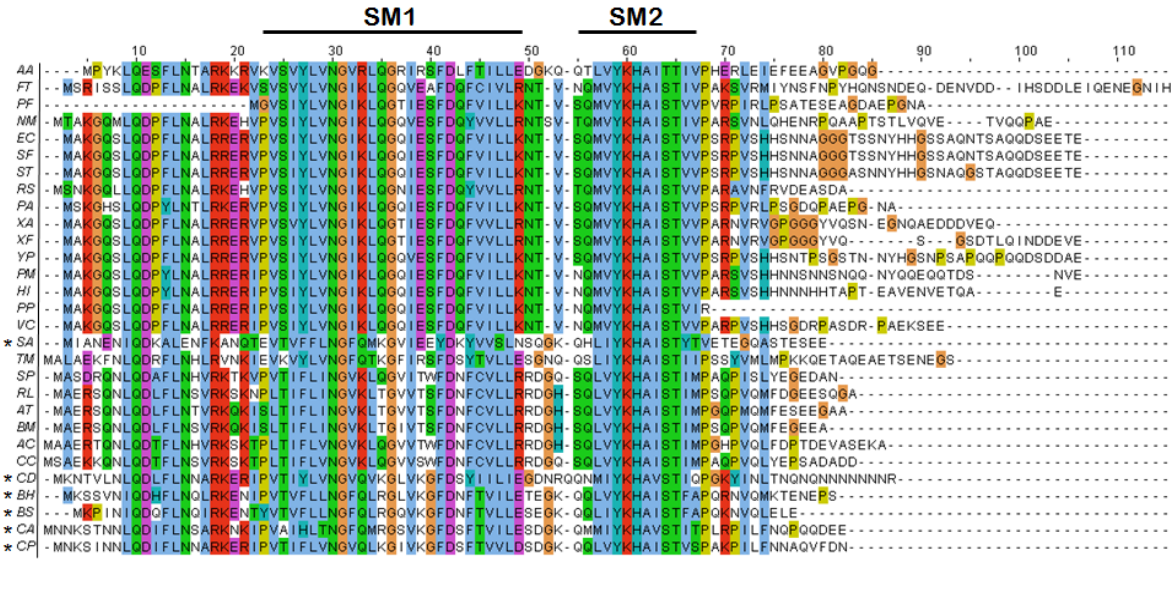


Figure 36. Alignment of Hfq Amino Acid Sequences From Diverse Bacterial Species. Sequences were aligned using CLUSTAL Omega. SM1 and SM2 are the conserved nucleic acid binding domains. Species with an * are gram positive. SA = *Staphylococcus aureus;* AT = *Agrobacterium tumefaciens;* AA = *Aquifex aeolicus;* AC = *Azorhizobium caulinodans;* BH = *Bacillus halodurans;* BS = *Bacillus subtilis;* BM = *Brucella melitensis;* CC = *Caulobacter crescentus;* CA = *Clostridium acetobutylicum;* CP = *Clostridium perfringens;* EC = *Escherichia coli;* HI = *Haemophilus influenza;* NM = *Neisseria meningitides;* PM = *Pasteurella multocida;* PP = *Photobacterium profundum;* PA = *Pseudomonas aeruginosa;* RS = *Ralstonia solanacearum;* RL = *Rhizobium loti;* ST = *Salmonella typhimurium;* SF = *Shigella flexneri;* TM = *Thermotoga maritime;* VC = *Vibrio cholera;* XA = *Xanthomonas axonopodis;* XF = *Xylella fastidiosa;* YP = *Yersinia pestis;* PF = *Pseudomonas fluorescens;* FT = *Francisella tularensis;* SM = *Silicibacter pomeroyi;* CD = *Clostridium difficile*

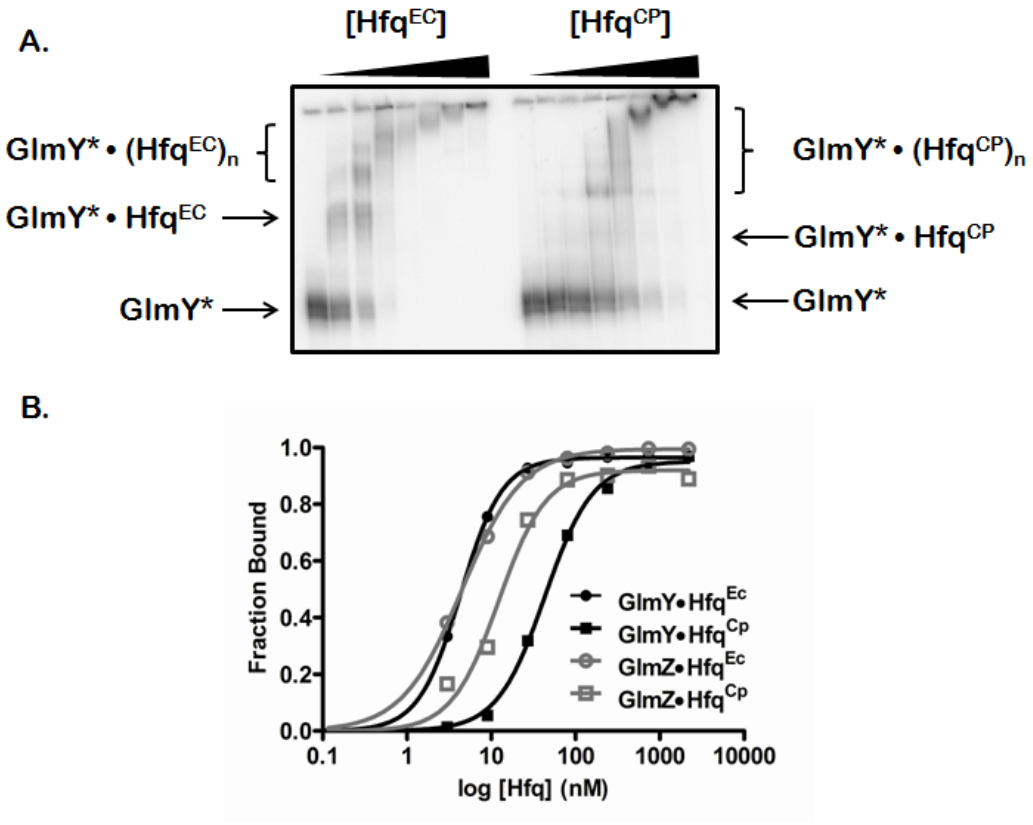


Figure 37. EMSA of GlmZ and GlmY binding to Hfq. A. Native gels were performed by titrating labeled sRNA with increasing amounts of either *E. coli* Hfq or *C. perfringens* Hfq. **B.** K_D determination for the gelshifts[85].

5.3 DISCUSSION

We studied the role of $(ARN)_x$ motifs in *glmS* regulation. A SHAPE derived structure revealed two potential $(ARN)_x$ motifs that were single stranded and surrounded by regions of structure. The structural context observed was similar to those shown in (Figure 16a). We observed Hfq footprinting at those sites and the loss of the footprint when a distal binding Hfq mutant was used indicates that Hfq binds to both $(ARN)_x$ motifs through distal surface interactions. The role of multiple $(ARN)_x$ sites has been an ongoing topic of interest throughout this thesis. The lead author on this work addressed the question by mutating each site separately and observing the ability of the message to be regulated in a GFP assay. We found that the site closest to the GlmZ interaction site was necessary for regulation while the other was not. So in this case only one motif was functional. It will be interesting to investigate similar cases in the future to see if a pattern emerges regarding the number of functional motifs and what features make one more important than another.

Nilshad Salim used a GFP assay to demonstrate that a truncated version of *E. coli* Hfq and Hfq from *C. perfringens* and *C. difficile*, which are naturally truncated, cannot facilitate the regulation of *glmS* by GlmY. This led us to further investigate the ability of GlmZ and GlmY to bind Hfq from *E. coli* and *C. perfringens*. A more significant defect in GlmY-Hfq^{CP} but not in GlmZ-Hfq^{CP} was observed using EMSA, leading to the conclusion that the two sRNAs bind differently to Hfq and that the C-terminal extension may play a more significant role in GlmY binding. This hypothesis suggests that the C-terminal region of Hfq may be important for Hfq activities beyond facilitating base-pair formation. It is interesting that *trans*-sRNA regulation of *glmS* occurs in *E. coli* and other

gram-negative bacteria, while riboswitch regulation occurs in gram-positive species. In fact, this duality fits a general trend where trans-sRNAs are much more prevalent in gram-negative bacteria and riboswitches are more common in gram-positive bacteria [16, 48, 205, 209]. It is not clear why the two have evolved different strategies but there are some key differences in RNA degradation and translation initiation that may contribute. Gram-positive bacteria use different degradation machinery than gram-negative, and while sRNAs can regulate translation, the regulation leads to RNA degradation less often than in gram-negative bacteria [10]. The mRNAs in Gram-positive species often have strong Shine-Delgarno sequences and the ribosomal protein S1 does not participate in translation initiation [10]. Another key difference is the role of Hfq in *trans*-sRNA regulation. Hfq is essential for this type of regulation in gram-negative bacteria but not always in gram-positive, although Hfq has been less well studied in gram-positive bacteria and it may be too early to draw conclusions [16, 48, 210, 211]. The difference in regulation of *glmS* may be related to the differences in Hfq observed between gram-negative (especially enterobacteria) and gram-positive bacteria. One could speculate that there may have been no evolutionary pressure to conserve the CTD in gram-positive organisms due to differences in regulation or the CTD was lost for some other reason leading to the evolution of riboswitch control for *glmS*.

In conclusion, the work described in this thesis has contributed to the overall knowledge in the field of *trans*-sRNAs in bacteria. Specifically, we more thoroughly defined the nature and the role of $(ARN)_x$ motifs in the 5'UTRs of regulated messages using a unique combination of *in silico, in vitro*, and *in vivo* techniques. This information

has increased the understanding of how the sRNA network functions and the role of Hfq in facilitating it. In addition, we were able use the characteristics $(ARN)_x$ motifs to develop a novel approach for target mRNA identification. This approach was able to identify *bona fide* targets with diverse and important functions in *E. coli*, one of which we went on to characterize extensively. Our technique is adaptable to other bacteria and can help to further increase our understanding of sRNA regulation in pathogens and other bacteria of importance.

5.4 MATERIALS AND METHODS

5.4.1 *glmS* SHAPE Analysis

SHAPE was performed on *glmS* as described in Chapter 2 Materials and Methods.

5.4.2 EMSA of GlmZ and GlmY

RNAs were amplified to include the entire sequence of the sRNA using primers with a T7 promoter incorporated. Transcription was performed using the PCR product as a template and purified on a denaturing PAGE gel. RNAs were ^{32}P labeled by first dephosphorylating with calf intestinal alkaline phosphatase and then phosphorylating with T4 polynucleotide kinase in the presence of ATP gamma ^{32}P. RNAs were gel purified. In preparation for binding, the RNAs (amount determined to provide 15,000CPM per lane) were heated to 95 °C for 3 minutes in 50 mM Tris-HCl pH 8, and 100 mM KCl, followed by cooling at room temperature for 15 minutes. Then 10 mM $MgCl_2$ was added, followed by Hfq, and the reaction was incubated at room temperature for 30 minutes. Hfq was added in varying amounts to achieve concentrations from 0 to

2.1 μM. Data were using a cooperative binding model using the equation: $Q_{bound} = [Hfq]^n / (K_d)^n + [Hfq]^n$

CHAPTER 6: CONCLUSIONS

In conclusion, the work described in this thesis has contributed to the overall knowledge in the field of *trans*-sRNAs in bacteria. Specifically, we more thoroughly defined the nature and the role of $(ARN)_x$ motifs in the 5'UTRs of regulated messages using a unique combination of *in silico, in vitro*, and *in vivo* techniques. This diversity of techniques provided many benefits. A computational approach allowed us to increase efficiency by identifying a set of high value potential target mRNAs from an entire genome. With that group in hand we didn't have to waste time or resources on mRNAs that had little chance of being real targets. As more knowledge about $(ARN)_x$ sites, sRNA interaction sites, and RNA processing events is acquired it can easily be incorporated into the existing framework to improve the hit rate of the bioinformatics scheme. The other value to the bioinformatics approach is its adaptability to other organisms and the potential to study regulons from pathogenic or difficult to grow bacteria in *E. coli*. Criteria in the work flow of the technique can be changed to accommodate species specific characteristics for any aspect, like Hfq-binding. Targets identified bioinformatically in other organisms can be validated in *E. coli* using the GFP reporter assay. The *in vitro* techniques that we used, SHAPE and EMSA, were particularly suited for validation and to study specific targets in more detail. SHAPE analysis of known target mRNAs allowed us to characterize the $(ARN)_x$ motif and also gave us confidence that computational folding was sufficient to identify single stranded

$(ARN)_x$ motifs on a genome wide scale. We also used SHAPE to determine accurate structures for known targets (*asnB, glmS*) that we wanted to investigate more thoroughly. EMSA is most useful for a detailed binding analysis. We initially tried to incorporate it into the validation stage but the ability of Hfq to bind RNA in general and the artificial *in vitro* conditions led to misleading results. We were able to determine that some of the predicted targets could bind Hfq and heat anneal to an sRNA partner but that did not necessarily mean that they were regulated *in vivo*, which was the most important question that we were asking. EMSA was very useful when defining the binding properties of targets of interest (*asnB, glmS*). Our goal was to use bioinformatics to predict novel target mRNAs and the best way to validate these targets was with an *in vivo* reporter assay. It was critical to study the predicted regulon in a cellular environment to be able to conclude that a *bona fide* regulatory event was occurring. A large number of the fusion constructs suffered from low fluorescence levels and were not testable; a problem that will certainly have to be addressed in the future. Regardless it did allow us to identify five new targets, for an overall 63% positive prediction rate.

In addition to using $(ARN)_x$ motif features as a bioinformatic tool, our characterization has increased the understanding of how the sRNA network functions and the role of Hfq in facilitating it. We hypothesize that $(ARN)_x$ motifs play an important role in facilitating the rapid response to stress and environmental conditions that is a feature of sRNA regulation. As a target mRNA is being transcribed it begins to fold and, if an $(ARN)_x$ motif is present, then Hfq can immediately bind the high affinity site. This interaction serves to mark the message for regulation. When a cognate sRNA is

transcribed in response to stress and it encounters the Hfq-mRNA complex the regulatory ternary complex can form immediately. The importance of the $(ARN)_x$ motif will become clearer as more studies are completed. Future work in the field holds the promise of learning about all aspects of sRNA regulation that make it so interesting and how we can use our knowledge in the pursuit of increasing human health and happiness.

REFERENCES

[1] Crick, F. Central dogma of molecular biology, Nature 227 1970, 561-563.

[2] Jacob, F., and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins, J Mol Biol 3 1961, 318-356.

[3] Britten, R. J., and Davidson, E. H. Gene regulation for higher cells: a theory, Science 165 1969, 349-357.

[4] Gilbert, W., and Muller-Hill, B. Isolation of the lac repressor, Proc Natl Acad Sci U S A 56 1966, 1891-1898.

[5] Mattick, J. S., and Makunin, I. V. Non-coding RNA, Hum Mol Genet 15 Spec No 1 2006, R17-29.

[6] Sorek, R., Kunin, V., and Hugenholtz, P. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea, Nat Rev Microbiol 6 2008, 181-186.

[7] Fedor, M. J., and Williamson, J. R. The catalytic diversity of RNAs, Nat Rev Mol Cell Biol 6 2005, 399-412.

[8] Siomi, H., and Siomi, M. C. On the road to reading the RNA-interference code, Nature 457 2009, 396-404.

[9] Mattick, J. S. The central role of RNA in human development and cognition, FEBS Lett 585 2011, 1600-1616.

[10] Romby, P., and Charpentier, E. An overview of RNAs with regulatory functions in gram-positive bacteria, Cell Mol Life Sci 67 2009, 217-237.

[11] Waters, L. S., and Storz, G. Regulatory RNAs in bacteria, Cell 136 2009, 615-628.

[12] Babitzke, P., and Romeo, T. CsrB sRNA family: sequestration of RNA-binding regulatory proteins, Curr Opin Microbiol 10 2007, 156-163.

[13] Breaker, R. R. Prospects for riboswitch discovery and analysis, Mol Cell 43 2011, 867-879.

[14] Grundy, F. J., and Henkin, T. M. From ribosome to riboswitch: control of gene expression in bacteria by RNA structural rearrangements, Crit Rev Biochem Mol Biol 41 2006, 329-338.

[15] Brantl, S. Regulatory mechanisms employed by cis-encoded antisense RNAs, Curr Opin Microbiol 10 2007, 102-109.

[16] Storz, G., Vogel, J., and Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers, Mol Cell 43 2011, 880-891.

[17] Aiba, H. Mechanism of RNA silencing by Hfq-binding small RNAs, Curr Opin Microbiol 10 2007, 134-139.

[18] Fender, A., Elf, J., Hampel, K., Zimmermann, B., and Wagner, E. G. RNAs actively cycle on the Sm-like protein Hfq, Genes Dev 24 2010, 2621-2626.

[19] Maki, K., Morita, T., Otaka, H., and Aiba, H. A minimal base-pairing region of a bacterial small RNA SgrS required for translational repression of ptsG mRNA, Mol Microbiol 76 2010, 782-792.

[20] Soper, T., Mandin, P., Majdalani, N., Gottesman, S., and Woodson, S. A. Positive regulation by small RNAs and the role of Hfq, Proc Natl Acad Sci U S A 107 2010, 9602-9607.

[21] Doetsch, M., Stampfl, S., Furtig, B., Beich-Frandsen, M., Saxena, K., Lybecker, M., and Schroeder, R. Study of E. coli Hfq's RNA annealing acceleration and duplex

destabilization activities using substrates with different GC-contents, Nucleic Acids Res 41 2013, 487-497.

[22] Lee, T., and Feig, A. The RNA-Protein Complexes of *E. coli* Hfq: Form and Function, In (Walter, N., Woodson, S., and Batey, R., Eds.) Non-Protein Coding RNAs Elsevier, New York 2008, pp 249-272.

[23] Chao, Y., and Vogel, J. The role of Hfq in bacterial pathogens, Curr Opin Microbiol 13 2012, 24-33.

[24] Ansong, C., Yoon, H., Porwollik, S., Mottaz-Brewer, H., Petritis, B. O., Jaitly, N., Adkins, J. N., McClelland, M., Heffron, F., and Smith, R. D. Global systems-level analysis of Hfq and SmpB deletion mutants in Salmonella: implications for virulence and global protein translation, PLoS One 4 2009, e4809.

[25] Christiansen, J. K., Larsen, M. H., Ingmer, H., Sogaard-Andersen, L., and Kallipolitis, B. H. The RNA-binding protein Hfq of Listeria monocytogenes: role in stress tolerance and virulence, J Bacteriol 186 2004, 3355-3362.

[26] Tsui, H. C., Leung, H. C., and Winkler, M. E. Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in Escherichia coli K-12, Mol Microbiol 13 1994, 35-49.

[27] Sittka, A., Pfeiffer, V., Tedin, K., and Vogel, J. The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium, Mol Microbiol 63 2007, 193-217.

[28] Ding, Y., Davis, B. M., and Waldor, M. K. Hfq is essential for Vibrio cholerae virulence and downregulates sigma expression, Mol Microbiol 53 2004, 345-354.

[29] Horstmann, N., Orans, J., Valentin-Hansen, P., Shelburne, S. A., 3rd, and Brennan, R. G. Structural mechanism of Staphylococcus aureus Hfq binding to an RNA A-tract, Nucleic Acids Res 40 2012, 11023-11035.

[30] Mikulecky, P. J., Kaw, M. K., Brescia, C. C., Takach, J. C., Sledjeski, D. D., and Feig, A. L. Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs, Nat Struct Mol Biol 11 2004, 1206-1214.

[31] Link, T. M., Valentin-Hansen, P., and Brennan, R. G. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA, Proc Natl Acad Sci U S A 106 2009, 19292-19297.

[32] Schumacher, M. A., Pearson, R. F., Moller, T., Valentin-Hansen, P., and Brennan, R. G. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein, EMBO J 21 2002, 3546-3556.

[33] Brescia, C. C., Mikulecky, P. J., Feig, A. L., and Sledjeski, D. D. Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure, RNA 9 2003, 33-43.

[34] Otaka, H., Ishikawa, H., Morita, T., and Aiba, H. PolyU tail of rho-independent terminator of bacterial small RNAs is essential for Hfq action, Proc Natl Acad Sci U S A 108 2011, 13059-13064.

[35] Sauer, E., and Weichenrieder, O. Structural basis for RNA 3'-end recognition by Hfq, Proc Natl Acad Sci U S A 108 2011, 13065-13070.

[36] Soper, T. J., and Woodson, S. A. The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA, RNA 14 2008, 1907-1917.

[37] Salim, N. N., and Feig, A. L. An upstream Hfq binding site in the fhlA mRNA leader region facilitates the OxyS-fhlA interaction, PLoS One 5 2010, e13028.

[38] Chen, C. Y., and Shyu, A. B. AU-rich elements: characterization and importance in mRNA degradation, Trends Biochem Sci 20 1995, 465-470.

[39] Someya, T., Baba, S., Fujimoto, M., Kawai, G., Kumasaka, T., and Nakamura, K. Crystal structure of Hfq from Bacillus subtilis in complex with SELEX-derived RNA aptamer: insight into RNA-binding properties of bacterial Hfq, Nucleic Acids Res 40 2012, 1856-1867.

[40] Beich-Frandsen, M., Vecerek, B., Konarev, P. V., Sjoblom, B., Kloiber, K., Hammerle, H., Rajkowitsch, L., Miles, A. J., Kontaxis, G., Wallace, B. A., Svergun, D. I., Konrat, R., Blasi, U., and Djinovic-Carugo, K. Structural insights into the dynamics and function of the C-terminus of the E. coli RNA chaperone Hfq, Nucleic Acids Res 39 2011, 4900-4915.

[41] Sauer, E., Schmidt, S., and Weichenrieder, O. Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition, Proc Natl Acad Sci U S A 109 2012, 9396-9401.

[42] Vincent, H. A., Henderson, C. A., Ragan, T. J., Garza-Garcia, A., Cary, P. D., Gowers, D. M., Malfois, M., Driscoll, P. C., Sobott, F., and Callaghan, A. J. Characterization of Vibrio cholerae Hfq provides novel insights into the role of the Hfq C-terminal region, J Mol Biol 420 2012, 56-69.

[43] Beisel, C. L., and Storz, G. Base pairing small RNAs and their roles in global regulatory networks, FEMS Microbiol Rev 34 2010, 866-882.

[44] Papenfort, K., and Vogel, J. Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level, Res Microbiol 160 2009, 278-287.

[45] Adamson, D. N., and Lim, H. N. Essential requirements for robust signaling in Hfq dependent small RNA networks, PLoS Comput Biol 7 2011, e1002138.

[46] Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C. M., and Vogel, J. An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs, EMBO J 31 2012, 4005-4019.

[47] Sittka, A., Lucchini, S., Papenfort, K., Sharma, C. M., Rolle, K., Binnewies, T. T., Hinton, J. C., and Vogel, J. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq, PLoS Genet 4 2008, e1000163.

[48] Majdalani, N., Vanderpool, C. K., and Gottesman, S. Bacterial small RNA regulators, Crit Rev Biochem Mol Biol 40 2005, 93-113.

[49] Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G., and Gottesman, S. Global analysis of small RNA and mRNA targets of Hfq, Mol Microbiol 50 2003, 1111-1124.

[50] Hussein, R., and Lim, H. N. Disruption of small RNA signaling caused by competition for Hfq, Proc Natl Acad Sci U S A 108 2011, 1110-1115.

[51] Moon, K., and Gottesman, S. Competition among Hfq-binding small RNAs in Escherichia coli, Mol Microbiol 82 2011, 1545-1562.

[52] Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J.,

Greenblatt, J., and Emili, A. Interaction network containing conserved and essential protein complexes in Escherichia coli, Nature 433 2005, 531-537.

[53] Carabetta, V. J., Silhavy, T. J., and Cristea, I. M. The response regulator SprE (RssB) is required for maintaining poly(A) polymerase I-degradosome association during stationary phase, J Bacteriol 192 2010, 3713-3721.

[54] Hu, P., Janga, S. C., Babu, M., Diaz-Mejia, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., Chandran, S., Christopoulos, C., Nazarians-Armavil, A., Nasseri, N. K., Musso, G., Ali, M., Nazemof, N., Eroukova, V., Golshani, A., Paccanaro, A., Greenblatt, J. F., Moreno-Hagelsieb, G., and Emili, A. Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins, PLoS Biol 7 2009, e96.

[55] Mohanty, B. K., Maples, V. F., and Kushner, S. R. The Sm-like protein Hfq regulates polyadenylation dependent mRNA decay in Escherichia coli, Mol Microbiol 54 2004, 905-920.

[56] Morita, T., Maki, K., and Aiba, H. RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs, Genes Dev 19 2005, 2176-2186.

[57] Carmichael, G. G., Weber, K., Niveleau, A., and Wahba, A. J. The host factor required for RNA phage Qbeta RNA replication in vitro. Intracellular location, quantitation, and purification by polyadenylate-cellulose chromatography, J Biol Chem 250 1975, 3607-3612.

[58] Kajitani, M., Kato, A., Wada, A., Inokuchi, Y., and Ishihama, A. Regulation of the Escherichia coli hfq gene encoding the host factor for phage Q beta, J Bacteriol 176 1994, 531-534.

[59] Masse, E., Vanderpool, C. K., and Gottesman, S. Effect of RyhB small RNA on global iron use in Escherichia coli, J Bacteriol 187 2005, 6962-6971.

[60] Papenfort, K., Pfeiffer, V., Mika, F., Lucchini, S., Hinton, J. C., and Vogel, J. SigmaE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay, Mol Microbiol 62 2006, 1674-1688.

[61] Vanderpool, C. K., and Gottesman, S. Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system, Mol Microbiol 54 2004, 1076-1089.

[62] Vogel, J., and Luisi, B. F. Hfq and its constellation of RNA, Nat Rev Microbiol 9 2011, 578-589.

[63] Vogel, J., Bartels, V., Tang, T. H., Churakov, G., Slagter-Jager, J. G., Huttenhofer, A., and Wagner, E. G. RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria, Nucleic Acids Res 31 2003, 6435-6443.

[64] Christiansen, J. K., Nielsen, J. S., Ebersbach, T., Valentin-Hansen, P., Sogaard-Andersen, L., and Kallipolitis, B. H. Identification of small Hfq-binding RNAs in Listeria monocytogenes, RNA 12 2006, 1383-1396.

[65] Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C., and Storz, G. The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs, Mol Cell 9 2002, 11-22.

[66] Sonnleitner, E., Sorger-Domenigg, T., Madej, M. J., Findeiss, S., Hackermuller, J., Huttenhofer, A., Stadler, P. F., Blasi, U., and Moll, I. Detection of small RNAs in Pseudomonas aeruginosa by RNomics and structure-based bioinformatic tools, Microbiology 154 2008, 3175-3187.

[67] Ramos, C. G., Grilo, A. M., da Costa, P. J., and Leitao, J. H. Experimental identification of small non-coding regulatory RNAs in the opportunistic human pathogen Burkholderia cenocepacia J2315, Genomics2012, 1-10.

[68] Messer, L. I., Currey, K. M., O'Neill, B. J., Maizel, J. V., Jr., Levin, J. G., and Gerwin, B. I. Functional analysis of reverse transcription by a frameshift pol mutant of murine leukemia virus, Virology 146 1985, 146-152.

[69] Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays, Genes Dev 15 2001, 1637-1651.

[70] Vanderpool, C. K., Balasubramanian, D., and Lloyd, C. R. Dual-function RNA regulators in bacteria, Biochimie 93 2011, 1943-1949.

[71] Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps, Nature 460 2009, 479-486.

[72] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 2010, 129-141.

[73] Milek, M., Wyler, E., and Landthaler, M. Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing, Semin Cell Dev Biol 23 2012, 206-212.

[74] Ule, J., Jensen, K., Mele, A., and Darnell, R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells, Methods 37 2005, 376-386.

[75] Geng, J., Song, Y., Yang, L., Feng, Y., Qiu, Y., Li, G., Guo, J., Bi, Y., Qu, Y., Wang, W., Wang, X., Guo, Z., Yang, R., and Han, Y. Involvement of the post-transcriptional regulator Hfq in Yersinia pestis virulence, PLoS One 4 2009, e6213.

[76] Roscetto, E., Angrisano, T., Costa, V., Casalino, M., Forstner, K. U., Sharma, C. M., Di Nocera, P. P., and De Gregorio, E. Functional Characterization of the RNA Chaperone Hfq in the Opportunistic Human Pathogen Stenotrophomonas maltophilia, J Bacteriol 194 2012, 5864-5874.

[77] Guisbert, E., Rhodius, V. A., Ahuja, N., Witkin, E., and Gross, C. A. Hfq modulates the sigmaE-mediated envelope stress response and the sigma32-mediated cytoplasmic stress response in Escherichia coli, J Bacteriol 189 2007, 1963-1973.

[78] Barra-Bily, L., Fontenelle, C., Jan, G., Flechard, M., Trautwetter, A., Pandey, S. P., Walker, G. C., and Blanco, C. Proteomic alterations explain phenotypic changes in Sinorhizobium meliloti lacking the RNA chaperone Hfq, J Bacteriol 192 2010, 1719-1729.

[79] Lorenz, C., Gesell, T., Zimmermann, B., Schoeberl, U., Bilusic, I., Rajkowitsch, L., Waldsich, C., von Haeseler, A., and Schroeder, R. Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts, Nucleic Acids Res 38 2010, 3794-3808.

[80] Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L., and Storz, G. A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator, Cell 90 1997, 43-53.

[81] Balbontin, R., Fiorini, F., Figueroa-Bossi, N., Casadesus, J., and Bossi, L. Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in Salmonella enterica, Mol Microbiol 78 2010, 380-394.

[82] Kierzek, E., Kierzek, R., Turner, D. H., and Catrina, I. E. Facilitating RNA structure prediction with microarrays, Biochemistry 45 2006, 581-593.

[83] Papenfort, K., Bouvier, M., Mika, F., Sharma, C. M., and Vogel, J. Evidence for an autonomous 5' target recognition domain in an Hfq-associated small RNA, Proc Natl Acad Sci U S A 107 2010, 20435-20440.

[84] Salim, N. N., Faner, M. A., Philip, J. A., and Feig, A. L. Requirement of upstream Hfq-binding (ARN)x elements in glmS and the Hfq C-terminal region for GlmS upregulation by sRNAs GlmZ and GlmY, Nucleic Acids Res 40 2012, 8021-8032.

[85] Soper, T. J., Doxzen, K., and Woodson, S. A. Major role for mRNA binding and restructuring in sRNA recruitment by Hfq, RNA 17 2011, 1544-1550.

[86] Lease, R. A., and Woodson, S. A. Cycling of the Sm-like protein Hfq on the DsrA small regulatory RNA, J Mol Biol 344 2004, 1211-1223.

[87] Olejniczak, M. Despite similar binding to the Hfq protein regulatory RNAs widely differ in their competition performance, Biochemistry 50 2011, 4427-4440.

[88] Lee, T., and Feig, A. L. The RNA binding protein Hfq interacts specifically with tRNAs, RNA 14 2008, 514-523.

[89] Kawamoto, H., Koide, Y., Morita, T., and Aiba, H. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq, Mol Microbiol 61 2006, 1013-1022.

[90] Jing, M., and Bowser, M. T. Methods for measuring aptamer-protein equilibria: a review, Anal Chim Acta 686, 9-18.

[91] Katsamba, P. S., Park, S., and Laird-Offringa, I. A. Kinetic studies of RNA-protein interactions using surface plasmon resonance, Methods 26 2002, 95-104.

[92] Feig, A. L. Studying RNA-RNA and RNA-protein interactions by isothermal titration calorimetry, Methods Enzymol 468 2009, 409-422.

[93] Perrin. Polarisation de la lumiere de fluorescence vie moyenne des molecules dans l'etat excite., journal de physique et le radium 7 1926.

[94] Arluison, V., Mutyam, S. K., Mura, C., Marco, S., and Sukhodolets, M. V. Sm-like protein Hfq: location of the ATP-binding site and the effect of ATP on Hfq-- RNA complexes, Protein Sci 16 2007, 1830-1841.

[95] Sun, X., and Wartell, R. M. Escherichia coli Hfq binds A18 and DsrA domain II with similar 2:1 Hfq6/RNA stoichiometry using different surface sites, Biochemistry 45 2006, 4875-4887.

[96] Updegrove, T. B., Correia, J. J., Chen, Y., Terry, C., and Wartell, R. M. The stoichiometry of the Escherichia coli Hfq protein bound to RNA, RNA 17 2011, 489-500.

[97] Wang, W., Wang, L., Zou, Y., Zhang, J., Gong, Q., Wu, J., and Shi, Y. Cooperation of Escherichia coli Hfq hexamers in DsrA binding, Genes Dev 25 2011, 2106-2117.

[98] Geissmann, T. A., and Touati, D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator, EMBO J 23 2004, 396-405.

[99] Desnoyers, G., Morissette, A., Prevost, K., and Masse, E. Small RNA-induced differential degradation of the polycistronic mRNA iscRSUA, EMBO J 28 2009, 1551-1561.

[100] Lindell, M., Romby, P., and Wagner, E. G. Lead(II) as a probe for investigating RNA structure in vivo, RNA 8 2002, 534-541.

[101] Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE), J Am Chem Soc 127 2005, 4223-4231.

[102] Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M., and Laederach, A. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis, Nucleic Acids Res 36 2008, e63.

[103] Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C., and Weeks, K. M. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states, PLoS Biol 6 2008, e96.

[104] Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M., and Giddings, M. C. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis, RNA 14 2008, 1979-1990.

[105] Reuter, J. S., and Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis, BMC Bioinformatics 11 2010, 129.

[106] Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), Proc Natl Acad Sci U S A 108, 11063-11068.

[107] Spitale, R. C., Crisalli, P., Flynn, R. A., Torre, E. A., Kool, E. T., and Chang, H. Y. RNA SHAPE analysis in living cells, Nat Chem Biol 9 2013, 18-20.

[108] Fratczak, A., Kierzek, R., and Kierzek, E. Isoenergetic microarrays to study the structure and interactions of DsrA and OxyS RNAs in two- and three-component complexes, Biochemistry 50 2011, 7647-7665.

[109] Chen, S., Zhang, A., Blyn, L. B., and Storz, G. MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli, J Bacteriol 186 2004, 6689-6697.

[110] Jorgensen, M. G., Nielsen, J. S., Boysen, A., Franch, T., Moller-Jensen, J., and Valentin-Hansen, P. Small regulatory RNAs control the multi-cellular adhesive lifestyle of Escherichia coli, Mol Microbiol 84 2012, 36-50.

[111] Frohlich, K. S., Papenfort, K., Berger, A. A., and Vogel, J. A conserved RpoS-dependent small RNA controls the synthesis of major porin OmpD, Nucleic Acids Res 40 2012, 3623-3640.

[112] Sharma, C. M., Papenfort, K., Pernitzsch, S. R., Mollenkopf, H. J., Hinton, J. C., and Vogel, J. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA, Mol Microbiol 81 2011, 1144-1165.

[113] Bailey, T. L., and Gribskov, M. Methods and statistics for combining motif match scores, J Comput Biol 5 1998, 211-221.

[114] Tjaden, B. TargetRNA: a tool for predicting targets of small RNA action in bacteria, Nucleic Acids Res 36 2008, W109-113.

[115] Li, W., Ying, X., Lu, Q., and Chen, L. Predicting sRNAs and Their Targets in Bacteria, Genomics Proteomics Bioinformatics 10 2012, 276-284.

[116] Hebrard, M., Kroger, C., Srikumar, S., Colgan, A., Handler, K., and Hinton, J. C. sRNAs and the virulence of Salmonella enterica serovar Typhimurium, RNA Biol 9 2012, 437-445.

[117] Postic, G., Dubail, I., Frapy, E., Dupuis, M., Dieppedale, J., Charbit, A., and Meibom, K. L. Identification of a novel small RNA modulating Francisella tularensis pathogenicity, PLoS One 7 2012, e41999.

[118] Zhao, Y., Li, H., Hou, Y., Cha, L., Cao, Y., Wang, L., Ying, X., and Li, W. Construction of two mathematical models for prediction of bacterial sRNA targets, Biochem Biophys Res Commun 372 2008, 346-350.

[119] Busch, A., Richter, A. S., and Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions, Bioinformatics 24 2008, 2849-2856.

[120] Eggenhofer, F., Tafer, H., Stadler, P. F., and Hofacker, I. L. RNApredator: fast accessibility-based prediction of sRNA targets, Nucleic Acids Res 39 2011, W149-154.

[121] Ying, X., Cao, Y., Wu, J., Liu, Q., Cha, L., and Li, W. sTarPicker: a method for efficient prediction of bacterial sRNA targets based on a two-step model for hybridization, PLoS One 6 2011, e22705.

[122] Modi, S. R., Camacho, D. M., Kohanski, M. A., Walker, G. C., and Collins, J. J. Functional characterization of bacterial sRNAs using a network biology approach, Proc Natl Acad Sci U S A 108 2011, 15522-15527.

[123] Mika, F., Busse, S., Possling, A., Berkholz, J., Tschowri, N., Sommerfeldt, N., Pruteanu, M., and Hengge, R. Targeting of csgD by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in Escherichia coli, Mol Microbiol 84 2012, 51-65.

[124] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, PLoS Biol 5 2007, e8.

[125] Masse, E., and Gottesman, S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli, Proc Natl Acad Sci U S A 99 2002, 4620-4625.

[126] Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H., and Altuvia, S. Novel small RNA-encoding genes in the intergenic regions of Escherichia coli, Curr Biol 11 2001, 941-950.

[127] Udekwu, K. I., Darfeuille, F., Vogel, J., Reimegard, J., Holmqvist, E., and Wagner, E. G. Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA, Genes Dev 19 2005, 2355-2366.

[128] Mandin, P., and Gottesman, S. Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA, EMBO J 29 2010, 3094-3107.

[129] De Lay, N., and Gottesman, S. A complex network of small non-coding RNAs regulate motility in Escherichia coli, Mol Microbiol 86 2012, 524-538.

[130] Mandin, P., and Gottesman, S. A genetic approach for finding small RNAs regulators of genes of interest identifies RybC as regulating the DpiA/DpiB two-component system, Mol Microbiol 72 2009, 551-565.

[131] Ulbrandt, N. D., Newitt, J. A., and Bernstein, H. D. The E. coli signal recognition particle is required for the insertion of a subset of inner membrane proteins, Cell 88 1997, 187-196.

[132] Papenfort, K., Pfeiffer, V., Lucchini, S., Sonawane, A., Hinton, J. C., and Vogel, J. Systematic deletion of Salmonella small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis, Mol Microbiol 68 2008, 890-906.

[133] Bouvier, M., Sharma, C. M., Mika, F., Nierhaus, K. H., and Vogel, J. Small RNA binding to 5' mRNA coding region inhibits translational initiation, Mol Cell 32 2008, 827-837.

[134] Guillier, M., and Gottesman, S. The 5' end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator, Nucleic Acids Res 36 2008, 6781-6794.

[135] Wadler, C. S., and Vanderpool, C. K. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide, Proc Natl Acad Sci U S A 104 2007, 20454-20459.

[136] Hershberg, R., Altuvia, S., and Margalit, H. A survey of small RNA-encoding genes in Escherichia coli, Nucleic Acids Res 31 2003, 1813-1820.

[137] Jin, Y., Watt, R. M., Danchin, A., and Huang, J. D. Small noncoding RNA GcvB is a novel regulator of acid resistance in Escherichia coli, BMC Genomics 10 2009, 165.

[138] Hobbs, E. C., Astarita, J. L., and Storz, G. 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 192 2010, 59-67.

[139] Urban, J. H., and Vogel, J. Translational control and target recognition by Escherichia coli small RNAs in vivo, Nucleic Acids Res 35 2007, 1018-1037.

[140] Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T., and Gottesman, S. 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 95 1998, 12462-12467.

[141] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res 31 2003, 3406-3415.

[142] Vogel, J., and Wagner, E. G. Target identification of small noncoding RNAs in bacteria, Curr Opin Microbiol 10 2007, 262-270.

[143] Panja, S., and Woodson, S. A. Hfq proximity and orientation controls RNA annealing, Nucleic Acids Res 40 2012, 8690-8697.

[144] Beisel, C. L., Updegrove, T. B., Janson, B. J., and Storz, G. Multiple factors dictate target selection by Hfq-binding small RNAs, EMBO J 31 2012, 1961-1974.

[145] del Val, C., Romero-Zaliz, R., Torres-Quesada, O., Peregrina, A., Toro, N., and Jimenez-Zurdo, J. I. A survey of sRNA families in alpha-proteobacteria, RNA Biol 9, 119-129.

[146] Gomez-Lozano, M., Marvig, R. L., Molin, S., and Long, K. S. Genome-wide identification of novel small RNAs in Pseudomonas aeruginosa, Environ Microbiol 14 2012, 2006-2016.

[147] Livny, J., Fogel, M. A., Davis, B. M., and Waldor, M. K. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes, Nucleic Acids Res 33 2005, 4096-4105.

[148] Rivas, E., and Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis, BMC Bioinformatics 2 2001, 8.

[149] Bradley, E. S., Bodi, K., Ismail, A. M., and Camilli, A. A genome-wide approach to discovery of small RNAs involved in regulation of virulence in Vibrio cholerae, PLoS Pathog 7 2011, e1002126.

[150] Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V., and Vogel, J. A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the Salmonella core genome, Mol Microbiol 66 2007, 1174-1191.

[151] Sharma, C. M., Darfeuille, F., Plantinga, T. H., and Vogel, J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites, Genes Dev 21 2007, 2804-2817.

[152] Peer, A., and Margalit, H. Accessibility and evolutionary conservation mark bacterial small-rna target-binding regions, J Bacteriol 193 2011, 1690-1701.

[153] Richter, A. S., and Backofen, R. Accessibility and conservation: general features of bacterial small RNA-mRNA interactions?, RNA Biol 9 2012, 954-965.

[154] Gaballa, A., Antelmann, H., Aguilar, C., Khakh, S. K., Song, K. B., Smaldone, G. T., and Helmann, J. D. The Bacillus subtilis iron-sparing response is mediated by a Fur-

regulated small RNA and three small, basic proteins, Proc Natl Acad Sci U S A 105 2008, 11927-11932.

[155] Cao, Y., Wu, J., Liu, Q., Zhao, Y., Ying, X., Cha, L., Wang, L., and Li, W. sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments, RNA 16 2010, 2051-2057.

[156] Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. WebLogo: a sequence logo generator, Genome Res 14 2004, 1188-1190.

[157] Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. Accurate SHAPE-directed RNA structure determination, Proc Natl Acad Sci U S A 106 2009, 97-102.

[158] Rudd, K. E. EcoGene: a genome sequence database for Escherichia coli K-12, Nucleic Acids Res 28 2000, 60-64.

[159] Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A. G., Mackie, A., Paulsen, I., Gunsalus, R. P., and Karp, P. D. EcoCyc: a comprehensive database of Escherichia coli biology, Nucleic Acids Res 39 2011, D583-590.

[160] Tjaden, B., Goodwin, S. S., Opdyke, J. A., Guillier, M., Fu, D. X., Gottesman, S., and Storz, G. Target prediction for small, noncoding RNAs in bacteria, Nucleic Acids Res 34 2006, 2791-2802.

[161] Moll, I., Afonyushkin, T., Vytvytska, O., Kaberdin, V. R., and Blasi, U. Coincident Hfq binding and RNase E cleavage sites on mRNA and small regulatory RNAs, RNA 9 2003, 1308-1314.

[162] Papenfort, K., Sun, Y., Miyakoshi, M., Vanderpool, C. K., and Vogel, J. Small RNA-mediated activation of sugar phosphatase mRNA regulates glucose homeostasis, Cell 153 2013, 426-437.

[163] Urban, J. H., and Vogel, J. Two seemingly homologous noncoding RNAs act hierarchically to activate glmS mRNA translation, PLoS Biol 6 2008, e64.

[164] Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., and Bryant, S. H. CDD: conserved domains and protein three-dimensional structure, Nucleic Acids Res 41 2013, D348-352.

[165] Lwoff, A. Interaction among virus, cell, and organism, Science 152 1966, 1216-1220.

[166] Cho, E. H., Gumport, R. I., and Gardner, J. F. Interactions between integrase and excisionase in the phage lambda excisive nucleoprotein complex, J Bacteriol 184 2002, 5200-5203.

[167] Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. Prophage genomics, Microbiol Mol Biol Rev 67 2003, 238-276, table of contents.

[168] Casjens, S. Prophages and bacterial genomics: what have we learned so far?, Mol Microbiol 49 2003, 277-300.

[169] Wang, X., Kim, Y., Ma, Q., Hong, S. H., Pokusaeva, K., Sturino, J. M., and Wood, T. K. Cryptic prophages help bacteria cope with adverse environments, Nat Commun 1 2010, 147.

[170] Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A., and Blattner, F. R. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7, Nature 409 2001, 529-533.

[171] Domka, J., Lee, J., Bansal, T., and Wood, T. K. Temporal gene-expression in Escherichia coli K-12 biofilms, Environ Microbiol 9 2007, 332-346.

[172] Wang, X., Kim, Y., and Wood, T. K. Control and benefits of CP4-57 prophage excision in Escherichia coli biofilms, ISME J 3 2009, 1164-1179.

[173] Tang, H., Rothery, R. A., Voss, J. E., and Weiner, J. H. Correct assembly of iron-sulfur cluster FS0 into Escherichia coli dimethyl sulfoxide reductase (DmsABC) is a prerequisite for molybdenum cofactor insertion, J Biol Chem 286 2011, 15147-15154.

[174] Richards, G. R., and Vanderpool, C. K. Molecular call and response: the physiology of bacterial small RNAs, Biochim Biophys Acta 1809 2011, 525-531.

[175] Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T., Mori, H., Perna, N. T., Plunkett, G., 3rd, Rudd, K. E., Serres, M. H., Thomas, G. H., Thomson, N. R., Wishart, D., and Wanner, B. L. Escherichia coli K-12: a cooperatively developed annotation snapshot--2005, Nucleic Acids Res 34 2006, 1-9.

[176] Schell, M. A. Molecular biology of the LysR family of transcriptional regulators, Annu Rev Microbiol 47 1993, 597-626.

[177] Kovacikova, G., and Skorupski, K. A Vibrio cholerae LysR homolog, AphB, cooperates with AphA at the tcpPH promoter to activate expression of the ToxR virulence cascade, J Bacteriol 181 1999, 4250-4256.

[178] Cao, H., Krishnan, G., Goumnerov, B., Tsongalis, J., Tompkins, R., and Rahme, L. G. A quorum sensing-associated virulence gene of Pseudomonas aeruginosa encodes a LysR-like transcription regulator with a unique self-regulatory mechanism, Proc Natl Acad Sci U S A 98 2001, 14613-14618.

[179] Sperandio, B., Gautier, C., McGovern, S., Ehrlich, D. S., Renault, P., Martin-Verstraete, I., and Guedon, E. Control of methionine synthesis and uptake by MetR and homocysteine in Streptococcus mutans, J Bacteriol 189 2007, 7032-7044.

[180] Maddocks, S. E., and Oyston, P. C. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins, Microbiology 154 2008, 3609-3623.

[181] Moon, K., Six, D. A., Lee, H. J., Raetz, C. R., and Gottesman, S. Complex transcriptional and post-transcriptional regulation of an enzyme for lipopolysaccharide modification, Mol Microbiol 89 2013, 52-64.

[182] Trent, M. S. Biosynthesis, transport, and modification of lipid A, Biochem Cell Biol 82 2004, 71-86.

[183] Zhou, J., and Rudd, K. E. EcoGene 3.0, Nucleic Acids Res 41 2013, D613-624.

[184] Tenorio, E., Saeki, T., Fujita, K., Kitakawa, M., Baba, T., Mori, H., and Isono, K. Systematic characterization of Escherichia coli genes/ORFs affecting biofilm formation, FEMS Microbiol Lett 225 2003, 107-114.

[185] Murata, M., Fujimoto, H., Nishimura, K., Charoensuk, K., Nagamitsu, H., Raina, S., Kosaka, T., Oshima, T., Ogasawara, N., and Yamada, M. Molecular strategy for survival at a critical high temperature in Eschierichia coli, PLoS One 6 2011, e20063.

[186] Lamichhane, T. N., Abeydeera, N. D., Duc, A. C., Cunningham, P. R., and Chow, C. S. Selection of peptides targeting helix 31 of bacterial 16S ribosomal RNA by screening M13 phage-display libraries, Molecules 16 2011, 1211-1239.

[187] Scofield, M. A., Lewis, W. S., and Schuster, S. M. Nucleotide sequence of Escherichia coli asnB and deduced amino acid sequence of asparagine synthetase B, J Biol Chem 265 1990, 12895-12902.

[188] Felton, J., Michaelis, S., and Wright, A. Mutations in two unlinked genes are required to produce asparagine auxotrophy in Escherichia coli, J Bacteriol 142 1980, 221-228.

[189] Larsen, T. M., Boehlein, S. K., Schuster, S. M., Richards, N. G., Thoden, J. B., Holden, H. M., and Rayment, I. Three-dimensional structure of Escherichia coli asparagine synthetase B: a short journey from substrate to product, Biochemistry 38 1999, 16146-16157.

[190] Hommais, F., Krin, E., Coppee, J. Y., Lacroix, C., Yeramian, E., Danchin, A., and Bertin, P. GadE (YhiE): a novel activator involved in the response to acid environment in Escherichia coli, Microbiology 150 2004, 61-72.

[191] Qian, G., Liu, C., Wu, G., Yin, F., Zhao, Y., Zhou, Y., Zhang, Y., Song, Z., Fan, J., Hu, B., and Liu, F. AsnB, regulated by diffusible signal factor and global regulator Clp, is involved in aspartate metabolism, resistance to oxidative stress and virulence in Xanthomonas oryzae pv. oryzicola, Mol Plant Pathol 14 2012, 145-157.

[192] Pulvermacher, S. C., Stauffer, L. T., and Stauffer, G. V. Role of the Escherichia coli Hfq protein in GcvB regulation of oppA and dppA mRNAs, Microbiology 155 2009, 115-123.

[193] Pulvermacher, S. C., Stauffer, L. T., and Stauffer, G. V. Role of the sRNA GcvB in regulation of cycA in Escherichia coli, Microbiology 155 2009, 106-114.

[194] Pulvermacher, S. C., Stauffer, L. T., and Stauffer, G. V. The small RNA GcvB regulates sstT mRNA expression in Escherichia coli, J Bacteriol 191 2009, 238-248.

[195] Coornaert, A., Chiaruttini, C., Springer, M., and Guillier, M. Post-transcriptional control of the Escherichia coli PhoQ-PhoP two-component system by multiple sRNAs involves a novel pairing region of GcvB, PLoS Genet 9 2013, e1003156.

[196] Min, B., Pelaschier, J. T., Graham, D. E., Tumbula-Hansen, D., and Soll, D. Transfer RNA-dependent amino acid biosynthesis: an essential route to asparagine formation, Proc Natl Acad Sci U S A 99 2002, 2678-2683.

[197] Horler, R. S., and Vanderpool, C. K. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence, Nucleic Acids Res 37 2009, 5465-5476.

[198] Huttenhofer, A., and Noller, H. F. Footprinting mRNA-ribosome complexes with chemical probes, EMBO J 13 1994, 3892-3901.

[199] Lu, X., Goodrich-Blair, H., and Tjaden, B. Assessing computational tools for the discovery of small RNA genes in bacteria, RNA 17 2011, 1635-1647.

[200] Boughammoura, A., Matzanke, B. F., Bottger, L., Reverchon, S., Lesuisse, E., Expert, D., and Franza, T. Differential role of ferritins in iron metabolism and virulence of

the plant-pathogenic bacterium Erwinia chrysanthemi 3937, J Bacteriol 190 2008, 1518-1530.

[201] Davis, B. M., Quinones, M., Pratt, J., Ding, Y., and Waldor, M. K. Characterization of the small untranslated RNA RyhB and its regulon in Vibrio cholerae, J Bacteriol 187 2005, 4005-4014.

[202] Ellermeier, J. R., and Slauch, J. M. Fur regulates expression of the Salmonella pathogenicity island 1 type III secretion system through HilD, J Bacteriol 190 2008, 476-486.

[203] Murphy, E. R., and Payne, S. M. RyhB, an iron-responsive small RNA molecule, regulates Shigella dysenteriae virulence, Infect Immun 75 2007, 3470-3477.

[204] Wilderman, P. J., Sowa, N. A., FitzGerald, D. J., FitzGerald, P. C., Gottesman, S., Ochsner, U. A., and Vasil, M. L. Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis, Proc Natl Acad Sci U S A 101 2004, 9792-9797.

[205] McCown, P. J., Roth, A., and Breaker, R. R. An expanded collection and refined consensus model of glmS ribozymes, RNA 17 2011, 728-736.

[206] Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench, Bioinformatics 25 2009, 1189-1191.

[207] Gopel, Y., Papenfort, K., Reichenbach, B., Vogel, J., and Gorke, B. Targeted decay of a regulatory small RNA by an adaptor protein for RNase E and counteraction by an anti-adaptor RNA, Genes Dev 27 2013, 552-564.

[208] Olsen, A. S., Moller-Jensen, J., Brennan, R. G., and Valentin-Hansen, P. C-terminally truncated derivatives of Escherichia coli Hfq are proficient in riboregulation, J Mol Biol 404 2010, 173-182.

[209] Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J. K., and Breaker, R. R. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control, Proc Natl Acad Sci U S A 101 2004, 6421-6426.

[210] Dambach, M., Irnov, I., and Winkler, W. C. Association of RNAs with Bacillus subtilis Hfq, PLoS One 8 2013, e55156.

[211] Mellin, J. R., and Cossart, P. The non-coding RNA world of the bacterial pathogen Listeria monocytogenes, RNA Biol 9 2012, 372-378.

ABSTRACT

INVESTIGATING HFQ-mRNA INTERACTIONS IN BACTERIA

by

MARTHA A. FANER

August 2013

Advisor: **Prof. Andrew L. Feig**

Major: **Chemistry (Biochemistry)**

Degree: **Doctor of Philosophy**

Regulatory RNAs (sRNAs) are essential for bacteria to thrive in diverse environments and they also play a key role in virulence [11]. *Trans*-sRNAs affect the stability and/or translation of their target mRNAs through complementary base-pairing. The base-pairing interaction is not perfect and requires the action of an RNA binding protein, Hfq. Hfq facilitates these RNA-RNA interactions by stabilizing duplex formation, aiding in structural rearrangements, increasing the rate of structural opening, and/or by increasing the rate of annealing [18-21]. Hfq has two well characterized binding surfaces: the proximal surface, which binds AU rich stretches typical of sRNAs, and the distal surface, which binds $(ARN)_x$ motifs typically found in target mRNAs [30, 33, 36]. Studies on Hfq-RNA interactions have focused largely on sRNAs until the more recent discovery of an $(ARN)_x$ motif within the 5'UTR of target mRNAs[36, 37]. The importance of this motif in facilitating Hfq-mRNA binding and its requirement for regulation of a couple well known target mRNAs led us to further characterize the motif in the work described in this thesis. We performed bioinformatic and *in vitro* analyses to investigate the prevalence, location, structural contexts, and Hfq-binding of $(ARN)_x$ motifs in known

target mRNAs. We found that the known targets contain single stranded $(ARN)_x$ sequences in their 5'UTRs that bind to Hfq. Two predominant structural contexts of the single stranded $(ARN)_x$ motifs became clear: they were either flanked by stem loop structures or within a loop of an internal bulge, multi-branch junction or hairpin. The key features of the motifs were then used as a bioinformatic tool on a genome wide scale to identify mRNAs that might bind to Hfq. We found that 21% of mRNAs have a suitable $(ARN)_x$ motif and therefore likely bind to Hfq. Messages that bind to Hfq may be novel sRNA targets so we investigated this possibility using an *in vivo* reporter assay and found that 63% of the mRNAs tested are regulated by a specific sRNA. The novel targets are involved in pathways including iron salvage, biofilm formation, and amino acid metabolism. Overall, we defined key features of $(ARN)_x$ motifs and were able to use those to predict novel target mRNAs in *E. coli*. This approach is efficient, effective and adaptable other bacterial species.

AUTOBIOGRAPHICAL STATEMENT

MARTHA A. FANER

EDUCATION:

2007 - 2013
Ph.D. in Biological Chemistry
Wayne State University Detroit, MI

2003 - 2007
Bachelor of Science (B.S.) Biological Chemistry
Northern Michigan University Marquette, MI

PUBLICATIONS:

- Salim, N N., Faner M, and Feig, A. L. (2012) Mechanisms of Hfq Mediated Gene Regulation of *glmS* by sRNAs GlmZ and GlmY, Nucleic Acids Research.
- Faner, M., Feig, A. L. (2013) Characterization of Hfq-RNA Interactions, Methods.

PRESENTATIONS:

Oral Presentations

- Faner, M., Swett, R., Kumar, A., Joiner, C., and Feig, A.L. "Presence of Hfq Binding Site Facilitates Identification of Functionally Important mRNA Targets" RNA Society Meeting, University of Michigan, 2012
- Faner, M. "PINK1 and Parkin: A dynamic duo in Parkinson's Disease" Departmental Literature Seminar, Wayne State University, 2011
- Faner, M. and Feig, A.L. "Investigating the role of mRNAs in Hfq-sRNA mediated gene regulation in *E. coli*" RNA club, Wayne State University, 2010
- Faner, M. and Feig, A.L. "Investigating structure and function of novel non-coding RNAs" RNA club, Wayne State University, 2009
- Faner, M. "The new functional non-coding RNA: CRISPR" Departmental Literature Seminar, Wayne State University, 2009

Poster Presentations

- Faner, M., Swett, R., Joiner, C., and Feig, A.L. "Predicting Targets of Riboregulation by Identifying New Hfq Binding Sites Within the 5'UTRs of mRNAs" University of Michigan Postdoc Poster Session, Ann Arbor, MI, 2011
- Faner, M.*,Salim, N., Swett, R., and Feig, A.L. "Investigating a Putative Hfq Binding Site in the 5'UTR of mRNAs in *E. coli"* The Michigan RNA Society Meeting, Albion College, Albion, MI, 2010
- Faner, M., Feig, A.L. "Discovery of novel sRNAs in *Clostridium Difficile*" The 10th Annual Graduate Symposium, Wayne State University, 2008

AWARDS:

Departmental Citation for Excellence in Teaching Service
Department of Chemistry, Wayne State University, 2012

ACS Women's Committee Travel Award Sponsored by Eli Lilly, 2012