ABSTRACT

SNYDER, ROBERT WESLEY. Comparative RNA Binding Thermodynamics and Information Content from Sequencing Data. (Under the direction of Donald Bitzer and Fred Breidt).

Producing proteins is an important industrial process. The best current method for making biopharmaceuticals is to modify a vector organism to produce the protein transgenically and scale up the process *en mass* in bioreactors. This usually requires modifying the DNA code for the protein such that the vector organism will accurately and efficiently translate the messenger RNA. To this end, we must understand the fundamentals of ribosome-mRNA interactions. Predicting the binding energy of sequence-dependent RNA duplex formations is crucial for understanding how the ribosome decodes the genetic information. Several models for predicting this phenomenon exist, however they do not always agree. Prior methods to quantify the process, such as melting curve experiments, compare only a small fraction of the possible interactions in order to infer the thermodynamic regulations. The scientific community needs a method that can compare binding models to determine which one is best so that proteins can be properly and efficiently synthesized. It was hypothesized that an experiment could be designed that could comprehensively compare the binding energies of RNA and rRNA to more specifically elucidate the relative energies of RNA binding for the case of protein translation. A more comprehensive, ribosome-specific method was developed, and seven accepted algorithms from the literature were used to predict the behavior of the ribosome and the relative accuracy of the models. Finally, a novel implementation of the biological information content algorithm was applied to the case of

variable background nucleotide distribution. These methods can be used to generate more accurate models of RNA binding and thus protein translation, which in turn can allow us to more efficiently and accurately produce proteins for medicine and agriculture.

Comparative RNA Binding Thermodynamics and Information Content from Sequencing Data

by
Robert Wesley Snyder

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Biomedical Engineering

Raleigh, North Carolina

2017

APPROVED BY:

Dr. Donald L. Bitzer Co-Chair of the Advisory Committee	Dr. Fred Breidt, Jr. Co-Chair of the Advisory Committee
Dr. Shawn Gomez	Dr. Albert Banes
Dr. Paul Wollenzien	

DEDICATION

Dedicated to Srishti who has kept me sane and loved, and to our unborn child which we are temporarily naming Watermelon

BIOGRAPHY

Robert Wesley Snyder hails from the city of Raleigh, North Carolina, and is the son of Roz and Wes Snyder. A south wind blew on that auspicious night, portending the facetious nature of the dissertation biography to follow… Ok enough silliness.

My first engagement with science that I can recall happened when I was 5 years old. I was quite enamored with birds, and particularly with their ability to fly, which I envied. My older brother Graham came home from school and told me "the reason birds grow wings and you grow arms is because of genetics". My imagination reeled with the possibility of grasping this magically instruction book called genetics and perhaps one day using it so that I could grow wings and fly like a bird. However, the childish desire to be the modern-day Icarus soon faded and was replaced by a new, and a stern, motivation toward science. I was 12 and my grandmother's slow decent into dementia grew ever more apparent and debilitating. This phenomenon was difficult to watch, and quite frankly, scary. It is a travesty, and yet an accepted part of life. Much of medicine focuses on the body, yet should the mind go, what was the point? Witnessing this first hand cemented my path toward science, and after a couple of excellent biology classes at Enloe High School, I knew I would be a scientist, and I knew I would do research. I knew that it would be genetics, and I knew that it would be to help people.

My current position as a PhD level scientist at a cancer research startup company could ostensibly be tracked back to one piece of advice I received: prepare for an interview. One piece of preparation got me my first lab tech job, which got me an engineering Co-op position, which got me a research assistantship, which got me a publication, which got me

into graduate school, which got me the career I have today. The piece of preparation that got me to where I am? I looked up the fact that mosquito egg shells can harden and make injecting viruses difficult. I didn't even get that job, but the lab manager was impressed that I'd taken the time to look up some cursory piece of esoterica and recommended me for my next position in a virology laboratory at the University of Maryland Biotechnology institute. From there I worked at The Institute for Genomic Research (TIGR) in the Venter Institute sequencing facilities, and then back to college and a research assistant position in chemical engineering, and finally graduate school. I was amazed that the random factoid I looked up before that interview actually got me a job, and though I may still have been successful if I hadn't looked it up, I never failed to do a lot of research before an interview.

Science has not always been easy, and finishing a PhD was particularly grueling. It was not the difficulty of the work or the frustration of failure that made it so, but rather the very isolating nature of the dissertation research process. For most of us, it is the biggest project we ever work on by ourselves. The lack of comradery with fellow students made successes more grey, and small hiccups in the research process became larger obstacles than they needed to be. I was bolstered by my committee, who were incredibly supportive. I always felt motivated and ready to grind on after meeting with my advisors. I was lucky to have their support, and the support of my family and loved ones. I could not have done it had I not had all of them in my life. Graduate school is now over and my career just beginning, and though I share the spirit of that little boy that wanted to fly, I am now a man with the nobler and perhaps less feasible goal of saving the world.

ACKNOWLEDGMENTS

So many thanks to my parents Roz and Wes without whom this would not have been possible. Of course, my advisors and my committee were extremely supportive and engaged, and I doubt my gratitude can be adequately conveyed in words. Acknowledgments are also due to Dr. Tatjana Shapkina who was instrumental in teaching me many biochemistry techniques in the laboratory, Dr. Jason Osborne from the statistics department who helped me with ΔG° normalization and comparisons, Dr. Jason Haugh from chemical and biomolecular engineering who helped with competitive reaction equilibria, and Dr. Tanya Kalich researcher at SAS Institute for great conversations which led to the tool to code adaptive regressions with. Also, I'd like to acknowledge Wade Colburn and Arsani Balamoun who helped me in the lab. There is always, naturally, the love of my life Dr. Srishti Lipsa Bhagat, M.D., Ph.D., that I must acknowledge played no small role in helping me get through this.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Background and Motivation

The complexity of the natural world is without compare to anything manufactured by humankind. Nature's evolution of the nearly infinite possible conformations of life makes full quantification of the vast and intricate processes intractable. There remain, however, the fundamental laws to which all matter and energy that we believe must adhere: the laws of thermodynamics. Therefore, given knowledge of an initial state, such as the atoms that make up DNA a sequence, a ribosome, and a desired protein product, it should be possible to not only predict the outcome, but to create new outcomes *de novo* so long as we understand the laws governing gene translation from fundamental principles.

Biological systems create order from a chaotic system, such as chemical, electromagnetic, or thermal energy. The organization of an ordered state from a chaotic one is defined as information. In organisms, this information is stored in genetic sequences. In the specific case of protein translation, a linear strand of information, the messenger RNA (mRNA), is "read" by a decoding device called the ribosome. Since the information is read linearly, it is reasonable to use signal processing analysis to understand how the mRNA encodes information, how the ribosome decodes it, and the final result of the process, the protein.

A method of inferring the magnitude of the signal is based on the energy arising from the biochemical interaction of RNA-RNA binding. The current algorithms for predicting these values are frequently based on empirical data from single oligonucleotide-

oligonucleotide binding events quantified by melting curve experiments. More recent advances in technology have allowed the analysis of simultaneously binding oligonucleotides by the hundreds or thousands using techniques such as microarrays [5] and microfluidic platforms [13], and some protein-DNA interactions have been quantified using a combination of sequencing with fluorescently labelled ligands [32]. We are not aware of any studies simultaneously measuring equilibrium binding events of millions of RNA-RNA sequences.

The first observation that motivated the following dissertation is that the models do not always agree with one another (Figure 1). Not only are the magnitudes of their predictions different, but the order in which one sequence may bind stronger than another also differ.

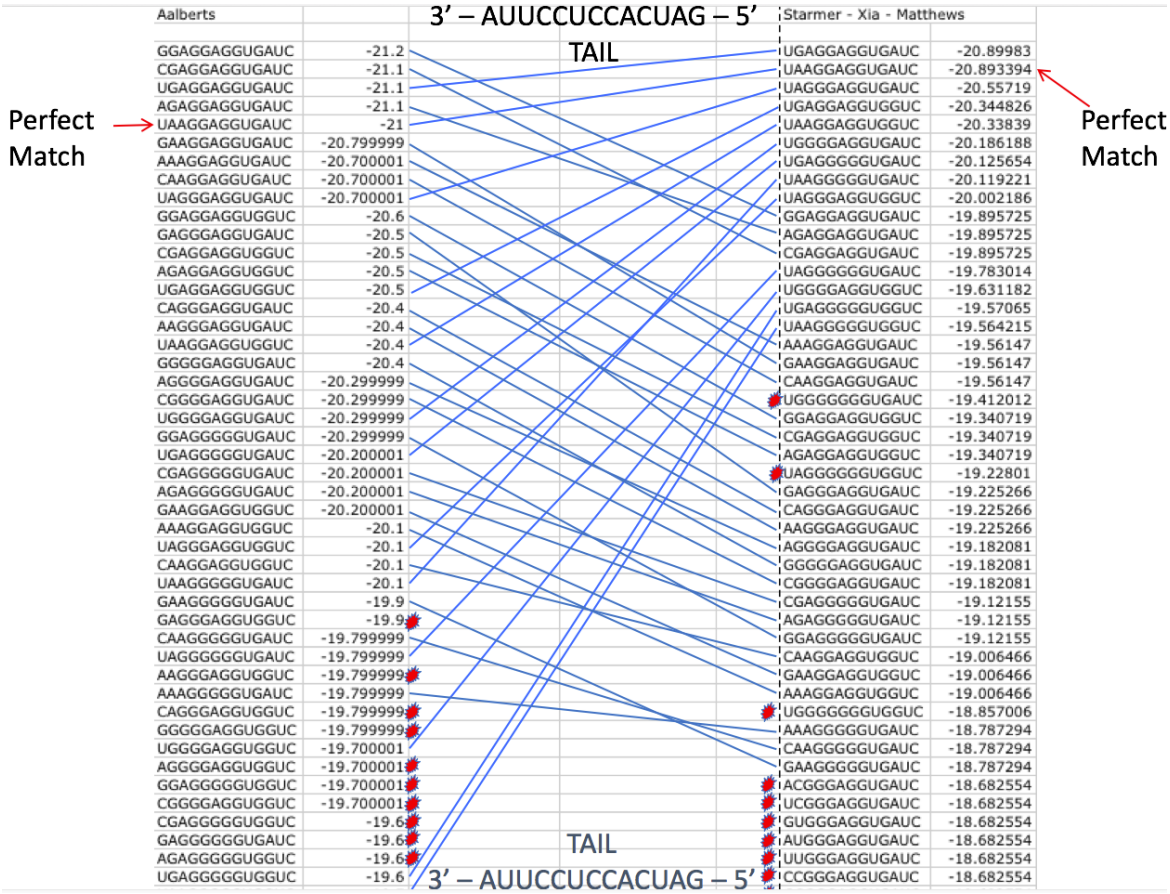

Figure 1. An ordered list of 13-mer oligonucleotides sorted by their predicted minimum free energy of binding with the 3' tail end of the 16S rRNA. The Aalberts' BINDIGO model is shown on the left and the free_scan Starmer model with the Xia/Mathews parameter set is on the right. The sequences that match each other are connected by lines. Sequences that are in the top list of one of the models, but not the other are denoted by a star.

It was desired to accurately determine which computer models best predict RNA binding energies so that we may be confident of the signal that we analyze when studying ribosome-mRNA interactions. It was hypothesized that a better method to compare the accuracy of RNA binding algorithms could be developed by analyzing binding ratios of randomized oligonucleotides reacting in competition with the RNA of the ribosome. To this

end a series of experiments that can simultaneously measure millions upon millions of binding events was designed, both in the context of the ribosomal RNA and more isolated and controlled RNA-RNA binding events. This required the utilization of new sequencing technologies and the development and application of a more general biological information content theory.

It was found that for sequences that are highly complementary with the 3' tail end of the 16S rRNA, the models analyzed were in general agreement, however the Unafold [25] model best predicted the data gathered for both experiments while the free_scan model with the combined Freier-Mathews parameter set [49] best predicted the rank order for *in vitro* ribosome experiments, whereas the ViennaRNA package [23] with the 2007 parameters best predicted the rank order for the nanobead-bound oligomer binding experiment.

1.1 Basics of Genetics

This section describes the physical structures and the processes they are involved in when prokaryotes decode DNA into RNA and subsequently RNA into protein.

1.1.1 Gene Structure

Deoxyribonucleic acid is a polymeric molecule made of four nucleic acids; deoxyadenosine, deoxyguanosine, deoxycytodine, and deoxythymidine, connected through phosphate diester linkages. The order, or sequence, of the nucleic acids, also called nucleotides or bases, is what encodes the information an organism needs to operate.

A *gene* consists of five primary parts, a transcription initiation site, called an operon, a ribosome binding site, a translation initiation site, the start codon, the coding region, and the stop codon.

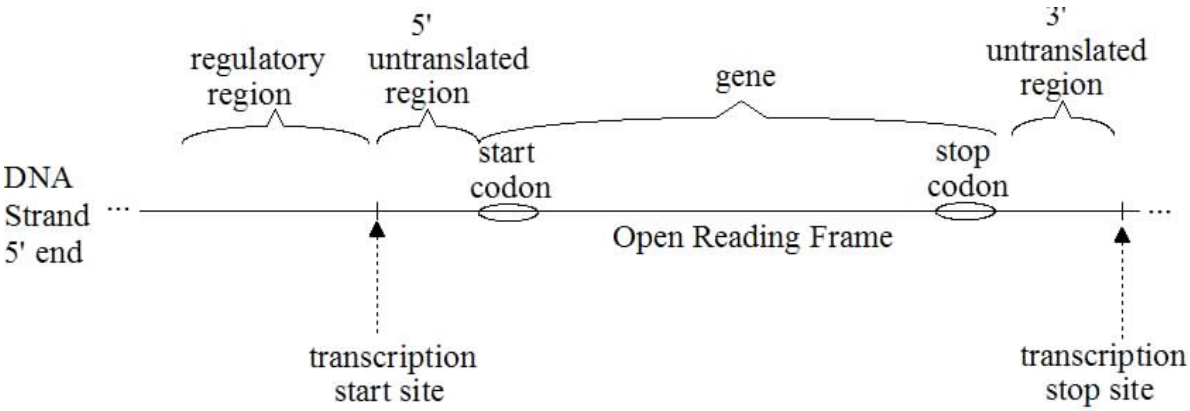

Figure 2. The structure of a prokaryotic gene. Image from Professor Mark Fienup, Dept. Computer Science, University of Northern Iowa. http://www.cs.uni.edu/~fienup/cs188s05/lectures/lec23_4-12-05.html

1.1.2 Gene Expression

The first step major step is called *transcription*, the process by which the organism copies a section of the DNA off the genome or plasmid into its complementary RNA. During transcription, an enzyme called RNA polymerase binds to a transcription initiation site and *transcribes* the DNA one base at a time, where a new RNA strand, called the messenger RNA, mRNA, or transcript, is created that is complementary to the DNA. The RNA complement to DNA is C to G, G to C, A to U, and U to A (there is no T in RNA, it is replaced by U).

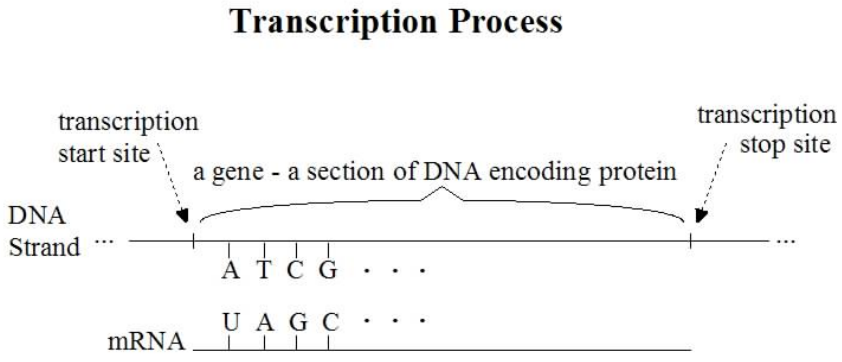

Figure 3. Diagram of the process of transcription.
http://www.cs.uni.edu/~fienup/cs188s05/lectures/4265182b.jpg

A *codon* is a sequence of three nucleotides. Since there are four possible nucleotides (C, U, A, G), there are $4^3 = 64$ possible codons. Of the 64 possible codons, only 61 are actually used to code information, and three are used as stop signals (Figure 2). That is, they indicate the end of translation and the point at which the ribosome will cease translating and fall off the mRNA.

The process after transcription is called translation, where the mRNA is read and proteins are produced according to the code on the mRNA. The mRNA has a region where a ribonucleoprotein called a ribosome can bind, called the ribosome binding site. This region is upstream (on the 5' side) of the start codon. The ribosome binds and travels down toward the 3' end, reading the codons of the mRNA as it goes. When the ribosome encounters the start codon, it initiates the elongation process by recruiting transfer RNAs (tRNAs) which bring the amino acids that correspond to each codon in succession. Each codon is read, and the corresponding amino acid added to the chain, until a stop codon is reached. At this point translation is terminated and the amino acid chain is released as a protein into the cell's cytoplasm.

1.2 The Bitzer Model

In 1974 Shine and Dalgarno noticed that a certain DNA sequence was frequently found upstream of the start codon of protein coding genes in bacteria [46]. This *Shine-Dalgarno sequence* is complementary to a short single-stranded RNA sequence found on the tail of the 16S ribosomal subunit. It was found that this region is responsible for the efficiency of protein translation initiation by recruitment between the ribosome and the messenger RNA [47]. The recruitment is achieved through standard Watson-Crick hybridization based on sequence complementarities. Further analysis in the Bitzer lab [34] extended the mRNA-16S tail interaction by calculating the theoretical Gibbs free energy of binding, $\Delta G°$, between the tail and all subsequent sequence downstream in "windows" of 13 bases (Figure 3). The results represent a minimum free energy (MFE) profile of ribosome-message interaction throughout the translation of the protein.

```
Position 0.      Free energy value  = 0.0
rRNA:            a u u c c u c c a c u a g
mRNA:            G G U A A A A G A A U A A U G G C…
Position 1.      Free energy value = 0.0
rRNA:            a u u c c u c c a c u a g
mRNA:            …G G U A A A A G A A U A A U G G C…
⋮
Position 63.     Free energy value  = −1.7
rRNA:            a u u c c u c c a c u a g
mRNA:            …U C A C C G A G A U C C U G G U C…
⋮
Position N-2.    Free energy value  = 0.0
rRNA:            a u u c c u c c a c u a g
mRNA:            …G C C G U C U G G U G A U G U A A
Position N-1.    Free energy value  = −0.7
rRNA:            a u u c c u c c a c u a g
mRNA:            …G C C G U C U G G U G A U G U A A
```

Figure 4. Alignment of the 16S rRNA tail with the mRNA sequence of gene aceF in *E. coli* using the free_scan model for minimum free energy (MFE) calculations with the Freier settings [37].

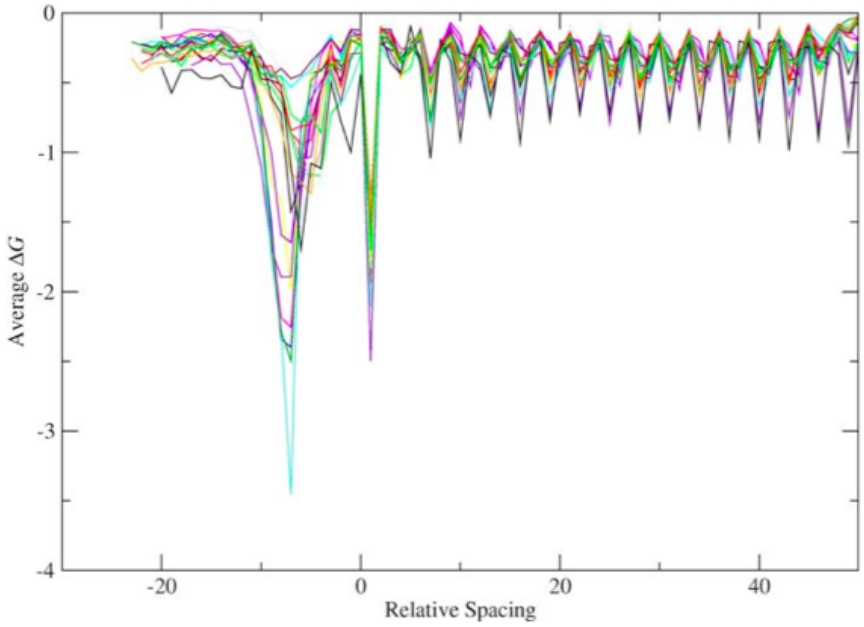

Figure 5. Ensemble average of free-energy of binding between the 16S tail and the messenger RNA of protein-coding genes in 18 different prokaryotes.

Signal processing of the energy profile nucleotide by nucleotide revealed a period-3 sinusoidal pattern of energy variation [37], which is not present in non-coding regions (Figure 4). This signal has species specific parameters, phase and displacement, that are vital to not only protein translational efficiency [51], but also to identifying or inducing programmed frameshifts [35], identifying introns and exons in eukaryotes [60], and even finding mistakes in genbank [49] such as 384 genes with mis-annotated start codons.

1.2.1 Ribosome structure

The prokaryotic ribosome is a ribonucleoprotein made up of two units of density 50S and 30S (Svedberg units refer to the rate of sedimentation during centrifugation) which combine to form the entire 70S ribosome, which is about 25 nm in diameter. The ribosome has three types of ribosomal RNA (rRNA), the 16S, 23S, and 5S. The 30S subunit contains the 16S RNA which has the single stranded 3' tail end that binds to the Shine-Dalgarno sequence during translation initiation (Figure 5), and it is the region we are interested in for this dissertation.

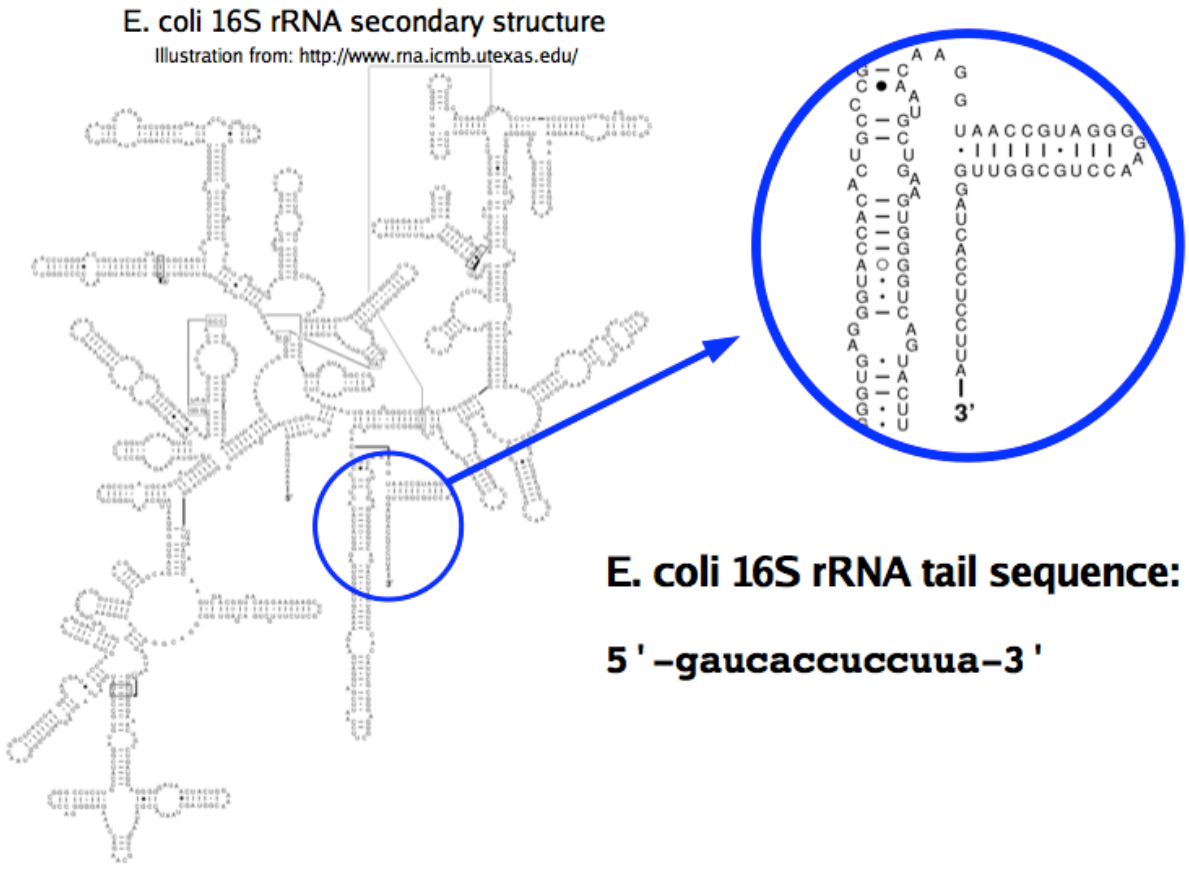

Figure 6. Secondary structure of 16S rRNA [15]. Note the single stranded 3' tail end emphasized above.

1.3 Free Energy Calculations

Ever since Watson and Crick published the structure of the double helix [56], the exact mechanisms of nucleotide base pairing have been of increasing interest due to the growing body of evidence supporting the importance of RNA folding and structures in multifarious biological systems. During a thermodynamically favorable chemical reaction a certain amount of energy known as the Gibbs free energy, ΔG° [62], is released. The new double helical structure is stable when the ΔG° is less than zero. Many researchers have

studied the free energy of these reactions in order to predict maximum likelihood structures *in silico*.

1.3.1 Free Energy: Review of RNA Binding Algorithms

The following section provides an overview of the current methods of predicting the structure and energetics of RNA folding. Models include the Nearest Neighbor Model, the HNNM, the algorithms based on those such as mfold [62], the RNA Vienna model [23], Freier [11], and Aalberts' BINDIGO algorithms [19].

RNA binding models may include secondary structure predictions when calculating $\Delta G°$. These models calculate possible binding conformations using previously known structures that have shorter sequences in common with the sequences being modelled [9]. The qualities of these predictions can be assessed by assigning a probability to it using a partition function [28]. A partition function is defined as the sum of all the equilibrium constants for the different secondary structures predicted for the sequences being modelled. The frequency of certain base pairs in all the different possible structures is calculated as base pairing probabilities, then structures with the highest combination of highly probable pairs are more likely to be correctly predicted ones.

For the specific situation of ribosome interaction, it is believed that certain secondary structure formation of the mRNA would inhibit translocation through the A site and P site, so the primary contributing factor to in-frame maintenance is the adjacent codon-anticodon interactions [46], and structures such as stem loops and hairpins are not present during this process (Figure 6).

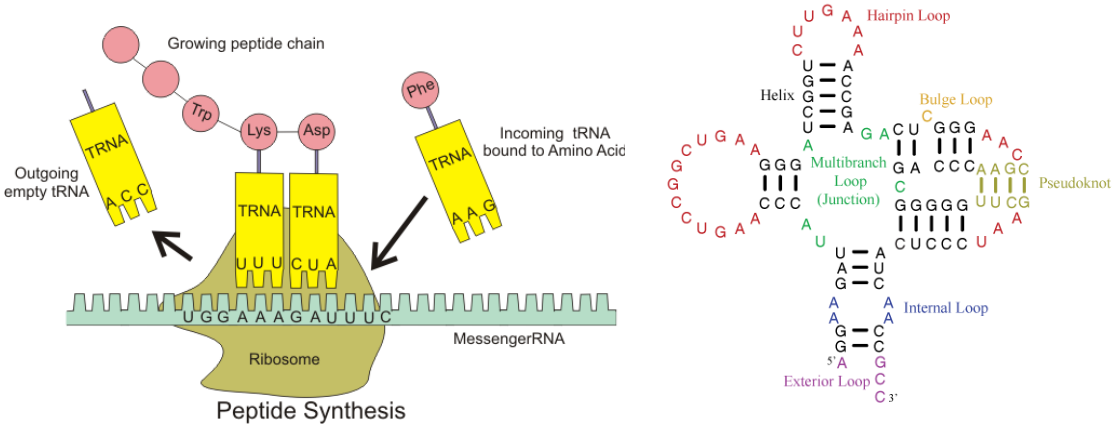

Figure 7. An artist's rendition of the process of translation, and RNA secondary structures. (From Lumelearning.com "Boundless Biology" chapter "Ribosomes and Protein Synthesis" and the website of the Turner laboratory, https://courses.lumenlearning.com/boundless-biology/chapter/ribosomes-and-protein-synthesis/, http://rna.urmc.rochester.edu/NNDB/help.html)

Predicting the relative binding strengths of 13-mer RNA hybrids is a fundamental calculation underlying the Bitzer model. Since there are discrepancies between the various models accepted in the literature, we must choose a model whose experimental data show it to be most accurate. Initial analysis compares the predictions of two models, Freier model and the Aalberts model. These are discussed below.

1.3.2 The Freier model – mfold

One long-standing implementation RNA binding algorithms, mfold, is a modified version of the original Turner rules for RNA-RNA complementary base-pairing [27]. mfold is frequently referred to as Unafold now. The rules make use of the individual nearest neighbor model (INN), where contiguous complementary base pairing contributes a greater

free energy than the sum of the individual base pairs, where the standard-state free energy of helix formation for an oligonucleotide is the sum of three terms: "(i) a free-energy change for helix initiation associated with forming the first base pair in the duplex, (ii) a sum of propagation free energies for forming each subsequent base pair, and (iii) a symmetry correction if the sequence is self-complementary" [11]. The model parameters were found by fitting absorbance vs. temperature melting curves for oligos of 8 bp or less. For most sequences it has fairly accurate estimates, however there were some oligos, particularly ones that are less stable, which fit the regression poorly and are not taken into account. The more recent INN hydrogen bonding (INN-HB) model improves mfold by penalizing terminal AU base pairs by 0.45 kcal/mol relative to terminal GC base pairs. One issue with the mfold software related to our problem is that it makes specific predictions about the penalties for free energy from the formation of certain secondary structures such as loops and hairpins. For ribosome binding to mRNA there are inherent space constraints due to the volume of the ribosome, and the secondary structure predictions are not believed to play a role. Another questionable parameter is the initial cost of helix formation that may not be significant during translation.

1.3.3 Starmer-modified version of mfold – free_scan

Dr. Josh Starmer wrote the free_scan program by modifying the Xia-Mathews or Freier parameters for the mfold algorithm to rule out loops and bulges, and eliminated the penalties for loops, bulges, and terminal AU base pairs [49].

1.3.4 Aalberts model – BINDIGO

The Aalberts model follows the mfold predictions except for the manner in which structural restrictions are enforced. Specifically, only inter-strand base pairs can be made. Additionally, the Aalberts model is asymptotically faster than the mfold algorithms, as it scales proportionally to the oligo lengths. By performing a sequence alignment analogous to the Smith-Waterman alignment first [64], and then using the experimentally measured free energies of the various structures, BINDIGO avoids the recursive folding of sub-sequences to find the optimal secondary structures. The method calculates the free energy at each alignment (Figure 7) and chooses the minimum as the answer to report.

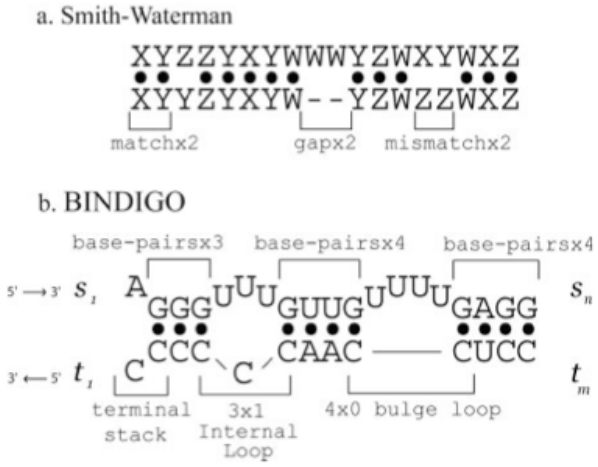

Figure 8. (a) The Smith-Waterman algorithm uses dynamic programming to align two sequences. The result is an optimal combination of matches, mismatches and gaps caused by insertions/deletions. (b) BINDIGO breaks pairings into secondary structures and scores each structural unit with measured free energies. The strings *s* and *t* are indexed left to right, but the program input is done in the 5' to 3' order.

1.3.5 RNA Vienna

ViennaRNA package is another widely implemented tool for RNA folding and binding predictions. Also based on the Turner 2004 rules [26], it is constrained by a set of rules for representing the edges between base pairs:

"1. base pair edges are formed only between nucleotides that form Watson-Crick or GU base pairs;

2. no two base pair edges emanate from the same vertex, i.e., a secondary structure is a matching;

3. base pair edges span at least three unpaired bases;

4. if the vertices are placed in 5' to 3' order on the circumference of a circle and edges are drawn as straight lines, no two edges cross [23]".

If the last condition is met, the algorithm excludes pseudo-knots. Since the oligonucleotides we are testing are quite short, pseudo-knots will not be a factor when modeling *in silico*. Base pairing probabilities are calculated using either an implementation of the Sankoff algorithm as part of the locarna package or using modified string alignment algorithms with the RNApaln and RNApdist packages [55].

1.3.6 Magnetic Streptavidin Nano-Beads and Biotinylated RNA

In the experiments shown in this dissertation, it was desired to have some control experiments that eliminated the possibility that other structures in the ribosome, such as

proteins or the other rRNA, may be affecting the results of the binding of RNA to the 3' tail sequence. To do this an experiment was designed that bound a 3' tail sequence to a magnetic streptavidin molecule such that no other proteins or RNA were in solution except the streptavidin and the RNA being tested. Streptavidin is a 159 residue homotetrameric protein (Figure 8) isolated from the organism *Streptomyces avidinii*. The extremely high affinity to biotin (Ka = 10^{15} M^{-1}) and ease of biotin coupling to other molecules makes streptavidin a useful substrate for tethered ligand experiments including RNA isolation and antibody capture [7, 13].

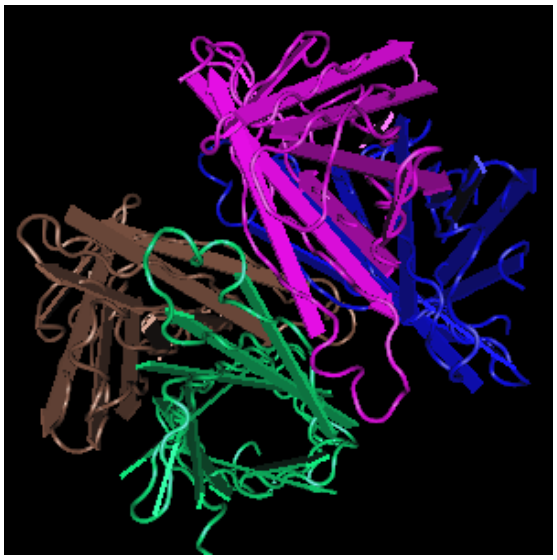

Figure 9. 3D rendering of the streptavidin molecule [13].

Streptavidin magnetic beads from New England Biolabs are superparamagnetic particles covalently bound to streptavidin. The magnetic property allows easy separation of bead-bound substrates-target molecules with the application of a magnetic field.

1.3.7 Next-Generation Sequencing: Illumina Hi-Seq

Sequencing is the process of determining in what the order the four bases (adenine, uracil, guanine, and cytosine for RNA) occur in an oligonucleotide. In the system created by Illumina, this process involves using fluorescently labeled nucleotides. The oligo to be sequenced, called the template, is modified by adding an adapter sequence containing a PCR promoter to the 5' phosphorylated end using T4 RNA ligase, and then adding a 3' adapter by reverse transcriptase polymerase chain reaction (RT-PCR) via SuperScript II reverse transcriptase. The template sequence with adapters on both sides is then bound to a plate by the ends and one end is free for the step called "bridge amplification". Fluorescent labelled bases, a different wavelength for each of the four bases, are added. Under a high-resolution camera, the labelled bases bind and fluoresce to their complementary base on the sequences being analyzed. As they bind the camera reads the wavelengths and determines which base is bound.

1.4 Information Theory Applied to Genetic Sequence

In the year 1948 Claude Shannon published his research on information theory [68]. The fundamental idea, derived from thermodynamics, is that there must exist a minimum amount of energy to change the statistical distribution of a system.

This minimum energy is $K_b T \ln(2)$, where K_b is the Boltzmann constant, equal to 1.38×10^{-23} J/K, and T is the Kelvin temperature.

The same idea was applied by Tom Schneider [41, 42, 43] that given a set of oligonucleotide sequences of length *n* with a probably of that each base is at a given state A, U, G, or C, is p_i, the conditional entropy of the state is defined as

$$H = p_i \sum_{i=0}^{n} log_2(p_i)$$

If, when the set undergoes some experimental treatment, the distribution of nucleotides, p_i, changes, then the conditional entropy of the system which was at H_{before} has changed to H_{after}, and the difference between the two is called the *information content* and it is expressed in units called "bits".

$$S = H_{after} - H_{before}$$

This change in p_i indicates that some nucleotides are being removed or added specifically. This implies that binding is happening. Thus, it can be concluded that the change in the entropy, the information content, could be directly related to the how frequently a certain sequence binds, and thus be correlated with its binding energy. This would allow researchers to more accurately determine the binding energies of oligos from sequence alone, rather than performing arduous binding experiments. If the information content can be related to the energy, then more accurate models of ribosome-mRNA interactions can be developed when optimizing genes for exogenous protein production.

1.5 Outline of Remainder of the Dissertation

Chapter 2 describes binding saturation curve experiments using radiolabeled oligonucleotides bound to ribosomes to measured ΔG˚. Chapter 3 describes a binding and sequencing experiment to simultaneously compare the binding energy of the *E.* coli 30S

ribosomal subunit with millions of short ssRNA oligonucleotides. Chapter 3 also describes a ribosome-free experiment to compare RNA binding models. Chapter 4 provides a derivation and the details of an information content algorithm for the calculation of the molecular efficiency of mRNA decoding. This is future work that is outlined but not implemented.

2 Preliminary Experiments

2.1 Introduction

The estimation of the parameters for models that predict RNA binding energy was based on empirically derived ΔG° values. The values came from experiments done *in vitro*. A frequently referenced method of determining the ΔG° of RNA binding include melting curve experiments [64,66,66] from which Turner derived many of their models' parameters. Another method is radiolabeled equilibrium experiments first described by Craven et. al in 1977 [69]. In these experiments, ribosomal protein-RNA complexes were radiolabeled and filtered through nitrocellulose which specifically filters proteins in the 30S ribosome and allows free RNA to flow through, retaining ribosome subunits and any RNA bound to them. With the RNA radiolabeled and the nitrocellulose filters easily exposed to film for measuring the RNA radioactivity, it was found that RNA-ribosome binding reactions could be maintained at equilibrium, the binding ratios measured, and the ΔG° of binding calculated. This chapter describes the use of this method in the context of radiolabeled 13-mer ssRNA oligonucleotides bound to 30S ribosomal subunits for the determination of specificity to, and activity of, the 3' tail end of the 16S rRNA.

It was hypothesized that these kinds of experiments would show that oligonucleotide binding to the 30S ribosome is a function of the complementarity with the 3' tail of the 16S subunit. Additionally, these experiments show that the rRNA of the ribosomes was present and active in the desired reaction conditions, so that when performing the sequencing experiments described in Chapter 3 it was known that they were testing the rRNA-RNA

interactions, and not an unknown or non-specific binding phenomenon. These experiments test single sequences of oligonucleotides bound to 30S subunits. The sequences have varying complementarity, and the amount of binding that occurs is used to calculate the ΔG°. These experiments measured the binding affinity of 13-mer RNA oligonucleotides to the *E.* coli ribosome. Specifically, the Gibbs free energy of binding, ΔG°, as the oligo sequence correlates to the reverse-complementarity to the 3' tail end of the 16S ribosomal subunit known as the Shine-Dalgarno sequence.

2.2 Methods

Five sequences were tested to cover a variety of potential binding strengths as well as to account for possible non-specific binding to other sections of the ribosome. The 13-mer ssRNA oligonucleotides (Table 1) were synthesized and lyophilized by Integrated DNA Technologies. Stock RNA was suspended in RNAse free water in 100 pM concentrations.

Table 1. Oligonucleotides chosen to test the binding affinity to the SD sequence. The letters in red indicate a Watson-Crick hybridization match. The rationale in sequence choice was to vary the complementarity, and thus the binding energy, to show that the relationship between the two was present and observable via saturation binding experiments. A non-complementary, low likihood of binding, sequence was used to control for non-specific binding (Nonspec - BG bind 3).

SD;	5′– UAAGGAGGUGAUC – 3′;	Anti Shine-Dalgarno
AntiSD (tail);	3′– AUUCCUCCACUAG – 5′;	
Partial 1 (P1);	5′- UAAGGAGGUCGAU – 3′;	3' end is different
AntiSD (ASD);	3′- AUUCCUCCACUAG – 5′	
Partial 2 (P2);	5′– CCAUUAGGUGAUC – 3′;	5' end is different
AntiSD;	3′– AUUCCUCCACUAG – 5′	
Partial 3 (P3);	5′– UAAGGUCCAGAUC – 3′;	Both ends bind
AntiSD;	3′– AUUCCUCCACUAG – 5′	
Nonspec;	5′– ACUAAGACACCCU – 3′;	BG bind 3

Under radiologically aware conditions (shield, hood, Geiger counter, dosimeter badge and ring), 10 pm of the AntiSD sequence (ASD), sequence Partial 1 (P1), P2, P3, and Non-specific sequence 3 (NS3), were labelled on the 5’ end with γ-ATP containing the radioactive isotope phosphorous-32. 10 µl of oligo solution of concentration 1 pm/µl were mixed in a 1.5 ml microcentrifuge tube with 2 µl γ-^{32}P (MP Biomedical cat num. 013502005), 5 µl 10x labelling buffer (0.5 M Tris-HCl, 100 mM MgCl2, 50mM DTT, pH 7.5), and 2 µl (20 U) of Optikinase enzyme (Affymetrix Cat No. 78334Y), for a total reaction volume of 50 µl. Contents were mixed by pipette, quick spun, and incubated at 37 C for 30 minutes.

Precursor ATP and enzyme existed in solution, and were removed from the oligo solution before use in further experiments. A filter was made by adding HPLC filter resin to

a cartridge (Creative Technology Systems Inc XTRX 20 Nuc Acid Purif. Cartridge, Cat No: X2NA50/10) as shown in the diagram (Figure 10).

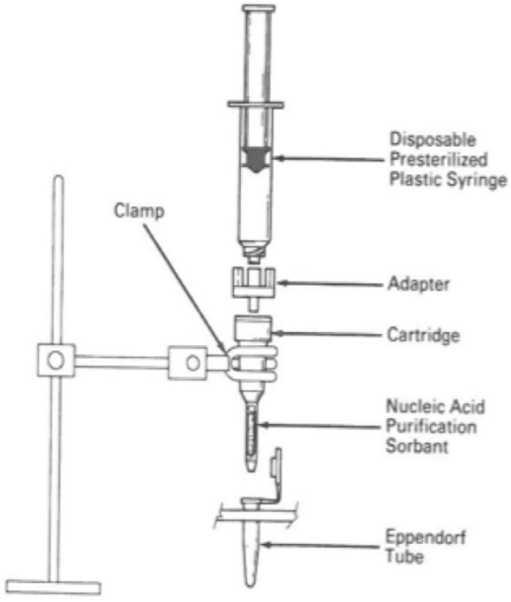

Figure 10. The diagram of the filter setup for purifying nucleic acids.

Resin was packed in the cartridge by lightly tapping it in. Using a Pasteur pipette, the cartridge was rinsed with 2 ml of 100% methanol. The adapter was securely pressed to the adapter to form an airtight seal. The methanol was pushed through, and the column was primed with Reagent A (0.1 M Tris-HCl, 10 mM TEA, 1 mM EDTA, pH 7.7) until the liquid reached the top of the bed. 200 µl of Reagent A was added to the reaction and gently mixed. Reaction mixture was added to the top of the cartridge and pushed through with air. The packed bed was washed with 3 ml of Reagent A. Nucleic acid was then eluted with 200 µl of 50% methanol. The effluent was collected, subjected to ethanol precipitation (1/5V NH3-Oac, 2.5V 95% EtOH, 1 ul glycogen), incubated at -20 C overnight, and spun in dry vacuum.

Labelled nucleic acid was resuspended in water, and an aliquot subjected to PAGE separation and phosphorimaging for analysis.

To measure the energy of binding of the oligos to the ribosomes, constant oligo concentrations were mixed with varying concentrations of ribosome 30S subunits, and the fraction of oligos bound measured by phosphorimaging.

Stock 30S ribosome subunits were prepared in-house according to Mahkno et al [24] by Dr. Tatjana Shapkina in the NCSU biochemistry lab of Dr. Paul Wollenzien, and were thawed and activated by incubation at 37 C for 10 minutes. 24 mm nitrocellulose filters with a pore size of 0.45 μm (Millipore Cat. No. HAWP02400) were soaked in HKAM7 (20 mM HEPES, pH 7.4, 30 mM KCl, 70 mM NH_4Cl, 7 mM $MgCl_2$) with 3 mM β-mercaptoethanol.

2 μl of the labelled oligo was added to 98 μl of unlabeled oligo with a concentration of 1 pm/μl. 2 μl of hot/cold oligo solution was mixed with varying amounts (2 – 200 pm) of ribosome in HKAM7 in a 50 μl reaction volume, and incubated for 30 minutes at 37 C. The reactions were then filtered through nitrocellulose filters by vacuum, and the filters dried and exposed to phosphorimaging screens. Filters not exposed to vacuum were used as total radioactivity, unlabeled oligos were used as controls, and labelled oligos filtered with no ribosome present were used as positive background. Samples were exposed to the screens for 18 hours and then scanned with a Molecular Dynamics model 445SI phosphorimager.

2.2.1 Calculation of Gibbs Free Energy, ΔG°, from the ratio of binding

The determination of ΔG° requires calculating the (apparent) equilibrium association constant K_D from a saturation binding curve. Increasing concentrations of ribosome were

added to a constant concentration of labeled RNA, and the percent of the RNA that bound was measured by its radioactivity. By fitting the saturation curve to the formula for K_D one can estimate K_D and calculate ΔG° from K_D.

b = pmol of complex formed

R = pmol of ribosomes added

A = pmol of ligand added

V = reaction volume in μl

Given a reaction between ribosome R and oligonucleotide A forming a ribosome-RNA complex b:

Reaction equation is

$$R + A \leftrightarrow b,$$

where

$$b = \frac{1}{2}(R + A + V * K_d - \sqrt{(R + A + V * (K_d)^2 - 4RA)}),$$

and

$$K_{\mathrm{d}} = \frac{[R-b][A-b]}{[b]} = \frac{(R-b_{A-b})}{b*V}.$$

From the data:

$$\text{Fraction Bound} = \frac{<b-background>}{<A>}$$

with the radiolabeled intensity denoted by < >. From the reaction equation:

$$\text{Fraction Bound} = \frac{1}{2}\left(\frac{R+VK_d}{A} + 1 - \sqrt{\left(\frac{R+VK_d}{A} + 1\right)^2 - \frac{4R}{A}}\right)$$

With K_D as a variable, the program KaleidaGraph was used to fit the equation for the fraction bound, and the ΔG˚ is calculated as:

$$\Delta G^o = RTln(k)$$

2.3 Results

The ΔG˚ for the five sequences was decreased the more complementary the sequence was with the ASD sequence, with a ΔG˚ of -9.1 kcal/mol for the matching sequence ASD to -7 kcal/mol for sample P3 which was only complementary to the ends of the 16S, and 0 kcal/mol for the oligo with no sequence complementarity. However, the magnitudes and range of the measurements are much lower than the predictions, varying from -7 to -9.1 kcal/mol, while the predictions ranged from -6.7 to -21 kcal/mol (Figure 11).

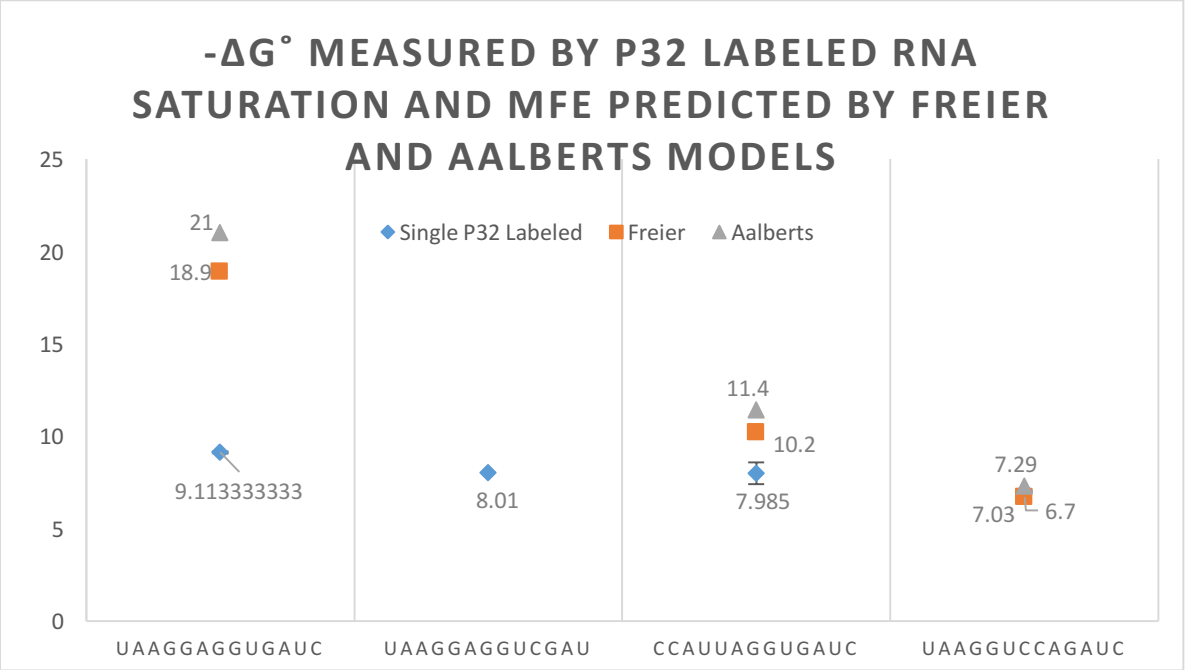

Figure 11. The ΔG˚ for the binding between four RNA oligos of varying complementarity to the 30S ribosome. The minimum free energies shown are predictions by two models, the Aalberts' BINDIGO model, and the free_scan program using the Freier settings.

The phosphorimaging results show that while most radiolabeled oligos pass through the nitrocellulose filter unabated, some radioactivity is still retained either by labeled oligos or unbound γ-ATP, and must be taken into account as background radiation that does not count as oligos bound to the ribosome (Figure 12).

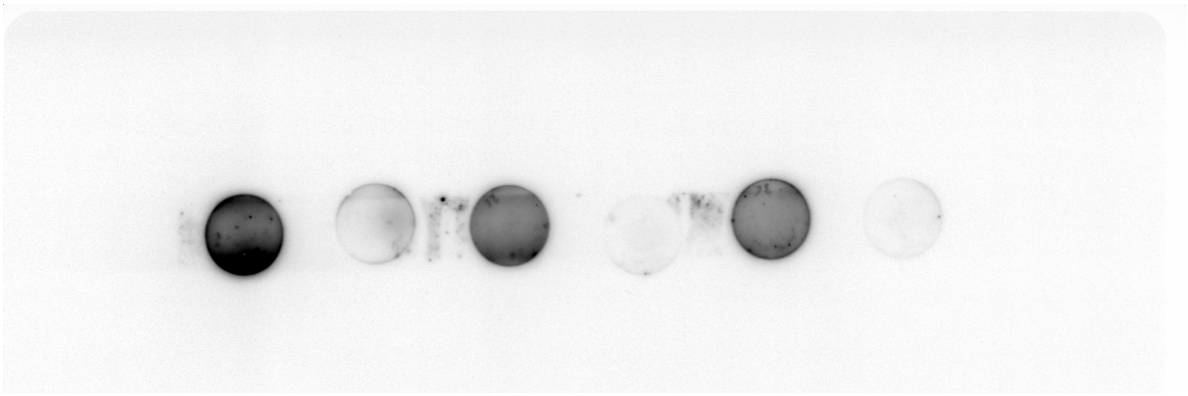

Figure 12. A test measurement of P32-labeled RNA, showing positive and background controls for three concentrations, 1, 0.75, and 0.5 pm/µl RNA. Dark circles are the positive controls of unfiltered RNA, and the background controls are the lighter circles indicating some radioactivity is still present after filtration.

The control sequence NS3 was run twice, with up to 240 pmols of ribosome, and showed no binding affinity to the ribosome (Figure 13).

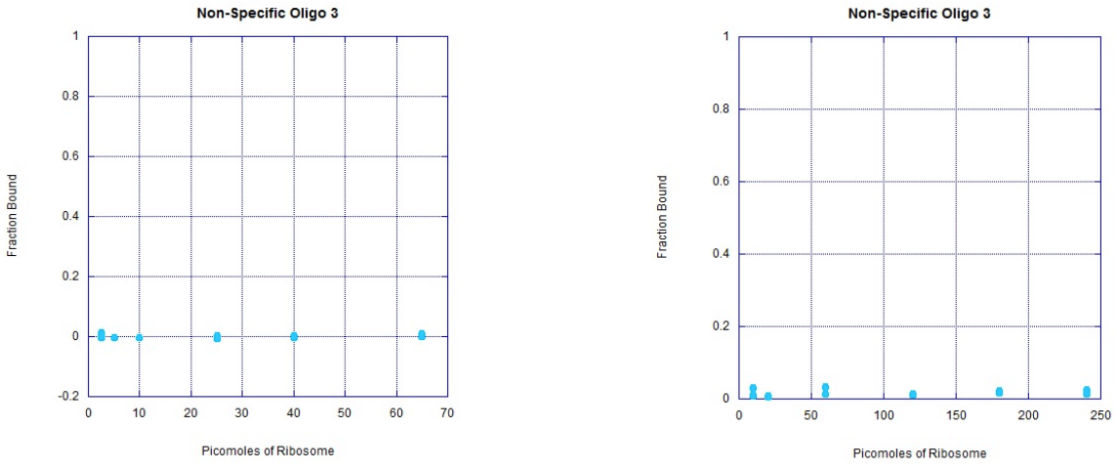

Figure 13. Two runs of the control sequence 5' – ACUAAGACACCCU – 3', showing no correlation between increasing ribosome concentrations and the fraction of oligos bound.

Sequence ASD, which was the complement to the 16S tail, was run in triplicate and had an average ΔG˚ of -9.1 +/- 0.05 kcal/mol. A consistent result was a lowered fraction bound at 80 pmol (1.6 pmol/μl) of ribosomes were in solution (Figure 14).

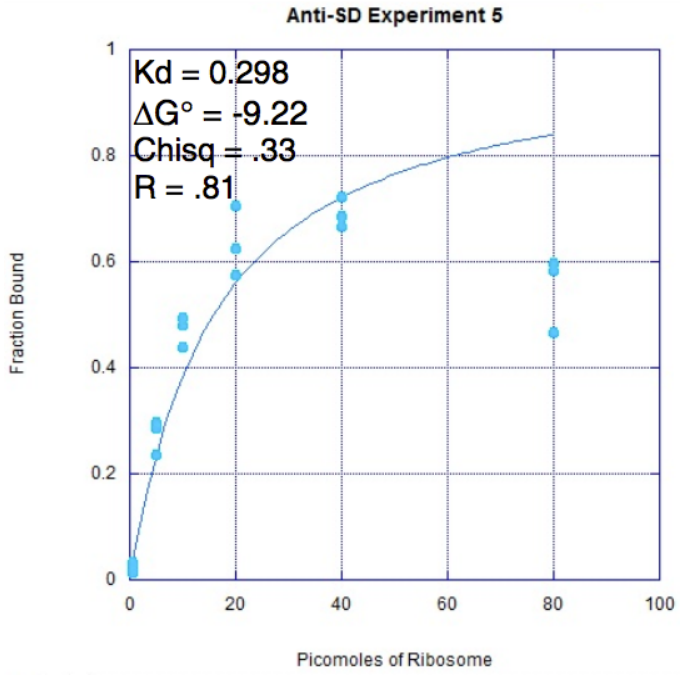

Figure 14. One of the three saturation experiments between sequence ASD (5' - UAAGGAGGUGAUC - 3') and the 30S ribosome. The apparent ΔG˚ was found to be -9.22 kcal/mol for this repeat. The fraction bound came to a maximum of about 0.7 with 40 pmol ribosomes to oligos, with higher ribosome concentrations resulting in lower fractions bound of about 0.57 at 80 pmol of ribosome.

Experimental samples P1, P2, and P3 were run under identical conditions resulting in ΔG˚ values of -8.0, -8.0, and -7.0 kcal/mol respectively. Similar lowering of fractions bound at higher ribosome concentrations were observed.

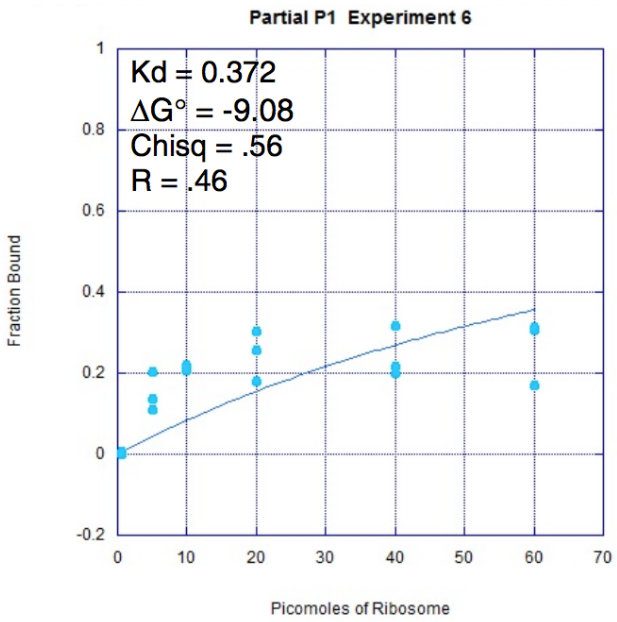

Figure 15. A binding saturation curve for sequence 5' – UAAGGAGGUCGAU – 3', sample P1. The fraction bound levels off around 0.25 to 0.3 at 60 pmols of ribosome.

2.4 Discussion

The binding fractions of the complementary sequences showed that binding occurred between these sequences and the ribosome. The lack of binding of the non-complementary sequences indicates that the binding fraction is a function of the nucleotide sequence. The significant difference between the ΔG˚ of the sequences correlates with their degree of sequence complementarity, which also follows the MFE predictions made by the models. However, while the observed trends follow those that are predicted by the models, there is a difference in the magnitude of the ΔG˚ and the MFEs for the more highly complementary sequences. When the ribosomes get to a higher concentration, the binding of the labelled oligos is reduced, which indicates that some sort of inhibition is occurring between ribosomes, reducing the ability of the oligos to

bind. It has been found that dimerization of inactive 30S ribosomal subunits can occur in 10 mM Mg++ [69]. More recent work describes crystal contacts between two subunits that result in the 3' rRNA end acting as a mRNA codon mimic and binding to the P- and E- sites [71]. These effects could block our radiolabeled oligos from binding when the ribosome concentrations get high enough. This may be happening at lower concentrations as well, but the effects may not be observed clearly until the ribosome concentrations are greater than 1.8 pmol/μl (80 pmol total in Figure 14). This effect may have reduced the binding fraction at all concentrations, and thus reduced the observed ΔG° as well. Regardless, the specificity of the ribosome was established, i.e. the ΔG° of binding was correlated with the degree of sequence complementarity to the 16S tail sequence, which had to be done to ensure a large scale (and more expensive) binding experiment would be fruitful.

For the purpose of establishing the accuracy of RNA binding predictions, this method is fairly limited in its scope, allowing only one, or possibly a few sequences to be tested at a time were one to use a fluorescence labelling quantification rather than beta radiation to measure the fraction bound. For that reason, this method was not pursued further to test the 67 million other possible 13-mer sequences that the 3' tail of the ribosome could potentially encounter during translation. For this reason, a more comprehensive method needed to be established. As a tool for measuring individual sequence binding strengths however, saturation binding curve experiments are robust and reproducible as shown here and in prior work. It is a recommended step take to establish reaction conditions and sequence specificity before taking the samples to another domain,

next generation sequencing comparisons during competitive binding, which is the subject of Chapter 3 of this dissertation.

3 RNA binding strength analysis determined by Next-Gen sequencing compared with RNA:RNA minimum free energy predictions

3.1 Abstract

Many algorithms exist for predicting the binding energy of RNA:RNA interactions. Algorithms based on the Individual Nearest Neighbor model have parameters derived from melting curve experiments (*i.e.*, Turner rules), but predictions of binding energy are frequently somewhat different depending on which model is used. For example, the BINDIGO RNA modeling program predicts the sequence 5' – GGAGGAGGUGAUC – 3' will bind the 16S tail sequence 5' – GAUCACCUCCUUA – 3' with a minimum free energy of -21.2 kcal/mol, and is the strongest possible sequence that could bind to the tail, while the free_scan model with the Mathews rules and the Turner 2004 parameters predicts this sequence binding with an MFE of -19.9 kcal/mol, and is the 10^{th} strongest possible. To model a ribosome interacting with single-stranded RNA in the process of translation, an accurate prediction of RNA:RNA binding energy is needed. Here we have used oligonucleotide libraries of ~10^7 unique 13-mer sequences to determine the relative binding strengths to the anti-Shine-Dalgarno (ASD) sequence immobilized on nanobeads or within the bacterial ribosome. Gibbs free energy ($\Delta G°$) values were estimated by analysis of sequencing data derived from the libraries before and after binding to the ASD oligomer and the ribosomes. Estimated $\Delta G°$ values were then compared to the minimum free energies predicted by several RNA free energy binding models. We found that the free_scan model with the combined Freier-Mathews parameter set best predicted the rank order for *in vitro*

ribosome experiments, whereas the ViennaRNA package with the 2007 parameters best predicted the rank order for the nanobead-bound oligomer binding experiment. However, each model had an optimal fit over different ranges of ΔG° values, indicating that model accuracy may be context dependent, *i.e.,* RNA binding to ASD oligos on beads *vs.* RNA binding to the ASD of ribosomes. Finally, competition between sequences and from intramolecular interactions was evaluated *in silico* using the NUPAC package of nucleic acid modeling and determined to be the likely cause of the compression of observed ΔG˚ values *in vitro*. This observation suggestion re-evaluation of traditional models for 13-mer (or greater) oligonucleotides using more comprehensive modeling that accounts for inter and intramolecular competition. Data from these experiments can be used to compare and improve RNA binding models and support protein translation simulation models requiring accurate prediction of binding energies.

3.2 Significance Statement

RNA has multiple functions, including structural and enzymatic functions in the framework of ribosomes and transfer RNA, regulation of gene expression (small regulatory RNAs), and as a carrier of genetic information (messenger RNA). These functions depend on inter and intra molecular binding of RNA molecules by hydrogen bonding of ribonucleotide bases. Several models for estimating binding energies have been developed, however, predicted Gibbs free energy (ΔG°) values vary between models. A protein translation model has been developed using ΔG° predictions. For making novel proteins as medications, industrial enzymes, and understanding disease, it is crucial to have accurate binding energy predictions. This work addresses the need to understand how computational models differ in the prediction of RNA oligonucleotide-ribosome interactions by utilizing sequencing technology for massively parallel analysis.

3.3 Introduction

The ubiquitous role of RNA in biology mandates the need to thoroughly understand its behavior for many applications. One medically and industrially important role occurs during the process of protein translation when a ribosome mediates messenger RNA and tRNA engagement and movement. A free energy model of mRNA translation (FET) [39] has been developed that uses a signal processing approach to describe an error-control system to maintain the ribosome in correct reading frame of the mRNA [30,27,31]. The model showed that the free energy binding between mRNA and the 3' end of the 16S rRNA of bacterial ribosomes (the anti-Shine-Dalgarno sequence (ASD) [46]) exhibits a sinusoidal signal as the ribosome traverses the length of the mRNA. It was hypothesized that the free energy of binding of the ASD sequence to mRNA as the ribosome moves along the mRNA helps enable the ribosome to maintain the correct reading frame. This model has been used to predict improvements in translational efficiency of selected mRNA sequences, to detect genome annotation errors, to identify programmed frameshifts, to locate pre-mRNA and intron-exon splice sites, and to identify putative protein coding regions [31,34,35,36,46,49,50,55,60].

Implementation of the FET model requires a calculation of the minimum free energy (MFE) between two short single-stranded RNA (ssRNA) oligonucleotides. For bacterial translation, this calculation would be performed with the 13 bp ASD sequence of the 16S rRNA of a given species and sequential 13 bp sequences along the length of the mRNA coding region in that organism. To calculate the MFE values for these 13 bp sequences and the ASD, the free_scan model [49,46, 50] has been used. The free_scan model for predicting

13-mer RNA:RNA MFE values gave robust results in finding +1 frameshifts during translation of mRNA sequences [50]. The free_scan model predictions are based on a combination of the Individual Nearest Neighbor Hydrogen Bond (INN-HB) algorithm with hybridization parameter values from Xia et al [58], G/U mismatch parameters from Mathews et al, and loop penalties from Freier et al. [11,27]. However, other models for MFE predictions are available, including UNAfold [25], BINDIGO [19], and ViennaRNA [23], which have been developed to improve RNA:RNA oligomer binding predictions. A comparative analyses of MFE by the free_scan model and the BINDIGO revealed differences in rankings of RNA:RNA binding strength. The FET model can be used to predict a variety of ribosome-mRNA interactions *in vivo,* including ribosome pause times that are partially dependent on the predicted MFE values [22]. In particular, there may be instances in which an inaccurate prediction of RNA binding results in a model of ribosome behavior that displaces the ribosome away from the correct reading frame. Therefore, we investigated RNA:RNA oligomer binding models to determine the accuracy of RNA binding models within the context of 13-mer interactions.

Previous methods to determine the thermodynamic parameters of RNA:RNA binding include UV melting curve experiments [8,12,16,17,45,59,44,2]. The temperature at which a change in UV absorbance occurs during RNA:RNA strand separation is used to calculate the ΔG°. Two limitations of optical melting experiments lie in their scope and their context. First, only one or two sequences can be tested at a time, making experimental analysis of multiple sequence changes intractable. Second, the binding and melting experiments are

performed free in solution which does not include the impact of competitive binding reactions by other RNA sequences on MFE calculation.

To determine ΔG° values for 13-mer binding, we used randomized, 13 base ssRNA oligomer libraries to simultaneously test over 10 million unique sequence binding events. Two independent experiments were performed, one using purified ribosomes with the ASD as the target for the randomized ssRNA pool and the other using a 13 bp ASD sequence tethered to magnetic streptavidin nanobeads. The resulting relative binding frequencies were then used to measure the free energy of binding based on sequence data of unbound oligos from the random pool. Unbound RNA data was used to avoid bias from elution procedures. The experimentally determined ΔG° values were then compared to MFE model predictions for the free_scan [49], UNAfold [25], ViennaRNA [23] with the Turner 2004 and the Andronescu [1] parameter sets [26,1], MultiRNAfold [3], and BINDIGO [19] models. *In silico* predictions of the effect of competition were performed using the NUPACK 3.2 [61] software package. The results are extrapolated to large data sets, effectively comparing normalized ΔG˚ measurements between oligonucleotide sequences to evaluate the fitness of RNA binding predictions. Our methodology and data can be used to improve the understanding of RNA:RNA interactions for a variety of applications.

3.4 Methods

3.4.1 Ribosome Experiments

Short (13-mer) single-stranded RNA oligonucleotides were synthesized by Integrated DNA Technologies (IDT) with 5' phosphorylation. The first 12 nucleotides were randomized during synthesis (25% chance of insertion of each nucleotide), although the 13^{th} nucleotide was a cytosine due to synthesis restrictions. The random oligonucleotide preparation was diluted prior to use to 689 pmol/μl, and designated RAND1. A second, similarly prepared sample was synthesized and designated RAND2. To calculate the randomized representation, the number of unique sequences was measured by sequencing. Ribosome experiments (see below) were conducted using sample RAND1 and bead experiments (below) were done with RAND2. The total number of possible sequences in each oligo preparation is 4^{12}, or 16.78 million.

E. *coli* 70S ribosomes were prepared according to Makhno et al. [24]. The 30S ribosomes were separated from 70S ribosomes by ultracentrifugation through a 10-40% sucrose gradient in buffer containing 20 mM TrisHCl, pH 7.5, 200 mM NH4Cl, 3 mM MgCl2 and 3 mM β-mercaptoethanol. The gradient was prepared with a Hofer SG/50 Gradient Mixer, and ribosomes were centrifuged at 27,000 RPM (131,000 x g) at 4°C under vacuum for 18 hours. The samples were fractionated (Gilson R105 Do1sy micro-fractionator, Rainin "Rabbit-Plus" peristaltic pump), and the presence of the subunits measured by ultraviolet visualization (Isco model T11 96206 UA-6 UV/VIS detector). The fractions containing the 30S subunits were marked on the UV-VIS output as the second spike of the detector output, the first spike indicating the 70S subunit. The 30S subunit fractions were

then pelleted by ultracentrifugation at 40,000 RPM (193,000 x g) overnight. Pellets were resuspended in buffer (20 mM TrisHCl, pH 7.5, 200 mM NH4Cl, 20 mM MgCl2 and 3 mM β-mercaptoethanol) for a final 30S ribosome concentration of 4.83 pmol/µl.

The 13-mer RAND1 oligo library was bound to 1.9 pm/µl of 30S subunit in 2.2X excess (oligo to ribosome) in HKAM binding buffer: 20 mM HEPES, pH 7.4, 30 mM KCl, 70 mM NH_4Cl, 7 mM $MgCl_2$, 3 mM β-mercaptoethanol. The ribosome oligo solution (final volume of 50 µl) was incubated at 37°C for 30 minutes to allow binding of the oligos to the ribosome subunit. The reaction mixture was then subjected to size exclusion separation on a 500 kDa molecular weight cutoff filter (Sartorius Vivaspin 500) and centrifuged at 16k RPM for 1 min. The reaction mixture and the filter were rinsed with 100 µl of HKAM7 buffer, centrifuged as above, and final rinse was done with 50 µl of 8 mM EDTA to elute weakly bound oligos from the 30S subunits which were in 7 mM MgCl solution. The flow-through from the EDTA rinse was collected, and the RNA concentration determined with an Agilent 2100 Bioanalyzer. Eluted RNA was then subjected to Illumina Hi-Seq 2500 small RNA next-generation sequencing. Three samples were sequenced: the initial randomized oligo (RAND), and two replicate experimental samples with the oligo recovered (following elution from the 30S ribosomes as described above) designated RCVD1 and RCVD2. The sample libraries were prepared for sequencing by the NCSU Genomic Science Laboratory according to the Illumina® TruSeq™ Small RNA Sample Preparation protocol, multiplexed with three unique barcodes, pooled and run on 11 lanes yielding a total of ~900 million reads.

3.4.2 Nanobead Experiments

In addition to oligo-30S ribosome binding experiments, 1 μm superparamagnetic streptavidin nanobeads (New England BioLabs S1420S) cross-linked to an ASD 13 bp RNA oligonucleotide were used as the target for binding of the RAND2 oligos. The ASD oligo (5'–AUAAGGAGGUGAUC–3') was synthesized (Integrated DNA Technologies) with a biotin molecule present on the 5' end. Two bead solutions labelled BRCVD1 and BRCVD2 were prepared aliquoting 120 μl of stock (4 mg/ml in PBS, 0.1% BSA, 0.02% NaN_3, pH 7.4), mixing by vortex, separating by magnetic field, and resuspending in 100 μl wash/binding buffer (0.5 M NaCl, 20 mM Tris-HCl, 1 mM EDTA, pH 7.5). 100 pmol of stock biotinylated oligonucleotide diluted 1:100 (1 μl) was mixed with 100 μl wash/binding buffer, then added to the bead solutions. The biotinylated RNA and streptavidin beads were incubated at room temperature for 5 minutes to allow binding. The beads were then separated by placing the tubes in a magnetic field for 3 minutes until the solutions were clear and the beads were all on one side of the tubes. The effluent was removed via pipette and the remaining beads were washed 3 times with 225 μl of wash buffer and resuspended in 50 μl HKAM7 buffer.

5.5 μl of RAND2 1:100 (100 pmol/μl) was added to BRCVD1 and BRCVD2 and incubated at 37°C for 30 minutes. The beads were then separated from solution by magnetic field and the effluent removed for sequencing. The beads were then resuspended in 50 μl wash/binding buffer. Bead-tail complexes were again separated from solution by magnetic field, leaving behind the oligonucleotides from the randomized group RAND2 that had previously hybridized. This effluent was also removed for sequencing. As a control, a 1 μl aliquot of RAND2 1:100 was added to 49 μl wash/binding buffer and sequenced.

3.4.3 Data Analysis

All sequencing data were subjected to FASTQ quality filtering using the FASTX-Toolkit [6] with a cutoff PHRED value of Q > 13, equivalent to 95% confidence of base calls for all nucleotides. A comparison of Q > 13 to Q > 30 (99% confidence) was performed and determined that 0.2% of reads contained at least one base call with a confidence between 95% and 99%. With such a low difference it was determined that it is more important to maintain a larger representation of the high energy oligos for which there are fewer examples by using a base call cutoff of 95%. Sequences were extracted from the filtered fastq files using the *kseq* source and library developed by Heng Li (S1). Sequence data for RAND1 and RAND2 (untreated randomized oligos) were used to normalize heterogeneity introduced during synthesis of the oligos using custom software (S1). The 13-mer sequence was extracted from raw sequence data to remove the sequencing adapters and all truncated or extended sequences were removed. A position-specific distribution matrix program was created to determine the relative frequencies of bases at each location. All possible 13-mer sequences were recreated *in silico*, and the number of times any given sequence was identified was tabulated and matched to the table of all possible sequences. The unique sequences were identified and a histogram of sequences was ordered by the MFE, based on model predictions for the oligo binding to the ASD.

The predicted MFE for the UNAfold 3.9, MultiRNAfold, BINDIGO and free_scan were calculated for all possible 13-mers present using compiled source code for each package (S1). The MFE values calculated by the Vienna model were run using RNAcofold

from the ViennaRNA Package version 2.1.8. Two versions of ViennaRNA were run using the Andronescu 2007 parameter set and the Turner 2004 parameters. Two versions of free_scan were also run using two parameter sets, the Freier version and the Xia versions. All MFE values were rounded down to one significant figure to allow proper comparative linear regression analysis.

The batch (run as a group) secondary structure predictions of all possible MFE tail binding events by the Vienna and UNAfold models were run and an analysis of unbound bases was performed. All custom programs were written in C, compiled and executed on a 64-bit machine under a Unix-based operating system (S1).

A disassociation constant (*K*) for each sequence was calculated from the ratio of concentrations, as determined before and after the binding event. The concentrations were based on the sequence frequency in the recovered samples and the measured RNA concentration, as described above. To calculate a concentration, the number of reads for any given sequence was divided by the total number of reads for that sample and multiplied by the total RNA concentration measured.

For each unique 13mer A_i and ribosome (or nanobead) B, the chemical equation is

$$A_i + B \rightarrow A_i B$$

and the disassociation constant K_i becomes

$$K_i = \frac{[A_i B]}{[A_i]_{free}[B]_{free}}$$

and the ribosome-oligo conjugate concentration can also be defined from the measured concentrations before and after binding as the total and free concentrations:

$$[A_iB] = [A_i]_{total} - [A_i]_{free}$$

$$[B]_{free} = [B]_{total} - [A_iB]$$

n = Total number of unique sequences in sample

$$[B]_{free} = [B]_{total} - \sum_{i=1}^{n} [A_i]_{total} - [A_i]_{free}$$

$$K_i = \frac{[A_i]_{total} - [A_i]_{free}}{[A_i]_{free}([B]_{total} - \sum_{i=1}^{n} [A_i]_{total} - [A_i]_{free})}$$

$$\Delta G_i^{\circ} = -RT \ln(K_i)$$

The calculations were performed in PTC Mathcad Prime 3.0. The sequences were ranked by their experimentally determined ΔG˚ and compared to the predicted values of the seven model implementations. To determine the ratio that $[A_i]_{free}$ would have to be reduced by to equal the computational models' predictions of ΔG˚, $[A_i]_{free}$ was defined as a function of K, and ΔG˚ is substituted. The original value for $[A_i]_{free}$ is then divided by the new.

Samples were then normalized by feature rescaling to allow comparisons between samples. The measured ΔG˚ was normalized by the minimum and maximum ΔG˚ in their set, such that:

$$\Delta G^{\circ'} = \frac{\Delta G^{\circ} - \min(\Delta G^{\circ})}{\max(\Delta G^{\circ}) - \min(\Delta G^{\circ})}$$

The sample data were combined, and the fitness of the models was evaluated by an adaptive linear regression script programmed in JSL scripting language and implemented by JMP version 12.2.0. The variable-range sliding-window standard least-squares linear regression on the normalized measured ΔG˚ was applied with dynamic start and end points.

The regression was weighted by the number of different sequences that were predicted by that model to bind at each MFE value. Regressions began at -2.0 kcal/mol, with a starting span of 3.0 kcal/mol. The window start point was incremented by 0.1 and a new regression conducted until the end point was equal to the largest magnitude MFE calculated by each model. Once the entire range was covered, the span of the window was increased by 0.1 and rerun until the span was 15.0 kcal/mol in width. This analysis was applied to all of the models tested resulting in between 5,500 and 10,900 unique regressions depending on the model.

To evaluate the effect of competition on the observed ΔG°, 100 strong binding (MFE < -10 kcal/mol) 13-mer sequences were randomly selected and bound to the tail *in silico* with 100, 300, 700, and 1180 random weak binding (MFE > -4 kcal/mol) sequences using the NUPACK package "complexes" program, and their equilibrium concentrations predicted using the "concentrations" program. The 100 sequences and all competitors were assigned the same starting concentration of 1.25e-13 M with the tail sequence assigned a concentration of 1.25e-12 M. The final bound tail complex concentration was used to calculate an observed ΔG°. The process was repeated using 100 random weak binding sequences with up to 1180 random strong binders.

3.5 Results

3.5.1 Next-Generation Sequencing

The nucleotide distribution for the synthesized libraries RAND1 and RAND2 were found to be skewed from the theoretical 25% per base at each position, based on sequence

data (Figure 16).

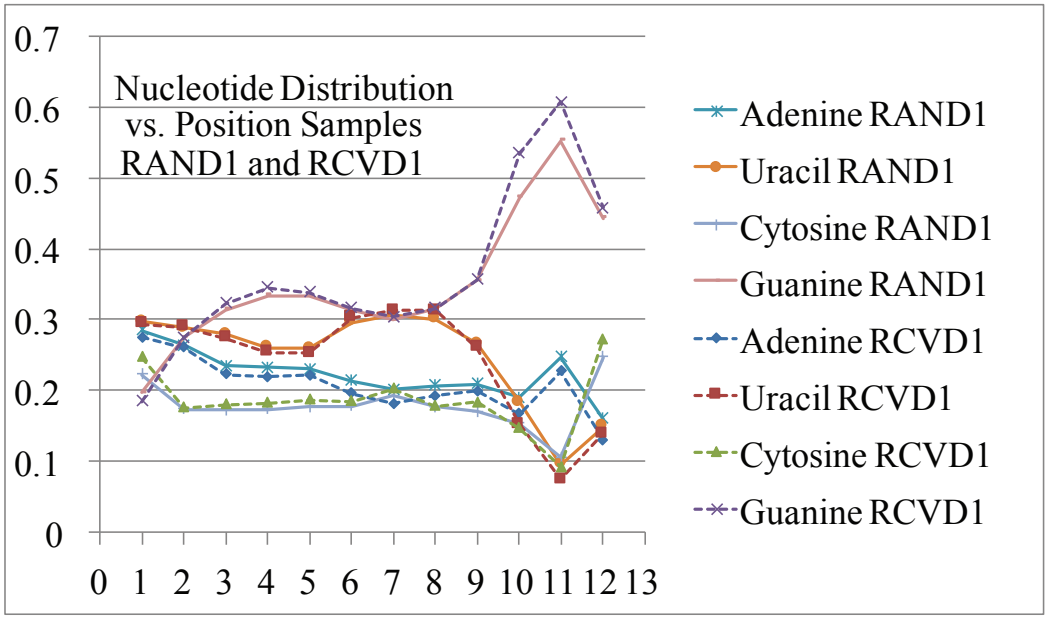

Figure 16. The starting nucleotide distribution of the "randomized" sample RAND1 and the distribution after binding to ribosome sample RCVD1 according to the location along the oligonucleotide. The plots for RAND1 and RCVD1 are constructed from the means of 2.1×10^{11} and 4.4×10^{11} data points.

The first nine bases of the RAND1 13-mer sequence had adenine (A), uracil (U), and cytosine (C) nucleotides distributions ranging from around 18% to 34%. Bases 10-12 had guanine (G) frequencies ranging from 36% to 54%, with correspondingly lower C, A, and U insertions at those positions. Similar results were seen for the RAND2 oligo (data not shown). Additionally, the change in the nucleotide distribution from RAND1 to RCVD1 indicates selection has occurred during oligonucleotide manufacture, with a strong selection for sequences with G bases at location 10 and 11, increasing the distributions by 6.3% and 5.5%, respectively. Statistics for the sequencing reactions are shown in Table 2.

Table 2. The statistics of the sequencing of each sample from each experiment, including the number of 13-mer reads after quality control for adapter dimers and PHRED value cut-off. The number of unique sequences read, with the mean times each unique sequence was read**.**

Sample	Description	Filtered Reads	Number Unique seen out of 17M	Mean times sequence was seen	Standard Deviation
RAND	Ribosome control	212,827,395	10,830,629	19.65	42.74
RCVD1	Ribosome sample 1	436,040,651	10,305,158	42.31	128.39
RCVD2	Ribosome sample 2	296,012,977	11,198,411	26.43	67.4
RAND2	Nanobead control	37,870,363	5,633,137	6.72	13.39
BRCVD1	Nanobead sample 1	53,120,905	5,657,813	9.39	21.97
BRCVD2	Nanobead sample 2	53,031,435	5,660,722	9.37	21.93

Data for RAND1 to RCVD1 or RCVD2 samples before and after binding to 30S ribosomes showed that the number of unique sequences recovered decreased (for example10.8M for RAND1 to 10.3M for RCVD1), although there were more than twice as many reads from sequencing (213M compared to 436M). Specific oligomer sequences in sample RAND1 bound to the target, thus they were not present in the RCVD1 sample. The standard deviation from the mean multiplied by each sequence was observed to increase from 42.7 to 128.4, showing a decrease of specific sequence concentrations and an increase in non-specific sequences, indicating that selection had occurred.

Analysis of the effect of competition performed by the NUPACK software package illustrated in Figure 17 shows the predicted reduction of the observed ΔG° as the number of competitors in solution increases.

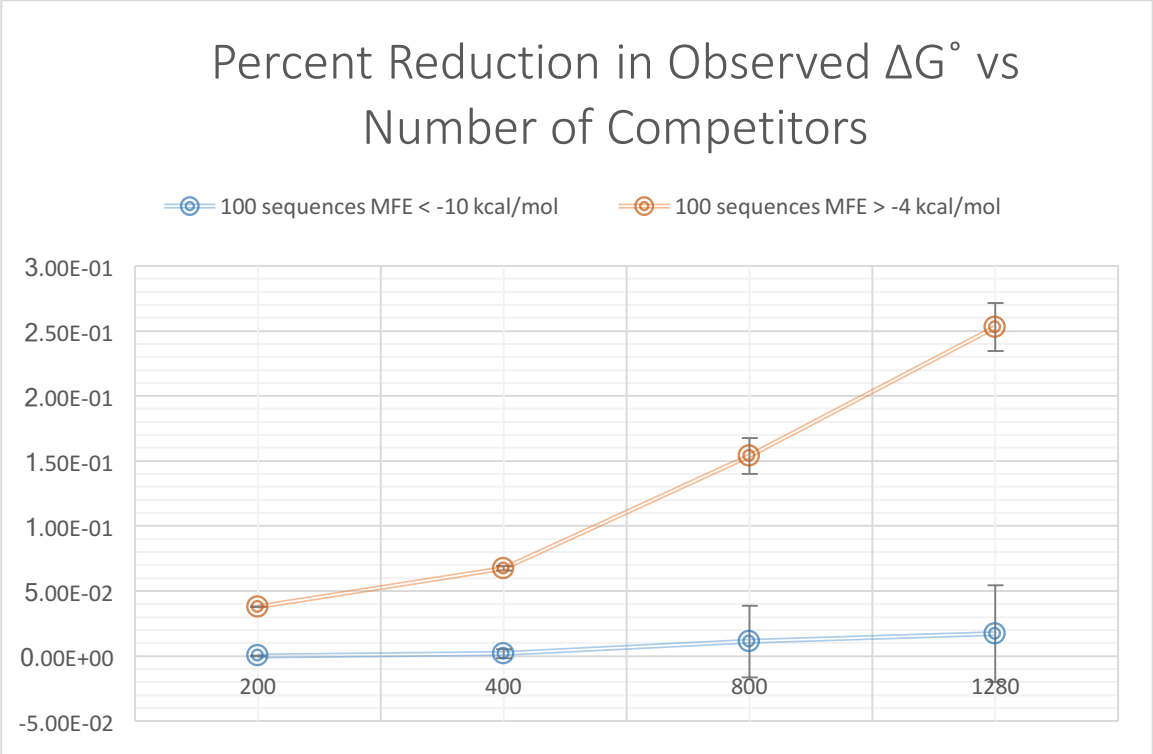

Figure 17. The percent reduction in the observed ΔG˚ of 100 randomly selected sequences binding to the tail sequence versus the number of competitors as predicted by the NUPACK "complexes" and "concentrations" programs. Only a 1380 sequences due to memory and constraints and processing time (O(N^4) time and $O(N^2)$ storage)

Strong binding sequence in competition show an average predicted reduction in observed ΔG˚ of 0.02%, while weak binding sequences show a greater reduction of 25.29%.

3.5.2 Model Comparisons

Figure 18 shows a comparison of the experimentally determined ΔG° and predicted MFE as calculated by the ViennaRNA package version 2.1.8.

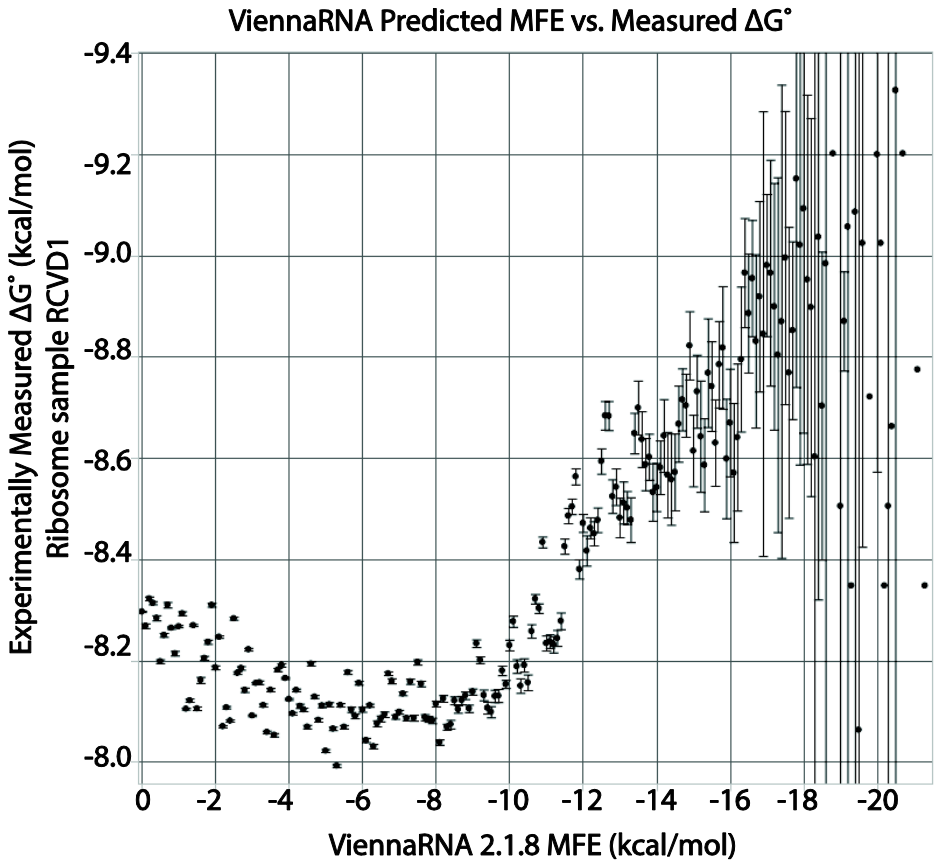

Figure 18. The experimentally determined ΔG° versus the MFE ΔG° predicted by the ViennaRNA version 2.1.8, using the Turner 2004 parameters for ribosome sample RCVD1. The figure includes 191 means of 8,708,232 data points. Each error bar is constructed using a 95% confidence interval of the mean value of the observed ΔG˚ on the y-axis for all the sequences that are predicted to have each MFE for the Vienna model on the x-axis. For the RNA Vienna MFE of -9, there are 25,114 unique sequences predicted to bind to the ASD with that energy. The mean observed ΔG˚ is 8.214 kcal/mol, with a 95% confidence that the mean is between -8.219 and -8.209.

The graph of observed *versus* predicted energy values shows three distinct regions, including low binding energy oligomers whose recovery was greater than expected (for predicted values of 0 to approximately -6 kcal/mol), a linear range (-8 to -18.9 kcal/mol), and a low confidence region (-18.9 to -21 kcal/mol). Similar trends were observed with all other models.

To determine if the low energy recovery region (Figure 19a) was from oligomer-oligomer dimers or multimers in which oligomers only partially bound to the ribosome, the number of unbound bases (compared to the ASD target) calculated with ViennaRNA secondary structure prediction program RNAcofold was compared with the MFE for each oligo in this region. The data show that the predicted energy for 50% of the nucleotides bound corresponds to the start of the linear range of -8.1 to -19.9 kcal/mol (Figure 19b).

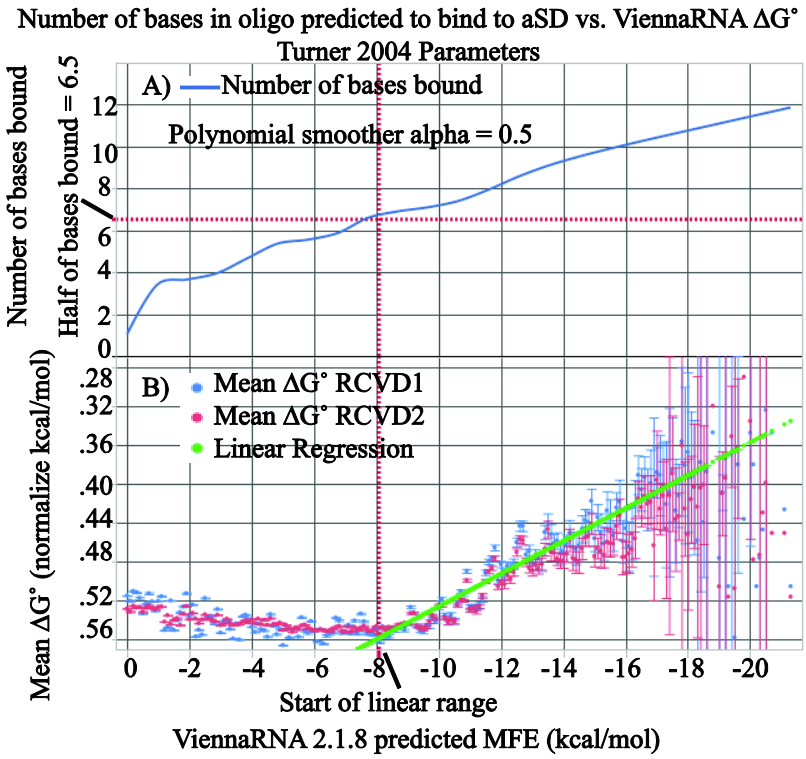

Figure 19. An analysis of the predicted secondary structures of the duplexes by ViennaRNA shows the average number of unbound bases present for all oligomers at each predicted minimum free energy (A). The experimentally measured ΔG° values for those oligomers are show above (B), with the optimal regression start location shown on the x-axis as the vertical dotted red line, and the horizontal dotted red line at which half of the bases are predicted to bind to the ASD.

For the stronger binding oligomers (ΔG° < -15 kcal/mol), an analysis of the confidence interval of the mean ΔG° for each value of predicted MFE was performed for the Vienna 2004 model with data from the bead-oligo binding experiments for sample BRCVD1 (Figure 20). The values for ΔG° beyond the linear range of -14.9 kcal/mol rose steeply as the number of sequences available to analyze dropped. The smaller interval of confidence for the strong binding oligomers correlated with the linear range cutoff.

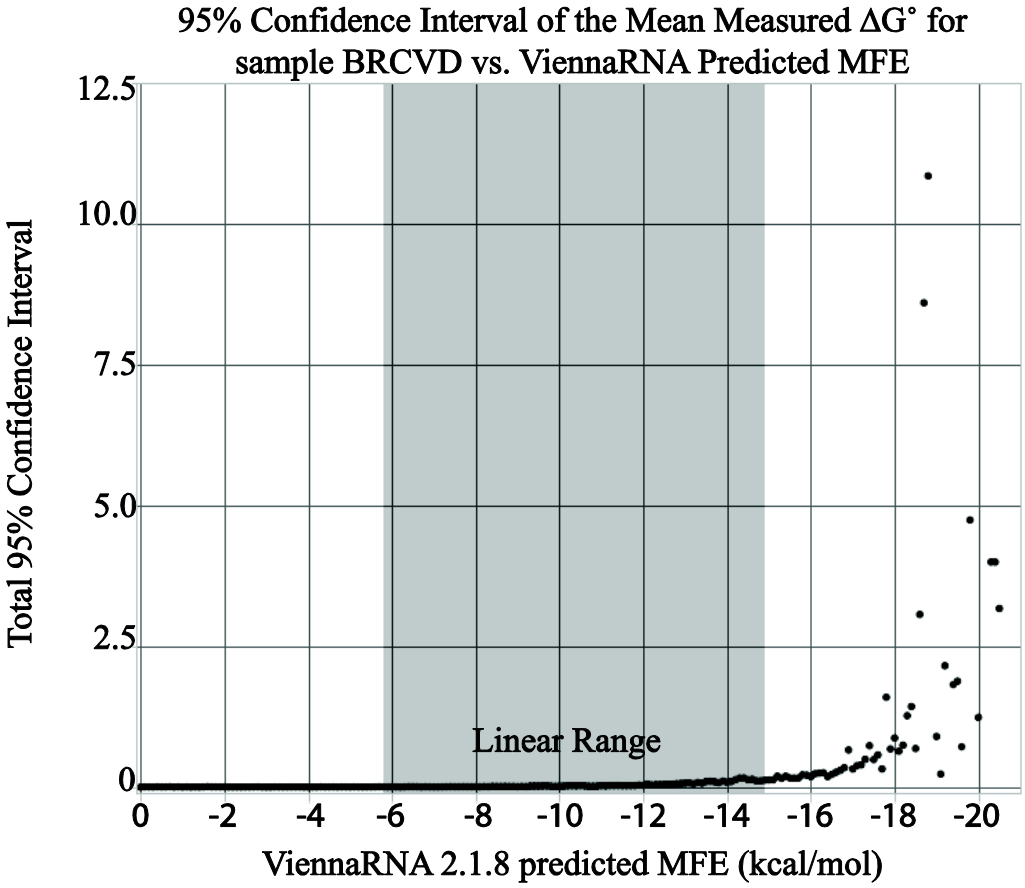

Figure 20. The 95% confidence interval of the mean of the 200 experimentally determined ΔG˚ values versus the predicted MFE. The linear range shown spans -5.8 to -14.9 kcal/mol for bead sample BRCVD1 with the ViennaRNA Package predictions using the Turner 2004 parameter set.

For comparative analysis of individual sequence binding strengths of oligomers to the ASD of the ribosomal subunits and the bead-bound ASD oligomers, the data were plotted as the predicted MFE versus the normalized mean ΔG° measured for each sample and model. Figure 21 and Figure 22 show the Unafold MFE predictions versus the measured mean ΔG°, the linear range and linear fit and the regression line for the ribosome and bead samples, respectively. The linear regressions for the ribosome (RCVD1 and RCVD2) and bead (BRCVD1 and BRCVD2) experiments had slopes of 0.015 and 0.012 with y-intercepts of 0.493 and 0.659.

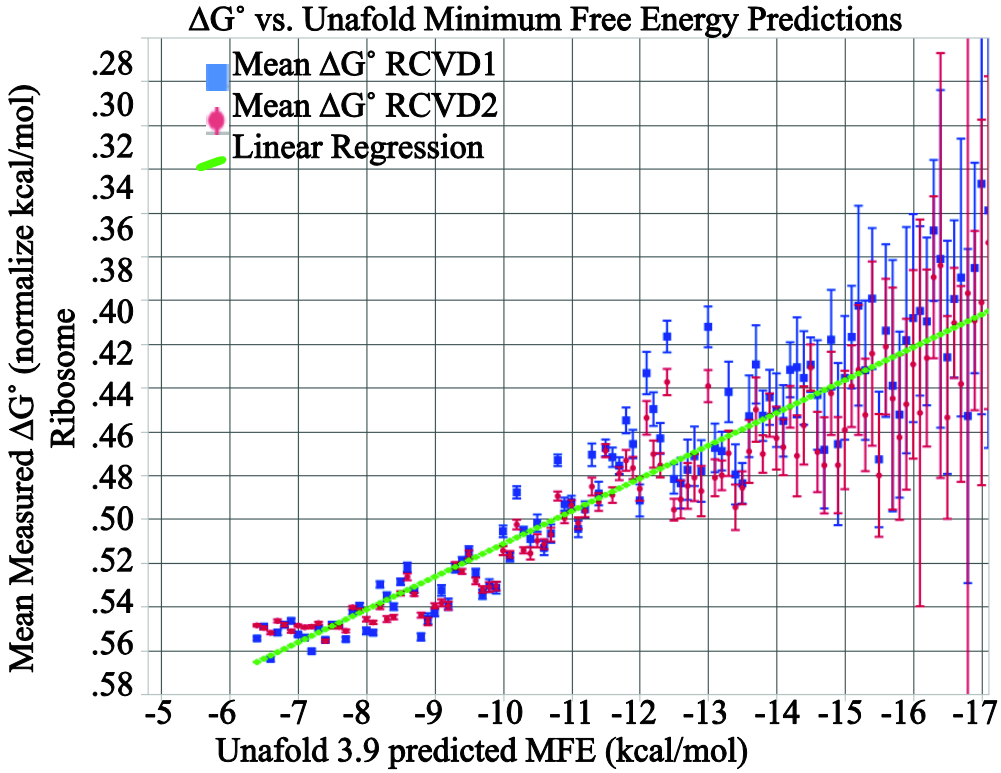

Figure 21. The predicted versus experimental results for ribosome samples RCVD1 and RCVD2 using the UNAfold model. The data included within the optimal linear regression are shown, with the axes set to match corresponding Figure 22. The figure is composed of 124 means of 1,930,428 data points. Each error bar is constructed using a 95% confidence interval of the mean.

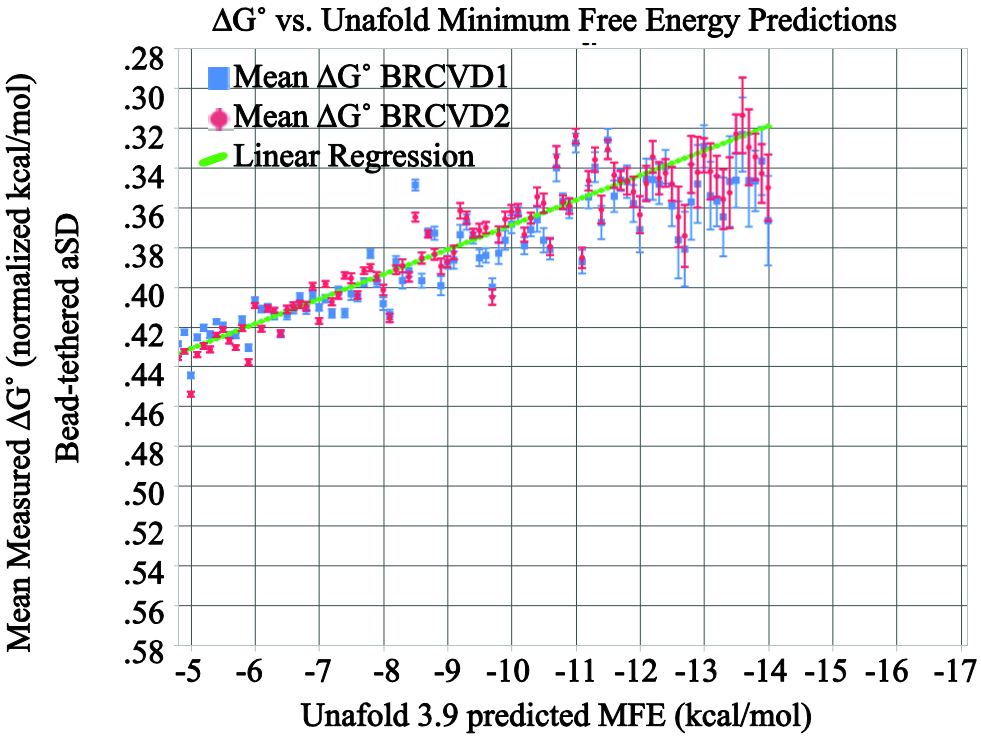

Figure 22. The predicted versus experimental results for bead samples BRCVD1 and BRCVD2 using the UNAfold model. The data included within the optimal linear regression are shown, with the axes set to match corresponding Figure 21. The figure shows the 253 means of 1,209,082 data points with each error bar constructed using a 95% confidence interval of the mean.

In addition to ranking the models by their respective linear regression fitness R^2 (Table 3), we calculated the range of ΔG° over which the linear regression fit.

Table 3. Results from the linear regression optimization for all models versus the experimental results for both ribosome experiments (A) and bead experiments (B) sorted by highest R^2 value, and listing the range over which the linear regression has the best fit.

A)	**Ribosome Combined Samples**			B)	**Nanobead Combined Samples**		
Model	ΔG° start	ΔG° stop	R^2	Model	ΔG° start	ΔG° stop	R^2
free_scan Xia/Mathews	-7.6	-16.9	0.918	Vienna 2007	-4.1	-18.2	0.901
Unafold	-6.3	-17.1	0.896	Unafold	-4.8	-14	0.84
Vienna 2007	-7.7	-10.8	0.883	Vienna 2004	-5.8	-14.9	0.796
free_scan Freier	-7.2	-13.1	0.84	free_scan Freier	-3.2	-14.9	0.768
Vienna 2004	-8.1	-18.9	0.837	free_scan Xia/Mathews	-4.6	-18.2	0.732
BINDIGO	-9.2	-18.3	0.822	MultiRNAfold	-5.9	-12.8	0.715
MultiRNAfold	-9.1	-18.3	0.764	BINDIGO	-5.5	-18.3	0.633

The free_scan model with the Xia 1998 parameters best fit the ribosome binding experiment with a coefficient of determination $R^2 = 0.918$, followed by UNAfold 3.9, ViennaRNA with Andronescu parameters, free_scan Freier, ViennaTurner 2004 parameters, BINDIGO, and MultiRNAfold with R^2 values of 0.896, 0.883, 0.84, 0.837, 0.822 and 0.764, respectively. For predicting the binding to a nanobead-bound 13-mer sequence, the ViennaRNA model with the 2007 parameters fit best with $R^2 = 0.901$. The other models followed slightly behind, with Unafold, ViennaRNA 2004, free_scan Freier, free_scan Mathews, MultiRNAfold and BINDIGO fitting the data with $R^2 = 0.84$, 0.796, 0.768, 0.732, 0.715, and 0.633, respectively. Overall the ViennaRNA 2007 fit best for both situations, followed by UNAfold, free_scan Freier, free_scan Xia, ViennaRNA 2004, BINDIGO, and MultiRNAfold. The model slopes and intercepts were shown in Table 4, along with the percent difference in between their slopes. The Unafold model was the most consistent between experiments with a 16% difference in slopes, while the free_scan Freier model had the largest variation with a 56% difference in slopes. The intercepts of the regression lines

show that the bead experiment had an overall stronger binding to the libraries (.66 to .75 for the ribosome and .47 to .52 for the beads).

Table 4. Regression details for each model, including the slope, intercepts, and percent difference between slopes for each model and experiment.

	Ribosome		Nanobead		%Δ Slopes
Model	slope	intercept	slope	intercept	
Vienna 2004	0.016572	0.691594	0.01057	0.487531	36.21771663
Vienna 2007	0.022067	0.72553	0.010292	0.47565	53.36022114
Unafold	0.014775	0.659263	0.012422	0.493167	15.92554992
free_scan Xia	0.014879	0.660479	0.010855	0.468165	27.04482828
free_scan Freier	0.024697	0.725475	0.010855	0.468165	56.04729319
MultiRNAfold	0.0207998	0.7521873	0.013167	0.523253	36.6965067
BINDIGO	0.015215	0.68709	0.010706	0.490719	29.63522839

Finally, it was noted that while some of the measured ΔG˚ samples equaled the predicted values, most did not as the experimental values are all compressed into a much smaller range. For sample RCVD1 (Figure 18) the mean ΔG° values ranged from -8.0 to -9.4 kcal/mol while the predicted values from the Vienna model were from 0 to -21 kcal/mol. Interestingly, the analysis of how much A_{free} values would need to be reduced by to equal the ΔG° value predicted by a model showed that the results had a log-linear relationship (Figure 23).

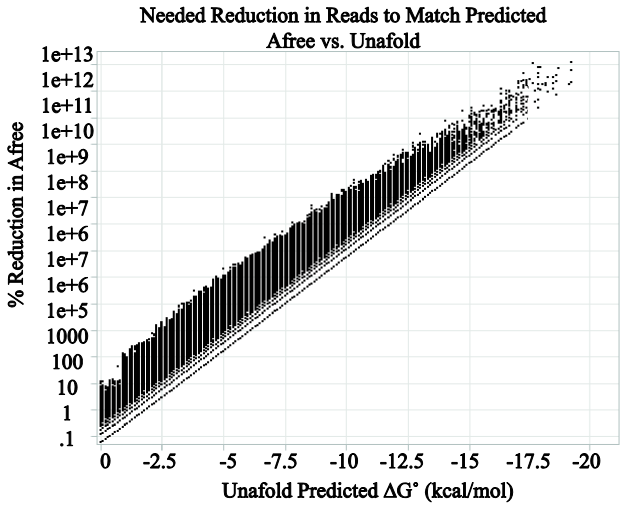

Figure 23. The needed reduction in the number of sequence reads A_{free} needed for the measured ΔG° to equal the predicted ΔG° with a logarithmic y axis.

3.6 Discussion

We have used an RNA sequencing approach to experimentally determine ΔG° values for 13-mer RNA:RNA binding, and to compare MFE predictions from several established models. In general, all of the models tested had a linear region in which stronger binding sequences had larger MFE values. However, there were areas of unexpected differences between the predicted and measured ΔG° values for both low binding energy oligonucleotides (MFE > -7 kcal/mol) and high binding energy oligonucleotides (MFE < -15 kcal/mol). It is likely that the oligomers with low binding energy for the ASD bound to one another in solution in addition to binding to the ASD. Therefore, we could not accurately measure their ΔG° for the ASD accurately, because free oligonucleotide concentrations in solution after the reaction were associated with this non-specific binding. The oligonucleotides with strong binding energies (MFE < -15 kcal/mol) had high variability in MFE, which was likely due to sampling error because of the limited number of sequences.

As a consequence, data for predicted MFE on both ends of the energy spectrum (i.e. > -7 kcal/mol to < -15 kcal/mol) were excluded from calculations.

The number of significant digits for MFE data in a given model output may affect the statistical analysis of model fitness for MFE calculations. ViennaRNA 2004, UNAfold, and free_scan Freier all had output to one tenth of a kcal/mol. Therefore, to compare the models it was necessary to round all MFE predictions to one tenth of a kcal/mol. Interestingly, we found that slopes for the MFE observed and predicted values for the bead and ribosome experiments varied, with the ribosome data having a larger slope for all models (.011 to .013 for bead data compared to .015 to .0 25 for ribosome data). These differences in the slopes varied from 16% for the Unafold model to 56% for the free_scan Freier model and is independent of the MFE value predicted. This difference may be due to physical differences between the ribosomes and beads to cause a change in the binding behavior of the oligomers in solution, such as non-specific binding of oligos to ribosome surfaces (other than the ASD sequence). The differences in slope may indicate preferential binding to the ASD sequence of the ribosome, although the ASD is GC-rich and, in general, may have stronger binding to the non-ASD sequences in the random RNA oligonucleotide pool. It is also possible that parts of the ribosome act to recruit ASD-like RNA sequences as part of its evolved function. The intercepts of the linear regression equations for the predicted versus measured data were different in all cases between the ribosome and nanobead experiments. This difference is most likely a result of the process that normalized maximum and minimum values to allow side-by-side comparisons of the regression parameters. This normalization process shifted the results on the y-axis, but did not affect the R^2 value comparison of fitness.

The mean values of the measured ΔG° were compressed into a smaller range than the predicted values. The average measured ΔG° for the ribosome experiments ranged from -8.0 to -9.4, while the predictions made by the models all range from 0 to around -20 kcal/mol. This discrepancy may be due to the simultaneous nature of the binding of the random oligonucleotide pool. Weak binding oligonucleotides may have served by competition to lower the binding of more strongly binding oligonucleotides. Additionally, the weak binding oligos may appear to have a stronger than predicted ΔG° due to chain binding to the single stranded portions of oligonucleotides that are partially bound to the ASD. To test this hypothesis, the change in concentration of each oligonucleotide sequence needed to equal the predicted value was calculated and plotted (Figure 23). The result showed an exponentially increasing concentration change as the predicted MFE magnitude increases in a log-linear relationship. This indicates that the compression of the measured ΔG° vs. the predicted MFE is due to a systematic bias, such as intermolecular binding.

The evaluation of competition and intermolecular binding using the NUPACK package quantified the predicted effect on the observed ΔG˚ for smaller (100-1280 species) groups of interacting oligonucleotides. As the number of competitors increases, the observed ΔG˚ decreases. For strong binders (ΔG˚ < -10 kcal/mol) the effect is less pronounced, as competitors for the tail that are likely to displace them are much rarer. While the effect of groups of competitors more numerous than a few thousand is computationally intractable due to memory and constraints and processing time (O(N^4) time and $O(N^2)$ storage), extrapolations can be made from smaller groups of competitors. On average stronger binders are less affected by competition than weaker ones, and the effect is systematic such that for a

homogenously distributed group of oligonucleotides of unique sequences competing for a target, normalizing their observed ΔG° will result in their respective ranks as to which binds more strongly, thus competition is not a barrier to establishing which model of RNA binding best predicts these data.

3.7 Conclusions

We have developed a method for high-throughput comparison of nucleic acid binding thermodynamics *in vitro* in the context of both artificial and naturally derived substrates. With measurements consisting of approximately ~10^9 total sequence reads, the binding landscape was elucidated, and the relative accuracy and precision of predictive computer algorithms for determining MFE for 13-mer oligonucleotide-ASD binding was determined. This approach may have utility for determining the MFE of RNA:RNA interactions for a wide array of cellular functions. Our measurements with *E.* coli 30S ribosomal subunits and a nanobead-tethered anti-Shine-Dalgarno sequences, indicated that the ViennaRNA 2.1.8 package using the Andronescu 2007 parameter set had the best precision for their binding kinetics with nanobead experiments, however, with ribosome data the free_scan Freier model had a higher coefficient of determination. By qualifying the accuracy of MFE predictions our findings have improved the utility of the FET model for predicting ribosome-mRNA interactions that are important in maintaining the proper reading frame during protein translation. These results may be useful for improving expression of recombinant proteins.

4 Information Content Analysis of Shine-Dalgarno Binding with Randomized Oligonucleotides

4.1 Abstract

The purpose of this chapter is to elucidate how one might better determine the quantitative or deterministic value of measuring the information content of nucleic acid binding to ribosomes in order to improve the benchmarking methods for RNA binding modeling packages. The focus is on the algorithms and derivations from information content that can be used in future work for analysis on data of this type, and does not contain the results of these analyses. The methods proposed can show correlations between individual sequence information content and measured binding energy. Additionally, correlations between the measured information content and predicted MFE by the RNA binding models can be calculated using these methods, and comparisons between the predicted and measured MFE and ΔG° can show the accuracy of RNA binding models.

4.2 Introduction

Common techniques to identify binding sites in DNA or RNA sequences *in silico* rely on finding the most common nucleotides that make up the consensus sequence, or "motif" [8]. This is a representation that shows the most common base found at each location. This method has shortcomings when attempting to discover new sites because the relative frequency of the other bases at each position is ignored. As Schneider et al [41] noticed as an

example: "…the first position of *Escherichia coli* translational initiation codons has 94% A, 5% G, 1% U, and 0% C, which is not represented precisely by the consensus 'A'".

Schneider goes on to present a method for evaluating the information content of macromolecular binding sites which does not ignore the variability of individual positions in the way that consensus sequences do. The calculation of the information content $R_{sequence}$ is based upon knowing the nucleotide distribution of the genome in general, and the specific distribution of nucleotides for that kind of binding site. If we look at several random locations and determine the difference in the distributions of the nucleotides from their distribution in the entire genome, we would usually see a difference close to 0, indicating there is no information present. However, should we find a difference, that could indicate something interesting at that location. Additionally, if we conduct the same analysis with known binding sites, we may discover new aspects to those sites, such as which regions of the binding sites are more important, and which regions can still work even when different nucleotides are present at some locations. For example, the ribosome binding site (RBS) upstream of the start codon in prokaryotes has a known sequence (the Shine-Dalgarno), yet does not always have to be exactly the Shine-Dalgarno sequence to be perfectly functional for purposes of binding ribosomes to mRNA for translation. By calculating the difference in uncertainty between 149 known RBSs and the *E. coli* genome, a sequence logo can be generated as illustrated in Figure 24.

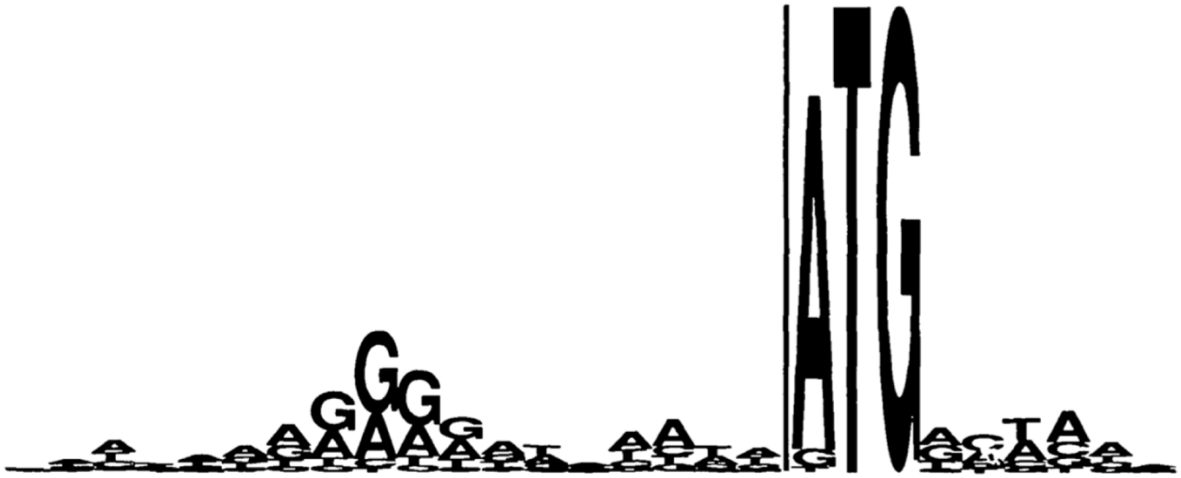

Figure 24. From Schneider et al: "Logo for *E. coli* ribosome binding sites. Only -18 to +8 of the -20 to +13 site is shown. The first translated codon is just to the right of the 2 bit high vertical bar. 149 known binding sites were used to create the logo."

This case relies on the assumption that every location of the genome sequence has an equiprobable nucleotide distribution. In certain cases, for example synthesized libraries, the distribution of nucleotides varies depending on which location one is examining. In such location-dependent skewing, an ad-hoc modification to $R_{sequence}$ was made which redefines the difference in entropy as the *relative entropy* [52], or the *Kullback-Leibler* distance [21] between two distributions. By modifying the calculation for entropy to be position dependent, information theory can be applied to synthesized libraries.

Previous work has analyzed the relationship between free energy and information in protein-DNA interactions with a method to estimate the binding energy of binding sites with a computational focus on alignments of known protein binding sites [54]. That analysis focused on the binding of the Mnt protein of *Salmonella* phage P22 to variations of its 17-base site [20], showing how changes in the sequence affect its binding affinity. They estimated the collection of predicted ΔG°s for every nucleotide switch that could occur at

each location along the binding site based on samples of known sites, and calculated the new binding site by summing the ΔG° for the substituted nucleotides. They made some assumptions, namely that ΔG°s are additive across positions, and that the starting entropy is approximated as a random sequence.

Neither of those situations is true in our case: we know that the rules for calculating ΔG° for our models are not additive and we must include Kullback-Leibler divergence as part of the calculation for location-dependent changes in entropy. We also are in the unique position of being able to directly assess the ΔG° of the variants in question and immediately compare them to predictions of ΔG° and to their relative entropy with respect to nucleotide distributions. Ultimately this method can enhance our conceptual understanding of, and thus our ability to improve, RNA binding.

4.3 Approach

As an example of how one might go about this kind of analysis, it is convenient to take the data set presented in this dissertation. In this case we would start with a large (k = ~8 million) library of RNA oligonucleotides interacting with the ribosome. Using next-generation sequencing we can determine the equilibrium concentrations of each oligonucleotide sequence before and after binding, and thus can calculate the ΔG°. Additionally, the sequencing process reveals the change in nucleotide probability from before and after the binding event. From this change in probability, we can calculate the information content. We can then compare the information content with the measured energy of binding and with minimum free energies predicted by various computer models.

4.4 Methods

The experimental methods shown here are what was used to gather a data set of the type being proposed for future information content analysis. Other methods of sequence binding data are available, but in this case a library of randomized 13-mer ssRNA oligonucleotides, called RAND, were synthesized (IDT) and bound to 30S *E. coli* subunits as described in the methods of Chapter 3 of this dissertation. The filtrate was collected and the concentration of each sequence before and after binding was determined by Illumina Hi-Seq 2500 small RNA next-generation sequencing.

A new set of randomized oligos named RAND2 was used in a magnetic bead experiment where the anti-Shine-Dalgarno (ASD) sequence (5' – AUUCCUCCACUAG – 3') was synthesized with a biotin molecule present on the 5' end and bound to magnetic streptavidin nanobeads. The oligos were bound to the ASD-bead complex, separated by magnetic field, and the filtrate collected and sequenced.

4.5 Data Analysis to be Conducted

The first step in analyzing the sequence data would be to determine the distribution of nucleotides at each location along the 13-mer. This involves counting the number of adenine, cytosine, guanine, and uracil acids present at locations 1 through 13 for each sample before and after the binding. What results is a position-specific scoring matrix (PSSM) of dimension 13x4 as illustrated below. The four rows represent the nucleic acids AUCG, and the columns are the distribution of the nucleotide at that position.

Table 5. The PSSM for sample RAND2, showing the distribution of each nucleotide at each location along the 13-mers present in the sample.

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0.26	0.24	0.22	0.21	0.21	0.19	0.19	0.20	0.20	0.19	0.25	0.21	0
U	0.28	0.28	0.27	0.25	0.26	0.30	0.31	0.30	0.25	0.18	0.10	0.19	0
C	0.15	0.12	0.12	0.12	0.13	0.13	0.14	0.13	0.12	0.11	0.08	0.18	1
G	0.32	0.36	0.39	0.42	0.41	0.38	0.36	0.37	0.43	0.52	0.56	0.43	0

Since the last nucleotide on the 3' end is always C, it (location 13) is generally ignored and not mentioned.

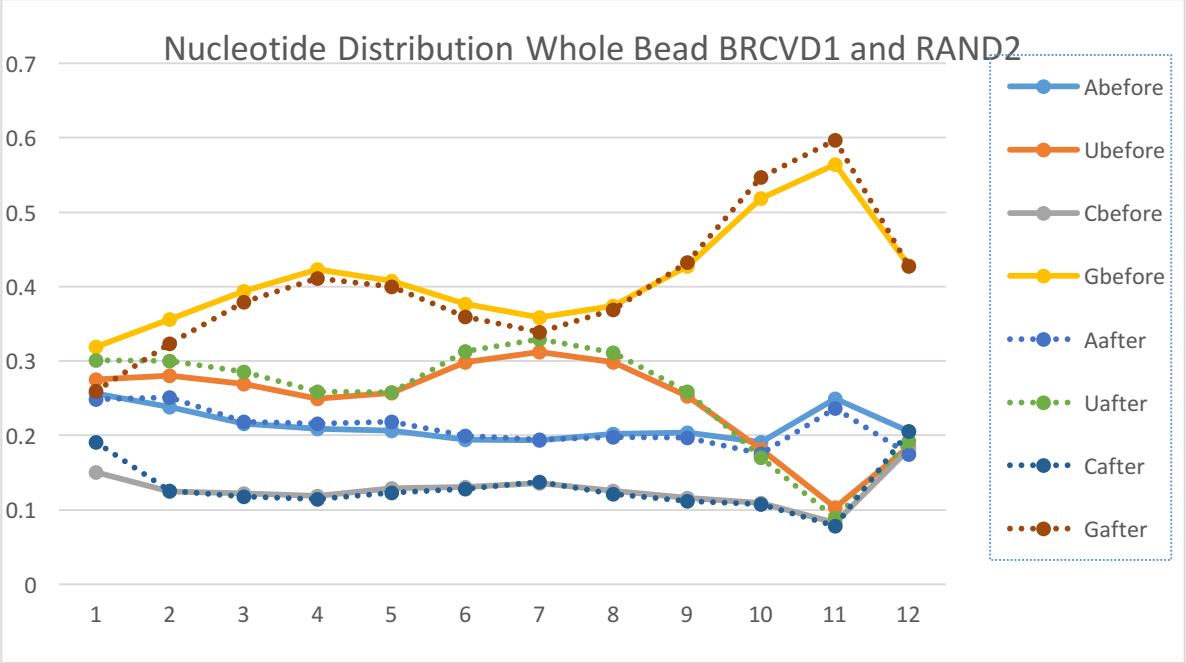

Figure 25. The distribution of nucleotides of the sequences that were in the random starting set (RAND2), and the filtrate of the sequences that did *not* bind, BRCVD1. Determining the distribution of those that did bind requires an additional step.

The fact that there are changes in the distribution of nucleotides, sometimes as high as 5% increase for G, indicates that selection has occurred that is selecting for the G nucleotide (Figure 25). This requires energy as specified by the second law of thermodynamics, and is probably related to the likelihood that oligonucleotides with certain sequences are more likely to bind than others. This specificity may be related to the Gibbs free energy of binding, ΔG°. It is hypothesized here that there exists a correlation between the information content and the ΔG° of our samples, and that if this correlation exists, measuring the strength of correlation between the measured information content and the predicted MFE by some computational model could provide an additional benchmark upon which to measure the reliability of that model.

The probability distribution of nucleotides before binding, is specified by a PSSM matrix

$$(1)\ \text{pRAND} = \begin{bmatrix} pA_b \\ pU_b \\ pC_b \\ pG_b \end{bmatrix}$$

pA_b, pU_b, pC_b, and pG_b denote probability vectors of length 13. The total number of oligos is denoted T_b, and the distribution of nucleotides that did not bind after is

$$(2)\ \text{pRCVD1} = \begin{bmatrix} pA_a \\ pU_a \\ pC_a \\ pG_a \end{bmatrix}$$

From these matrices we can calculate the number of oligos that bound since we know the number of target molecules T_r (ribosomes or beads). We also know the total number of each nucleotide before and after binding: T_{nRAND}, T_{nRCVD1}.

(3) $T_{nRAND} = T_b$ pRAND

(4) TnRCVD1 = Ta pRCVD1

The nucleotide distribution of those oligos that *are* bound, p_{Bound}, could then be calculated by determining how many of each nucleotide bound by subtracting from the free unbound oligos then dividing by the number of ribosomes. We added 209 pmol of oligonucleotides to 95 pmol to ribosome, so we would have 114 oligonucleotides that did not bind if every single ribosome bound exactly one oligo:

(5) $T_{nbound} = T_{nrand} - T_{nrcvd1}$

(6) pBound = T_{nbound}/T_r

This works only when the assumption that each target (ribosome or bead) binds exactly one oligonucleotide. However, we know from the results of Chapter 2 that there exists a phenomenon where the weak binders seem to bind more often than is predicted. The result is that the 1:1 binding ratio assumption is wrong, and indeed use of equation (6) may sometimes result in distributions of nucleotides that are impossible (greater than 1 or less than 0).

To estimate the true target-to-oligo binding ratio, a prediction is made by extending the linear range of binding and calculating the average difference in the number of oligos actually bound from the number that theoretically would have bound for that model.

To do this we would determine the observed ΔG° by calculating the association constant, K_i for each sequence, *i*:

(7) $$K_i = \frac{[A_i]_{total} - [A_i]_{free}}{[A_i]_{free}([B]_{total} - \sum_{i=1}^{n}[A_i]_{total} - [A_i]_{free})}$$

where $[A_i]_{total}$, $[A_i]_{free}$, and $[B]_{total}$ are the concentrations of the oligo with sequence *i* before binding (samples RAND and RAND2), oligos that did not bind (RCVD and BRCVD samples), and the concentrations of the ribosomes (or beads) respectively, for a total number of sequences, *n*.

To calculate what the model predicted the concentration $[A_i]_{free}$ must be in order to result in that model's predicted MFE, we would substitute the model's prediction of K_i and solve for $[A_i]_{free}$ from equation **Error! Reference source not found.**:

$$[A_i]_{free-predicted} = [A_i]_{total} \frac{[A_i]_{total}}{e^{\frac{-\Delta G^\circ}{R^* T}} * \{[B] - [A]_{total} - [A]_{free} + 1\}} \frac{n}{[A]_{free}}$$

Note that the absence of the subscript *i* for a concentration indicates the total concentration of RNA molecules in that sample, not the concentration of any individual oligo sequence. To find out how many oligos per ribosome, we would calculate:

$$OligosPerRibosome_i = \frac{[A_i]_{free-predicted}}{[A_i]_{free}} \tag{8}$$

Now the average oligos bound per ribosome or bead (for a given model) would be

$$\frac{\sum_{i=0}^{n} OligosPerRibosome_i}{n} \tag{9}$$

Each model can be analyzed in this manner, and since we are not certain of the accuracy of any one model, the mean of the seven models would be used. Substituting the new oligo per ribosome ratio into equation (4), the distributions of each sequence bound to the targets can be estimated, and from there a more accurate distribution can be used to

calculate the mutual information content. The distribution of nucleotides in the linear range of the samples before and after would be used in the equations

$$R = -\Delta H = -[(H_{before} + e_{nbefore}) - (H_{after} + e_{nafter})] \quad (10)$$

where

$$H = \sum_{i=A,U,C,G} p_i \times \log_2(p_i) \quad (11)$$

where p_i is the distribution of nucleotide i, and the approximation for the small sample correction is

$$e_n = \frac{1}{\ln(2)} \times \frac{s-1}{2n}. \quad (12)$$

The number n is in the millions for our samples, so the error is negligibly small.

In equation above, there is an assumption that p_i is the same for each nucleotide regardless of the location along the oligonucleotide. This is frequently used for analyzing sequences in organisms where the total nucleotide distribution of the genome is known. However, in our case we know that the distribution of each nucleotide varies depending on the location 0 through 12 of the oligonucleotide. Thus, we would modify equation (10) for each location j, 0 through 12:

$$H_{j-before} = \sum_{i=A,U,C,G} p_{i-before} * \log_2(p_{i-before})$$
$$H_{j-after} = \sum_{i=A,U,C,G} p_{i-after} * \log_2(p_{i-after}) \quad (13)$$

with the information content for each location j equal to

$$R_j = -\Delta H = -[H_{j-before} - H_{j-after}]$$

The result is the amount of information in bits at each location, which can be represented as a sequence logo.

An additional analysis can be made by calculating the information content contributed by every unique sequence. In this case, the summation is not made over every nucleotide distribution, but rather given a certain sequence, use the distributions of that sequence to calculate the individual mutual information magnitude:

$$R_{sequence} = \sum_{k=0}^{12} \left| p_{k-before} * \log_2(p_{k-before}) - p_{k-after} * \log_2(p_{k-after}) \right| \quad (14)$$

4.6 Discussion

4.6.1 Sequence Logo:

For comparing sequence logos between various models, it can be difficult to quantify which is more "correct". The logo may be obfuscated by the large proportion of weak-binding oligos having bound to the target. Rather what is possible is to qualitatively describe the information content for the groups of sequences in the linear ranges of those models as appearing more a less similar to the compliment, or a section of the compliment of the tail sequence.

A short region of complementarity may bring up another confounding phenomenon when analyzing sequence logos: that oligos are allowed to bind to their target at any alignment. For example, if you have oligo sequences containing the subsequence CAUGCAUG flanked by some arbitrary base •.

Table 6. A hypothetical example of subsequence alignment obfuscating information content. The subsequence CAUGCAUG is present at different positions within the oligo, and flanked by arbitrary bases denoted by •

•	•	•	•	C	A	U	G	C	A	U	G	•	•	•	•
•	•	•	•	•	C	A	U	G	C	A	U	G	•	•	•
•	•	•	•	•	•	C	A	U	G	C	A	U	G	•	•
•	•	•	•	•	•	•	C	A	U	G	C	A	U	G	•
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

Suppose that the four oligo sequences in Table 5 above is determined to have bound to a target in equal proportion. Note that every oligo contains the subsequence CAUGCAUG however this sequence is present at different locations within the oligo. Thus, in this hypothetical example the information gained from locations 8 through 12 above would be 0, since there is an even distribution of each base bound.

Finally, the amount of information that can be gained can be reduced by a skewed starting distribution. Put in simple terms, there is no room to travel if we are already at the destination. How much room was taken out of the way, however can be calculated as the amount of potential information "lost". This is the starting entropy of the system that we know, since it is not uniformly randomized, is not the maximum entropy. We can calculate the difference between the starting entropy of the system H_{before} and the maximum entropy. Any number of bits above or below 0.5 indicates the amount of information that could potentially have been detected, if the starting distribution were not skewed.

Conclusions

Analysis of the information content data could reveal what can be quantitatively determined with regard to the effectiveness of the various RNA binding models. A correlation between the mutual information of individual sequences and the measured ΔG° could be found and shown to be a useful tool for RNA energetic analysis. However, skewed starting distributions and obfuscation by low-energy sequences binding more often than expected introducing noise all contribute to difficulties in assessing the usefulness of information content analysis. If the synthesized nucleotide pool were more randomly distributed it could reduce uncertainty in the information content. Additionally, there may be experimental measures that could improve the efficacy of an analysis. For example, implementing a method to remove the weakly bound sequences that are attached via the "chain" phenomenon described in Chapter 3 the dissertation could greatly improve the resolution. An example could be a step-wise chelation and melting whereby magnesium ions are chelated with increasing concentrations of EDTA and the reaction is heated by small increments while samples are taken for sequencing. For a method like that one would expect to see libraries that start off looking like the post-binding samples we created, RCVD and BRCVD. As the dissociation conditions become more favorable we would begin to see libraries that look like the inverse, where we are sequencing only oligos that did bind. These could be correlated with the melting temperatures as an additional measure of ΔG° in conjunction with the binding ratios. Removal of the weak binder noise could paint an even clearer picture of the information content. This method of analysis in conjunction with the

novel methods described in this dissertation all pave the way to a better understanding how RNA works from fundamental principles.

REFERENCES

1. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinf* 23(13):i19-i28. doi: 10.1093/bioinformatics/btm223.
2. Andronescu M, Condon A, Turner DH, Mathews DH (2014) The determination of RNA folding nearest neighbor parameters. *Methods Mol Bio* 1097:45-70. doi: 10.1007/978-1-62703-709-9_3.
3. Andronescu M, Zhang ZC, Condon A (2005) Secondary structure prediction of interacting RNA molecules. *J Mol Biol* 345(5):987-1001. doi: 10.1016/j.jmb.2004.10.082.
4. Blankenberg D, Gordon A, Kuster Von, G, Coraor N, Taylor J, Nekrutenko, A, Galaxy Team. (2010). Manipulation of FASTQ data with Galaxy. *Bioinformatics*, *26*(14), 1783–1785. doi:10.1093/bioinformatics/btq281
5. J. D. Buenrostro, C. L. Araya, L. M. Chircus, and C. J. Layton, "Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes : Nature Biotechnology : Nature Research," *Nature*, 2014.
6. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucl Acids Res* 38(6):1767-1771. doi: 10.1093/nar/gkp1137.
7. Cuddy K, Foley B, Jaffe JS, & Gillespie D (1993). RT-PCR with affinity-captured mRNA. *Nucleic Acids Research*.
8. Davidson EH, Jacobs HT, and Britten RJ, "Eukaryotic gene expression: Very short repeats and coordinate induction of genes," *Nature*, vol. 301, pp. 468–470, Feb. 1983.).
9. C. B. Do, D. A. Woods, and S. Batzoglou, "CONTRAfold: RNA secondary structure prediction without physics-based models.," *Bioinformatics*, vol. 22, no. 14, pp. e90–8, Jul. 2006.
10. Fink TR, Crothers DM (1972) Free energy of imperfect nucleic acid helices. I. The bulge defect. *J Mol Biol* 66(1):1-12.
11. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, & Turner DH (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(24), 9373–9377.
12. Fresco JR, Klemperer E (1959) Polyriboadenylic acid, a molecular analogue of ribonucleic acid and desoxyribonucleic acid. *Annl New York Acad Sci* 81:730-741.
13. Freitag S, Le Trong I, Klumb L, Stayton PS, Stenkamp R. E. (1997). Structural studies of the streptavidin binding loop. *Protein Sci.* 6 p.1157
14. M. Geertz, D. Shore, and S. J. Maerkl, "Massively parallel measurements of molecular interaction kinetics on a microfluidic platform.," *Proceedings of the National Academy of Sciences*, vol. 109, no. 41, pp. 16540–16545, Oct. 2012.
15. S. Ghosh and S. Joseph, "Nonbridging phosphate oxygens in 16S rRNA important for 30S subunit assembly and association with the 50S ribosomal subunit.," *RNA*, vol. 11, no. 5, pp. 657–667, May 2005.

16. Giese MR, Betschart K, Dale T, Riley CK, Rowan C, Sprouse KJ, Serra MJ (1998) Stability of RNA hairpins closed by wobble base pairs. *Biochem* 37(4):1094-1100. doi: 10.1021/bi972050v
17. Groebe DR, Uhlenbeck OC (1988) Characterization of RNA hairpin loop stability. *Nucl Acids Res* 16(24):11725-11735.
18. He M, & Taussig MJ (1997). Antibody-ribosome-mRNA (ARM) complexes as efficient selection particles for in vitro display and evolution of antibody combining sites. *Nucleic Acids Research*, *25*(24), 5132–5134.
19. Hodas NO, Aalberts DP (2004) Efficient computation of optimal oligo-RNA binding. *Nucl Acids Res* 32(22):6636-6642. doi: 10.1093/nar/gkh1008.
20. Knight KL and Sauer RT, "Biochemical and genetic analysis of operator contacts made by residues within the beta-sheet DNA binding motif of Mnt repressor.," *EMBO J.*, vol. 11, no. 1, pp. 215–223, Jan. 1992.
21. S. Kullback and R. A. Leibler, 1951 "On information and sufficiency," The annals of mathematical statistics, 79-86.
22. Li G-W, Oh E, Weissman, JS (2012) The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nat* 484(7395):538-541. doi: 10.1038/nature10965
23. Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26.
24. Makhno VI, Peshin, NN, Semenkov YP and Kirillov SV (1988). Modified method of producing "tight" 70S ribosomes from *Escherichia coli*, highly active in individual stages of the elongation cycle, *Mol. Biol. 22,* 528-537.
25. Markham NR, Zuker M (2008) UNAfold: Software for nucleic acid folding and hybridization. *Methods Mol Biol (Clifton, N.J.)* 453:3-31. doi: 10.1007/978-1-60327-429-6_1.
26. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA* 101(19):7287–7292. doi: 10.1073/pnas.0401799101.
27. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911-940. doi: 10.1006/jmbi.1999.2700.
28. D. H. Mathews, W. N. Moss, and D. H. Turner, "Folding and finding RNA secondary structure.," *Cold Spring Harb Perspect Biol*, vol. 2, no. 12, p. a003665, Dec. 2010.
29. May EE, Vouk MA, Bitzer DL, Rosnick DI (2004) An error-correcting code framework for genetic sequence analysis. *J Franklin Inst 341*(1):89-109.
30. May EE, Vouk MA, Bitzer DL (2006) Classification of *Escherichia coli* K-12 ribosome binding sites. An error-control coding model. *IEEE Eng in Med Biol Mag: Q Mag Eng Med Biol Soc* 25(1):90-97.
31. Mishra M, Vu SK, Bitzer D L, Vouk MA (2004) Free energy periodicity in prokaryotic coding and its role in identification of +1 ribosomal frameshifting in the *Escherichia Coli*

K-12 gene prfb. *Conf Proc: Annu Int Conf IEEE Eng Med Biol Soc* 4:2848-2851. doi: 10.1109/IEMBS.2004.1403812.
32. Nutiu R, Friedman R.C, Luo S, Khrebtukova I, Silva D, Li R, Zhang L, Schroth GP, Burge CB (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotech*, 29(7),659–664.

33. Philippe C, Bénard L, Eyermann F, Cachia C, Kirillov SV, Portier C, Ehresmann B, et al. (1994). Structural elements of rps0 mRNA involved in the modulation of translational initiation and regulation of E. coli ribosomal protein S15. *Nucleic Acids Research*, *22*(13), 2538–2546.
34. Ponnala L, Barnes TM, Bitzer DL (2004) Ribosome tail ends as "signal detectors" for protein production in prokaryotes. *Conf Proc: Symp Biotechnol Bioinf* pp 15-23. doi: 10.1109/SBB.2004.1364361
35. Ponnala L, Bitzer DL, Stomp A, Vouk MA (2006) A computational model for reading frame maintenance. *Conf Proc: Annu Int Conf IEEE Eng Med Biol* 1:4540-4543. doi: 10.1109/IEMBS.2006.259717
36. Ponnala L, Stomp A, Bitzer DL, Vouk MA (2006) Statistical significance and biological relevance of a sinusoidal pattern detected in translational free energy signals. *Genomic Signal Proc Stat, GENSIPS'06. IEEE Int Workshop* pp 55-56.
37. Ponnala, L, Stomp AM, Bitzer DL, & Vouk MA (2006). Analysis of Free Energy Signals Arising from Nucleotide Hybridization Between rRNA and mRNA Sequences during Translation in Eubacteria. *EURASIP Journal on Bioinformatics and Systems Biology*, *2006*, 1–9. doi:10.1155/BSB/2006/23613
38. Ponnala, L, Bitzer D, & Stomp A (2006). A computational model for reading frame maintenance. *Engineering in Medicine and Biology*
39. Rheinberger HJ (1991) The function of the translating ribosome: allosteric three-site model of elongation. *Biochimie*, vol. 73, no. 7, pp. 1067-1088.
40. Rosnick DI, Bitzer DL, Vouk MA, May EE (2000) Free energy periodicity in *E. coli* coding. In *Eng Med Biol Soc, Proc 22nd Annu Int Conf IEEE* 4:2470-2473.
41. Schneider TD, Stormo GD, Gold L, and Ehrenfeucht A, "Information content of binding sites on nucleotide sequences," *Journal of Molecular Biology*, vol. 188, no. 3, pp. 415–431, Apr. 1986.
42. Schneider TD and Stephens RM, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, 1990.
43. Schneider TD, "Evolution of biological information.," *Nucleic Acids Research*, vol. 28, no. 14, pp. 2794–2799, Jul. 2000.
44. Schroeder SJ, Turner DH (2009) Optical melting measurements of nucleic acid thermodynamics. *Methods Enzymol* 468: 371-387. doi: 10.1016/S0076-6879(09)68017-4
45. Serra MJ, Lyttle MH, Axenson TJ, Schadt CA, Turner DH (1993) RNA hairpin loop stability depends on closing base pair. *Nucl Acids Res* 21(16):3845-3849.
46. Shine J, Dalgarno L (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA* 71:1342-1346.

47. P. Stanssens, E. Remaut, and W. Fiers, “Alterations upstream from the Shine-Dalgarno region and their effect on bacterial gene expression.,” *Gene*, vol. 36, no. 3, pp. 211–223, 1985.
48. Starmer J, Stomp A, Vouk M, Bitzer D (2005) Predicting fidelity of protein synthesis in *E. coli*. *Proc IEEE Int Workshop on Genomic Signal Proc Stat GENSIPS 2005*, May 22 - 24, New Port, Rhode Island (CD-ROM).
49. Starmer J, Stomp A, Vouk M, Bitzer D (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol* 2(5):e57:0455-0466. doi: 10.1371/journal.pcbi.0020057.
50. Starmer, JM (2006). What can RNA hybrids tell us about translation? (Doctoral dissertation). *Retrieved from NCSU Institutional Repository http://www.lib.ncsu.edu/resolver/1840.16/3369*. 2010-04-02T18:29:27Z
51. Starmer J, Stomp A, Vouk M, Bitzer D (2005) “Predicting fidelity of protein synthesis in E. coli,” Proc. IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS 2005)
52. Starmer J, A.-M. Stomp, Vouk MA, and Bitzer DL, “Predicting Shine-Dalgarno sequence locations exposes genome annotation errors,” PLoS Computational Biology, vol. 2, no. 5, pp. 454–466, 2006.
53. Stormo GD, Information content and free energy in DNA–protein interactions. 1998 *Journal of Theoretical Biology*, 195, 135-137
54. Stormo GD and Fields DS, “Specificity, free energy and information content in protein-DNA interactions,” *Trends in Biochemical Sciences*, vol. 23, no. 3, pp. 109–113, Mar. 1998.
55. Vu SK, Belloti AA, Gabriel CJ, Brochu HN, Miller ES, Bitzer DL,Vouk MA (2014) Modeling ribosome dynamics to optimize heterologous protein production in *Escherichia* coli. *Signal Inf. Process GlobalSIP, IEEE Global conf.* pp. 1422-1425.
56. J. D. Watson and F. Crick, “Molecular structure of nucleic acids,” *Nature*, 1953.
57. Will S, Reiche, K, Hofacker IL, Stadler PF, & Backofen R (2007). Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. *PLoS Computational Biology*, *3*(4), e65. doi:10.1371/journal.pcbi.0030065
58. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochem*, *37*(42):14719-14735. doi: 10.1021/bi9809425.
59. Xia T, McDowell JA, Turner DH (1997) Thermodynamics of nonsymmetric tandem mismatches adjacent to GC base pairs in RNA. *Biochem* 36(41):12486-12497. doi: 10.1021/bi971069v.
60. Xing, C., Bitzer, D. L., Alexander, W. E., Stomp, A. M., & Vouk, M. A. (2006). Free energy analysis on the coding region of the individual genes of Saccharomyces cerevisiae. *Conference proceedings : ... Annl Intn.l Conf. IEEE Eng. Med. Bio. Soc.*, *1*, 4225–4228. doi:10.1109/IEMBS.2006.259972

61. Zadeh JN, Steenber CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA (2011) NUPACK: Analysis and Design of Nucleic Acid Systems. *J Comput Chem* 32: 170-173. doi: 10.1002/jcc.21596
62. PAC, 1996, 68, 149. A glossary of terms used in chemical kinetics, including reaction dynamics (IUPAC Recommendations 1996), doi: 10.1351/pac199668010149
63. M. Zuker, D. H. Mathews & D. H. Turner. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide (1999) *RNA Biochemistry and Biotechnology*, 11-43, J. Barciszewski and B. F. C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, Dordrecht, NL
64. Smith, Temple F. & Waterman, Michael S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology* 147: 195–197. PMID 7265238. doi:10.1016/0022-2836(81)90087-5
65. Schroeder SJ, Turner DH (2009) Optical melting measurements of nucleic acid thermodynamic. *Meth Enzymol* 468: 372-387. doi: 10.1016/S0076-6879(09)68017-4
66. Clanton-Arrowood, K., McGurk, J., and Schroeder, S. J. (2008). 3’-Terminal nucleotides determine thermodynamic stabilities of mismatches at the ends of RNA helices. *Biochemistry* 47, 13418–13427
67. Turner, D. H. (2000). Conformational changes. In ‘‘Nucleic Acids: Structures, Properties, and Functions,’’ (V. A. Bloomfield, D. M. Crothers, and I. Tinoco Jr., eds.) pp. 259–334. University Science Books, Sausalito, CA
68. C. E. Shannon (1948). A mathematical theory of communication. *Bell Syst Technol J* 27: 379–423.
69. Spicer E, Schwarzbauer J, Craven GR (1977). Isolation of ribosomal protein-RNA complexes by nitrocellulose membrane filtration: equilibrium binding studies. *Nucleic Acids Research*
70. Guérin MF, Hayes DH (1987). Comparison of active and inactive forms of the *E. Coli* 30S ribosomal subunits. *Biochimie* vol. 69:9, pp 965-974. doi 10.1016/0300-9084(87)90230-6
71. A. P. Carter, W. M. Clemons, and D. E. Brodersen, “Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics,” *Nature*, 2000.