THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF CHEMICAL ENGINEERING

# DETERMINING THE RIBOSOME BINDING FOOTPRINT IN ESCHERICHIA COLI TO IMPROVE THE RIBOSOME BINDING SITE CALCULATOR MODEL OF TRANSLATION INITIATION

ASHLEE SMITH
FALL 2014

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Chemical Engineering
with honors in Chemical Engineering

Reviewed and approved* by the following:

Howard M. Salis
Assistant Professor of Chemical Engineering and Assistant Professor of Agricultural and Biological Engineering
Thesis Supervisor

Themis Matsoukas
Professor of Chemical Engineering
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

# ABSTRACT

In the last 20 years, the technology used to create synthetic DNA has greatly advanced. Today, DNA can now be synthesized 3 times as quickly and costs 5 times less. The availability of recombinant DNA systems provides flexibility for researchers in academia and industry trying to solve metabolic engineering problems and elucidate biological functions. One common metabolic engineering problem is developing a system to control the expression of a protein of interest over a range from very little to very large expression. Instead of generating large combinatorial libraries to find a sequence with a specific function, which is time consuming and inefficient, researchers are rationally altering the DNA sequence to achieve a desired attribute using information gained from biophysical modeling.

One such alteration is the design of synthetic ribosome binding sites to improve the rate of translation initiation in order to increase the expression of protein. The Ribosome Binding Site Calculator (RBS) is software that utilizes a biophysical model of translation initiation to allow researchers to predict the translation initiation rate of a particular sequence or generate a synthetic RBS with a specified translation initiation rate. There are many interactions that take place during translation initiation. In order to improve the accuracy of the model's predictions, one such interaction that we studied is the ribosomal footprint. The ribosome must unfold a section of messenger RNA (mRNA) after the start codon and load it into an mRNA channel. This channel can only accommodate single stranded mRNA. The number of nucleotides in the coding section the ribosome must unfold and load into the channel is the ribosomal footprint (FP). We hypothesized that a secondary structure contained in the ribosomal footprint must be unfolded and would greatly decrease protein expression, and that a secondary structure outside of

the footprint region would not have to be unfolded and protein expression would be higher. We identified the ribosomal footprint for *Escherichia coli* by creating a high expressing RBS sequence and a modified coding section appended to the red fluorescent protein (RFP) reporter gene. The modified coding section featured a movable hairpin structure that was moved from immediately after ATG (position 3) to position 39 to show the changes in protein expression. We found that translation initiation gradually increased from very low levels when the hairpin was close to the start codon to over 1000-fold higher after the hairpin had passed the footprint region. The footprint was found to be 13 or 14 nucleotides including the start codon. Discrepancies between the model predictions and the data revealed the free energy the ribosome must expend to load the mRNA into the channel. The average error for constructs with a hairpin located 3 to 10 nucleotides into the coding sequence was determined to be 7.7 kcal/mol. This value represents the free energy cost of secondary structure within the footprint region, $\Delta G_{load}$. This information improves the biophysical model by determining the free energy change of loading the mRNA into the channel and will provide researchers with more accurate predictions of translation initiation.

# TABLE OF CONTENTS

List of Figures ..... iv

List of Tables ..... v

Acknowledgements ..... vi

Chapter 1: Introduction ..... 1

- Translation Initiation ..... 3
- The Ribosome Binding Site Calculator ..... 5
- The Ribosomal Footprint ..... 8

Chapter 2: Design ..... 11

- Standby Site Hairpin ..... 12
- Shine-Dalgarno Sequence ..... 14
- Restriction Enzymes and Spacing ..... 15
- Coding Section Hairpin ..... 16
- Codon Usage ..... 18
- Final Constructs ..... 20

Chapter 3: Experimental Methods ..... 24

- Plasmid Construction ..... 24
- Storing the Plasmids ..... 25
- Growth and Fluorescence Measurements ..... 26

Chapter 4: Results ..... 27

- Downstream structures greatly affect translation initiation ..... 27
- Model predictions for different footprint assumptions ..... 28
- Ribosomal footprint unfolds 13 or 14 nucleotides downstream of the start codon ..... 29
- The 30S ribosomal subunit follows thermodynamic minimization ..... 31
- A correct ribosomal footprint improves the model's predictions ..... 33
- Ribosome requires additional free energy input to accommodate hairpins in the footprint ..... 36

Chapter 5: Discussion and Conclusions ..... 38

Appendix A Detailed Design Information and Considerations ..... 40

- Codon Usage in *E. coli* ..... 40
- Table of Restriction Enzymes ..... 40
- pFTVP vector sequence with hairpin at position +3 ..... 41

REFERENCES ..... 43

# LIST OF FIGURES

Figure 1: Structure of mRNA Channel in 30S Ribosome ..... 8

Figure 2: RBS and CDS Design for a Varying Footprint Length ..... 11

Figure 3: Ribosome Binding Site Hairpin (Gruber et al., 2008) ..... 13

Figure 4: Coding Section Hairpin (Gruber *et al*., 2008) ..... 16

Figure 5: Codon Usage in Leading Sequence (Footprint Length 15) ..... 19

Figure 6: Final Design of RBS ..... 22

Figure 7: Minimum Free Energy Structure of Final Construct 12 with RBS and CDS Hairpins (Gruber *et al*., 2008) ..... 23

Figure 8: pFTVP Vector Expression System ..... 25

Figure 9: Fluorescence Data for a Movable Hairpin at Varying Distances Away from the Start Codon ..... 27

Figure 10: Comparison of Model Predictions and Experimental Data ..... 28

Figure 11: Average Error of Model Predictions over Varying Footprint Lengths ..... 30

Figure 12: Comparison of Fluorescence and Model Predictions for Variation of Base Pairs at the Bottom of the CDS Hairpin ..... 32

Figure 13: Linear Relationship between the Predicted TIR at a Footprint of 13and the Fluorescence Data ..... 33

Figure 14: Linear Relationship between Model Predictions and Fluorescence Data for a Footprint of 14 ..... 34

Figure 15: Nonlinear Relationship between the Model Predictions and Fluorescence Data at a Footprint of 35 ..... 34

Figure 16: Disparity between the Model Predictions and the Data for Coding Section Hairpins within the Footprint Region ..... 36

Figure 17: Codon Usage in *E. coli* ..... 40

# LIST OF TABLES

Table 1: Included Codons ..... 19

Table 2: Final RBS and CDS Sequences ..... 21

Table 3: Used Restriction Enzymes and Recognition Sequences ..... 40

# ACKNOWLEDGEMENTS

I would like to thank my Thesis Supervisor, Dr. Howard M. Salis, for sponsoring me for the completion of my thesis in his laboratory and giving me encouragement and instruction along the way. Thank you to graduate students Iman Farasat and Amin Espah Borujeni. Iman devoted his time and energy to train me in synthetic biology and guide me through completing this research. Amin duplicated the data and designed additional sequences providing a more complete view of the mechanism. He also gave me great advice in the writing of this thesis and provided the data shown in Figures 9-16 from his continuation of my research. I would also like to thank former undergraduate researcher Natasha Lundgren for her previous work troubleshooting designing the mRNA constructs and determining that rare codons should be avoided.

Additionally, I would like to extend a thank you to our other lab members, Manish Kushwaha, Tian Tian, Long Chen, and Chiam Ng Yu for their day-to-day laboratory advice. Thank you to my parents, Robert and Janet Smith, for encouraging my love of research and providing for me to attend Penn State.

# Chapter 1: Introduction

Organisms synthesize proteins by the cellular process of translation. Ribosomes, the protein-synthesizing machinery, bind to the messenger RNA (mRNA) which encodes the protein's sequence and produce a polymer of amino acids, the building blocks of proteins. Translation occurs in three phases: initiation, elongation, and termination (Kozak, 1999). Previous research has shown that translation initiation step is often regulated and is rate-limiting. It takes seconds for the mRNA and ribosome complex to assemble. Elongation occurs at a much faster rate, adding multiple amino acids per second to the growing polypeptide chain (Simonetti *et al*., 2009). Initiation controls which start codon is chosen, the reading frame for the protein, and the amount of ribosomes that can start translation (Kozak, 1999; Salis *et al.*, 2009). Due to the importance of translation initiation, it is a step that can be exploited to manipulate protein expression.

In the last 20 years, it has become significantly easier and cheaper to synthesize DNA, allowing many researchers to create synthetic DNA sequences to fit a desired function. The cost of creating synthetic gene sequences has decreased by almost three orders of magnitude, while the base pairs of DNA that an individual can assemble in one day has increased by five orders of magnitude (Carlson, 2009). By combining DNA from various natural and synthetic sources, known as recombinant DNA, we can design DNA with specific features and use it to evaluate the function of biological processes (Malik *et al.*, 1981). In metabolic engineering and pharmaceutical research, it is desirable to overexpress a protein of interest, which could be a therapeutic drug, vaccine component, an enzyme in laundry detergent, a biofuel, and many more

applications. There are various ways to increase the expression of a protein: increase the rate of transcription by changing the promoter sequence, increase the rate of translation initiation by changing the ribosome binding site, optimizing the codons of the protein to include fast and common codons, and to change the copy number of a plasmid or genome. Here, we focus on increasing protein expression by improving the translation initiation rate (TIR) through controlled manipulation of DNA near the ribosome binding site.

Previously, the only method available to researchers seeking to change protein expression was inducing mutations by trial-and-error to generate large libraries of clones with different properties. This process is inefficient, requires a high labor cost, and would be too time consuming for larger genetic systems and metabolic pathways. Strategically manipulating the DNA sequence of a desired protein allows for "rational control" over protein expression by altering the rate of translation initiation (Salis *et al.*, 2009). Translation initiation is affected by the individual mRNA sequence that is translated. The expression of a desired protein can be selectively changed with respect to the rest of the cellular proteins by changing the DNA (and thus mRNA) sequence of the ribosome binding site (RBS) and other regulatory sequences (Salis *et al.*, 2009; Salis, 2011). By altering these sequences systematically, one can efficiently search the DNA space to achieve a desired function (Farasat *et al*., 2014). Development of a biophysical model of translation initiation allows these changes to be made using statistical thermodynamics instead of trial-and-error mutations (Salis *et al*., 2009).

**Translation Initiation**

In bacteria, translation is carried out by the 70S ribosomal machinery and several initiation, elongation, and termination factors. The 70S ribosomal complex is composed of the small 30S subunit and the large 50S subunit. Initiation of translation depends on the binding of 30S to the mRNA transcript and the recruitment the initiator transfer RNA (tRNA), fMet-$tRNA^{fMet}$. This is considered the rate limiting step translation. Initiation establishes the reading frame by forming the first codon-anticodon interaction (Simonetti *et al*., 2009). Formation of the translation initiation complex is dependent on the binding of the 30S subunit to the mRNA and the activity of initiation factors IF1, IF2, and IF3. IF1 and IF3 function to supply 30S ribosomes to bind to the mRNA by promoting 70S dissociation and binding to the 30S subunit to prevent 50S binding, respectively. IF2 is a GTP-binding protein responsible for recruiting fMet-$tRNA^{fMet}$ to the ribosome, and the binding of the tRNA to the peptidyl (P) site of the 30S ribosome is stabilized by IF1 and IF3. The initiator tRNA and 30S binds the start codon, which always codes for the amino acid methionine. Interaction between the anticodon of fMet-$tRNA^{fMet}$ and the start codon is most stable when the start codon sequence is AUG. Weaker interactions occur between less common start codons, GUG and UUG, and the initiator tRNA (Kozak, 1999). This forms the 30S initiation complex.

The ribosome binding site consists of several sequences that are necessary for ribosome binding and translation initiation. Salis et al. (2009, 2011) define the ribosome binding site sequence as the 35 nucleotides upstream of the start codon and the beginning of the coding sequence, where the A in AUG is considered (+1). This definition includes a standby site, the Shine-Dalgarno (SD) sequence to which the 16S ribosomal RNA (rRNA) binds, the spacer region between the SD sequence and the start codon, and the start codon (Salis 2011).

The standby site is the sequence at the 5’ end of the long untranslated region (UTR) which is initially bound by the ribosome. The 30S ribosome contains a positively-charged platform domain that binds to the mRNA sugar-phosphate backbone at the 5’ UTR. The amount of available single-stranded mRNA for binding to the platform surface of the 30S subunit affects the translation initiation rate (TIR), which mainly depends on the geometry of the mRNA at the 5’ UTR. An upstream distal binding site, hairpin, and a downstream proximal binding site characterize a standby site module (Espah Borujeni, *et. al*., 2014).

The 30S ribosomal subunit consists of a core 16S rRNA which is surrounded by more than 20 small proteins. The last nine nucleotides at the 3’ end of the 16S rRNA bind to the Shine-Dalgarno sequence on the mRNA. The Shine-Dalgarno sequence is a three to nine base purine-rich sequence (A or G) upstream of the start codon. Interaction of the ribosome with the SD sequence keeps the ribosome positioned near the start codon and keeps the start codon available to ribosomes by preventing the reforming of secondary structures (Kozak, 1999). The consensus SD sequence in bacteria is AAGGAGGU (Mironova *et al*., 1999).

The ribosome contains an mRNA channel that the mRNA strand feeds through (Kozak, 1999; Yusupova *et al*., 2001). The channel was observed by x-ray crystallography as a groove on the 30S ribosomal subunit that accommodates approximately 30 single-stranded mRNA nucleotides. The channel consists of two sections: a section upstream of the start codon that accommodates the SD sequence interactions and a downstream section where the mRNA coding section has to be loaded. The middle of this channel is where the start codon associates with the initiator tRNA in the P site (Yusupova *et al*., 2001). Once the mRNA is fed through the channel and the 16S rRNA binds to the SD sequence, the ribosome is correctly positioned with the start codon base pairing with the anticodon of the initiator tRNA in the P site. The spacer region is the

nucleotide distance between the SD sequence and the start codon. Previous results have determined that 5 nucleotides is the optimal spacing (Chen *et al*., 1994), and that the efficiency of the complex is reduced when the ribosome has to accommodate a longer or shorter spacing region (Salis *et al*., 2009). Initiation is complete upon 50S subunit binding. The GTP bound molecule of IF2 is hydrolyzed, and the initiation events are now irreversible. The start site and reading frame are permanently chosen, and elongation begins (Simonetti *et al*., 2009).

**The Ribosome Binding Site Calculator**

A biophysical model has been developed using statistical thermodynamics to rationally control protein expression through the design of synthetic ribosome binding sites. The Ribosome Binding Site (RBS) Calculator allows scientists to enter a user-designed sequence and receive a prediction of protein production rate, or the calculator can design a synthetic ribosome binding site to achieve the desired protein expression (Salis, 2011). The RBS Calculator predictions are based on an algorithm that takes into account multiple free energies of molecular and sequence-dependent interactions during translation initiation (Salis *et al*., 2009).

The model uses thermodynamics to assess the free energy changes during the formation of the 30S-mRNA complex, defined as $\Delta G_{total}$. Salis *et al*. defined the relationship between the free energy change of the initiation complex formation to the translation initiation rate in equation 1. The proportionality constant, K, was experimentally determined to be 2500, and $\beta$ is the apparent Boltzmann constant, which was experimentally determined to be 0.45 kcal/mol (Salis *et al.,* 2009).

$$TIR = K\, e^{(-\beta * \Delta G_{total})} \quad (1)$$

The total free energy change is predicted based on the summation of the specific free energy changes that occur during complex formation (equation 2). The reference state is a single-stranded (unfolded) mRNA molecule with a folding energy of zero ($\Delta G_{mRNA}$ = 0 kcal/mol).

$$\Delta G_{total} = \Delta G_{spacing} + \Delta G_{mRNA:rRNA} + \Delta G_{start} + \Delta G_{standby} - \Delta G_{mRNA} \quad (2)$$

The free energy of spacing, $\Delta G_{spacing}$, is defined as the free energy penalty the ribosome must incur to accommodate a ribosome binding site with spacing between the SD sequence larger or shorter than the optimal five nucleotides (Chen *et al*., 1994). The spacer sequence can be as long as 13 nucleotides, but the efficiency of ribosome binding will be reduced (Kozak, 1999). This free energy is positive because the non-optimal spacing distorts the 30S initiation complex and the ribosome must incur a penalty to bind to this less entropically desirable site.

The free energy of the mRNA and rRNA interactions ($\Delta G_{mRNA:rRNA}$) is the free energy release that occurs when the 16S rRNA interacts with the SD sequence in the ribosome binding site ($\Delta G_{mRNA:rRNA} < 0$). The algorithm minimizes this free energy change of rRNA and mRNA binding and the penalty from the free energy from spacing between the SD sequence and the start codon. In this way, the model predicts the best binding possible using the last nine nucleotides of the 16S rRNA (3'-AUUCCUCCA-5') with any of the consensus SD sequence nucleotides an mRNA molecule contains (Salis *et al*., 2009).

The free energy of the start codon, $\Delta G_{start}$, is the energy released upon the start codon binding to the anticodon loop of the initiator tRNA.

The free energy of binding to the 5' UTR standby site, $\Delta G_{standby}$, reflects the work the ribosome must expend to accommodate the 5' UTR on its platform surface. The first version of the model defined the standby site as the four nucleotides upstream of the 16S rRNA binding site (Salis *et al*., 2009). Further research has shown that this definition was incorrect. Latest versions

of the model define the standby site module as an mRNA hairpin surrounded by upstream (distal) and downstream (proximal) single-stranded regions. The surface area of single stranded mRNA available to or unfolded by the ribosome affects the translation initiation rate. This area depends on the length of the distal and proximal binding sites and the height of the hairpin. Extensive characterizations of many structured standby sites showed that $\Delta G_{standby}$ is a function of mRNA's geometry and not its energetics (Espah Borujeni, *et al*., 2014).

$\Delta G_{mRNA}$ is the work that the ribosome must perform to unfold the mRNA's inhibitory structures sequestering the ribosome binding site. mRNA molecules can adopt favorable secondary structure, known as the minimum free energy structure ($\Delta G_{mRNA} < 0$), and the ribosome must change this conformation to bind to its binding site.

The first version of this model was experimentally validated on 119 natural and synthetic ribosome binding site predictions in *Esherichia coli*. The model can design sequences that manipulate translation initiation levels within a factor of 2.3 over a range of 100,000-fold (Salis *et al.*, 2009). Updates to this model include developing a comprehensive model for the surface area available to the ribosome at the standby site (Espah Borujeni *et al*., 2014).

Although the model considers the interactions of the upstream standby, spacing, and Shine-Dalgarno sequences of the ribosome binding site, there are no data on the interactions downstream of the start codon that may influence translation initiation. The question we seek to answer with this research is *what is the role of downstream structures in translation initiation?*

**The Ribosomal Footprint**

One downstream interaction is the ribosomal footprint – the length mRNA after the start codon that is unfolded by the 30S ribosome and loaded into the ribosomal mRNA channel after the P-site (Figure 1).

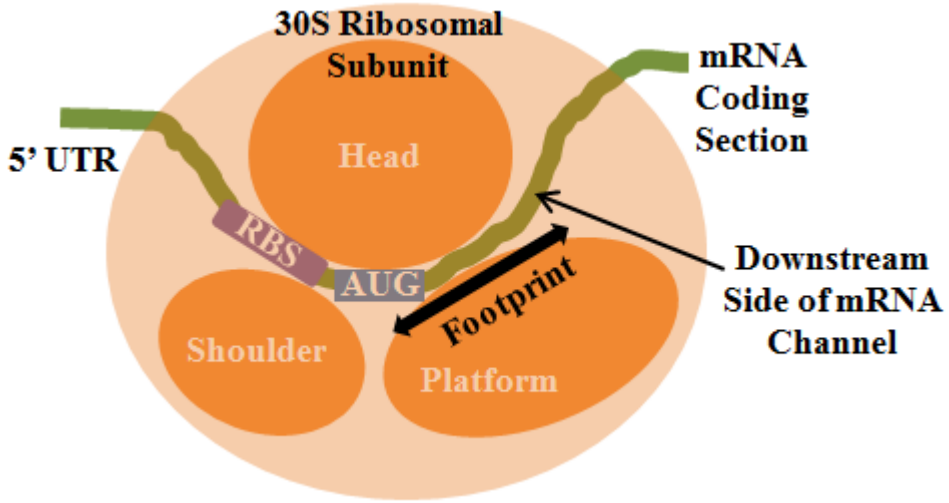



**Figure 1: Structure of mRNA Channel in 30S Ribosome**

In the original version of the model, there was no knowledge concerning the effect of downstream structures on the translation initiation. For simplification, the footprint was defined such that it was sufficiently large as to exclude mRNA secondary structures where the ribosome would bind. Current versions of the model have included a footprint of $n = 35$ nucleotides (Salis *et al*., 2009, 2011). Previous estimates place the size of the footprint at 19 nucleotides from the start codon (Hüttenhofer *et al*., 1994).

We sought to determine the role of the ribosome on unfolding downstream structures. Defining the exact number of nucleotides that the ribosome is required to unfold will improve the accuracy of the model's predictions. If there is a hairpin within the footprint, the ribosome

must expend a lot of free energy to load this hairpin into the mRNA channel, which has a diameter that can only accommodate single-stranded mRNA (Yusupova *et al*., 2001). Previous research has shown that the ribosome has helicase activity after elongation starts, as the ribosome must unfold structures it encounters along the mRNA molecule for translation elongation to proceed. The ribosome itself can be considered as a helicase, with helicase activity in the middle of the mRNA channel. Ribosomal subunits S3, S4, and S5 also confer helicase activity. They form the mouth of the tunnel which the downstream mRNA enters. Mutations in these subunits have been shown to decrease helicase activity. However, it does not appear to function during translation initiation (Takyar *et al*., 2005). Thus, the ribosome must expend more energy to unfold the downstream mRNA structures during translation initiation when helicase activity is not available. During translation elongation, the ribosome must input free energy to unfold downstream mRNA structures, but the helicase activity acts as a catalyst. However, in the absence of helicase activity during translation initiation, the ribosome most likely unfolds these structures by energy-intensive random, ratchet-like motions (Shuwirth *et al*., 2005). This will add an additional, very positive free energy change, $\Delta G_{load}$ to the code for the RBS calculator.

Secondary structures in the mRNA that exclude the ribosome from binding will yield a lower translation initiation rate estimate, and RBS sequences generated by the calculator that take into account this footprint will more closely resemble observed protein expression. We sought to determine the ribosomal footprint in *Escherichia coli* by varying the distance of a singular secondary structure in the coding section from the start codon, ATG, and by characterizing the resulting protein production rate. Sequences with the secondary structure close to the start codon are expected to produce little protein, and sequences with the hairpin located at a large distance from the start codon are expected to express significantly more protein. It is

hypothesized that protein expression will gradually increase as the hairpin moves closer to the footprint length, as the ribosome may be able to unfold part of the hairpin. Protein production is expected to level off and be constant once the hairpin is placed at the footprint length and beyond. In this research, we define the footprint (FP) as the length of nucleotides including ATG that the ribosome would have to unfold to load into the mRNA channel.

# Chapter 2: Design

In order to determine the length of the footprint, it was necessary to design a protein coding sequence (CDS) with a movable hairpin structure that would remain the same in all constructs. The distance the hairpin was away from the start codon would represent the footprint, N. It was also necessary to design an RBS with a high level of protein expression to show if there was a significant difference between translation initiation when the ribosome footprint is unobscured and the decreased translation initiation that was expected when a hairpin was within the footprint length. The RBS would also remain constant throughout all constructs. The ribosome binding site was designed to have a strong 5' UTR hairpin, a Shine-Dalgarno sequence with a translation initiation rate of approximately 100,000 au, and a spacing of five nucleotides between the SD sequence and the start codon AUG. The most common start codon (AUG) was used to maximize expression and be accurately predicted by the calculator. The design of these constructs is shown in Figure 2.

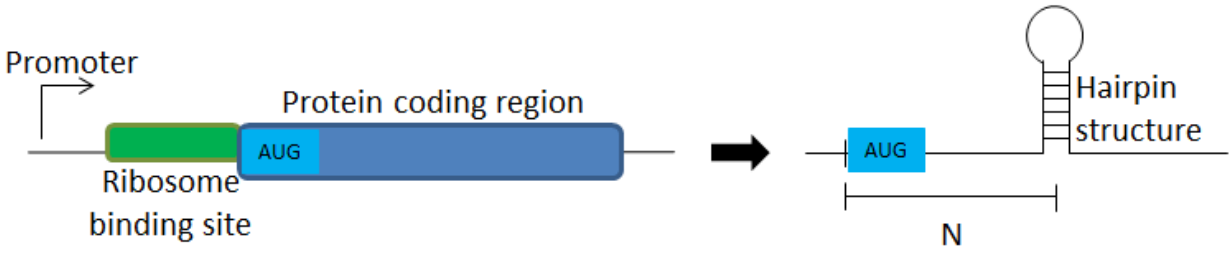



**Figure 2: RBS and CDS Design for a Varying Footprint Length**

We used red fluorescent protein (RFP) as the protein that would be expressed. Using RFP enabled us to visually observe changes in translation initiation rate in bacterial colonies that were producing little protein (white) to expressing a lot of protein (deep pink) and this would be confirmed quantitatively by fluorescence measurements. We chose to use *Escherichia coli* (strain

DH10B) as the chassis for this experiment. *E. coli* is well characterized, easily manipulated, and was the original host in which RBS Calculator model was tested. A sequence of 44 nucleotides was appended to the beginning of the RFP protein that the hairpin would be moved within. This would ensure that the structure and function of RFP was not altered by the addition of a hairpin containing different amino acids and would allow for comparison between all the constructs based only on the position of the hairpin.

The criteria for design are listed below:

1. Standby Site Hairpin
2. Shine-Dalgarno Sequence
3. Restriction Enzymes and Spacing
4. Coding Section Hairpin
5. Codon Usage
6. Final Constructs

**Standby Site Hairpin**

We designed a strong hairpin for the upstream 5' UTR standby site of the RBS. The main purpose of this hairpin is to preserve the structure of the system and prevent any undesirable folding. Additionally, this standby site provides sufficient single-stranded surface area for binding to the ribosomal platform (Espah Borujeni *et al*., 2014). Creating a favorable hairpin upstream of the Shine-Dalgarno and within standby site sequences ensures that this structure will form and other weaker associations that could obscure the SD sequence or the start codon will not form. This strong hairpin was designed with the stem containing six G:C base pairs. Guanine and cytosine nucleotides associate with three hydrogen bonds, versus two hydrogen bonds for adenine and uracil/thiamine, so G:C base pairs were chosen to create a stronger hairpin. The loop nucleotides were chosen to be adenine and thiamine and cytosine bases to prevent additional

base pairing. The sequence of the hairpin is “CCCGCCatatacGGCGGG”, where the capitalized letters form the stem and the lowercase letters form the loop.

We chose six nucleotides each in the stem and loop to ensure they would not be cut by ribonucleases (RNases). Previous research has shown that RNase III recognizes double stranded mRNA with no bulges (mismatched pairs) of 12 nucleotides or greater and cleaves them. The presence of this strong hairpin sequestering the RBS also protects the 5’ end of the mRNA from RNase degradation (Pertzev *et al*., 2006). The strength of the hairpin is approximately -11.8 kcal/mol $\Delta G_{fold}$ as predicted by the University of Vienna’s RNAfold WebServer using the parameters of the Turner 1999 model, minimum free energy (MFE) and partition function, avoid isolated base pairs, and no dangling end energies (Gruber *et al*., 2008). This hairpin is shown in Figure 3.

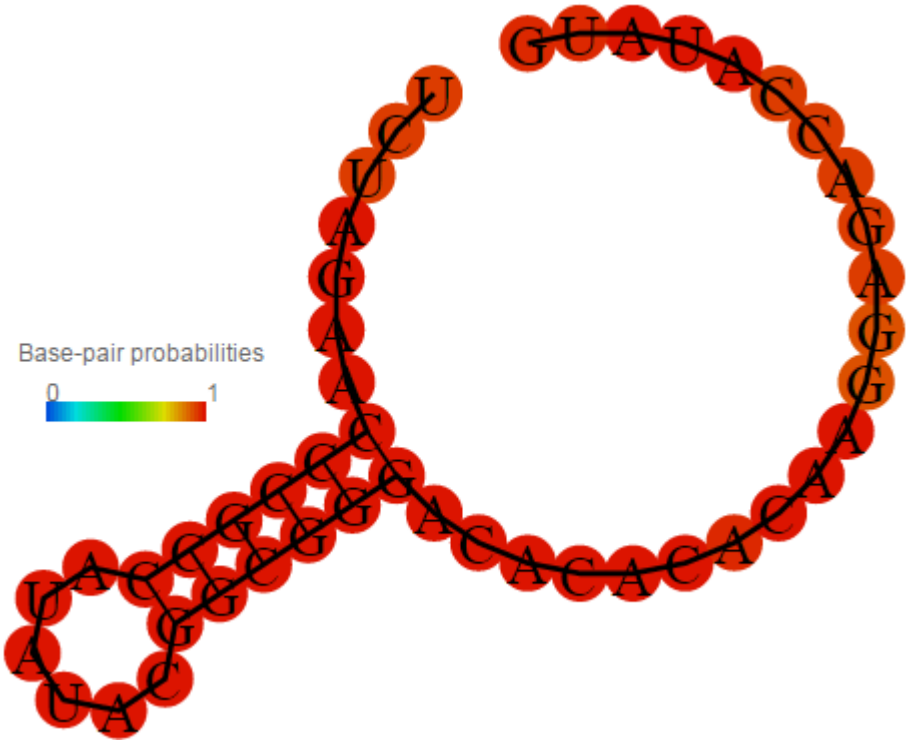



**Figure 3: Ribosome Binding Site Hairpin (Gruber et al., 2008)**

The color of the structure denotes the probability of the base pairings, where low probability of base pairing is indicated by blue and high probability of base pairing is indicated by red. For unpaired regions, the red color symbolizes the high probability of the region remaining unpaired (Gruber *et al*., 2008). Additionally, we designed an XbaI ("TCTAGA") restriction enzyme site at the 5' end UTR because it was unique to the plasmid and would allow simple insertion of the RBS via digestion and ligation.

**Shine-Dalgarno Sequence**

The complete SD sequence, fully complimentary to the last nine nucleotides of the 16S rRNA in *E. coli* is "TAAGGAGG," and we truncated the sequence to "AAGGAG", achieving a translation initiation rate (TIR) of 104,166.8 au and minimizing secondary structure formation with the coding section (Ringquist *et al*., 1992). The Ribosome Binding Site Calculator version 2.0 was used to predict the TIR (Salis, *et al*., 2009). We sought to produce a TIR that was sufficiently high to show the difference between unhindered ribosome binding and hindered ribosome binding. We suspected that a hairpin close to the length of the footprint would most likely be unfolded by the ribosome and the energy penalty of unfolding would not be great enough to prevent protein expression. As the hairpin became closer to the start codon, we expected it would incur a greater energy penalty for the ribosome to unfold, producing progressively less protein and progressively lighter-colored colonies. We expected constructs with the hairpin immediately adjacent to the start codon to have little or no expression, and these colonies would be white. Using an RBS with a high TIR ensured that we could see a range of colony colors and corresponding range of fluorescence measurements.

The SD sequence was chosen carefully not to have secondary structure and interact with the 5' UTR hairpin. This proved difficult because repetitive sequences (AA, GG) can base pair easily with other repetitive sequences (TT, CC). We included 8 nucleotides between the last base of the hairpin and the start of the Shine-Delgarno (SD) sequence, the majority of which were adenine and cytosine nucleotides in order to maintain a single-stranded structure. This sequence prevented the SD sequence from hybridizing with other upstream sequences.

**Restriction Enzymes and Spacing**

While the ribosome binding site remained constant throughout the experiment, the coding section was variable. For ease of creating the different constructs, adding restriction sites by the start codon and at the end of the leading nucleotide sequence was necessary. However, it was desired to keep the spacing between the SD sequence and the start codon at five nucleotides. Restriction sites were evaluated for their nucleotide sequence and length (see APPENDIX for list of restriction sites). Some restriction sites were unsatisfactory because they were too closely related to the SD sequence and increased the predicted TIR, while others were too long or required too many nucleotides to separate the site from the SD sequence. We chose the NdeI restriction site, "CATATG", because it contains the start codon ATG (AUG in mRNA) and fits well within the five nucleotide limit. The five nucleotides after the SD sequence are "ACCAT." We chose the restriction enzyme SacI (GAGCTC) to place at the end of the leading coding sequence because it was unique to the plasmid we were working with, the pFTVP vector (See APPENDIX).

**Coding Section Hairpin**

The hairpin design for the coding section was similar to that of the RBS hairpin, with a $\Delta G_{fold}$ of -11 kcal/mol as predicted by the ViennaRNA webserver (Gruber *et al*., 2008). This hairpin was designed to be strong enough so it would be the most favorable structure to form and prevent undesirable or additional structures from forming in the coding section. Unless otherwise noted, the CDS hairpin stem and loop lengths were six nucleotides so the hairpin could be moved effectively without altering the reading frame of the coding sequence or being degraded by RNases. The stem contains four complementary codons containing G and C bases, while two codons were used in the loop. The sequence of the hairpin is "GGCGCGattatcCGCGCC" where the capitalized letters form the stem and the lowercase letters form the loop. It was necessary for the coding section hairpin geometry to contain 18 nucleotides so that the hairpin, when unfolded, would form 6 codons. The CDS hairpin was designed to be moved between each set of three nucleotides (codons) to preserve the reading frame. If the number of nucleotides was not a multiple of three, the reading frame would be altered. This would change the proteins produced and could have made the red fluorescent protein nonfunctional. This hairpin is shown in Figure 4.




**Figure 4: Coding Section Hairpin (Gruber *et al*., 2008)**

This hairpin was used to map the ribosomal footprint in a codon-based resolution. To map the ribosomal footprint function in a nucleotide-based resolution, we need to insert the CDS hairpin at out-of-frame positions. In order to keep a constant reading frame, we needed to modify the loop sequences, change the base pair sequence of the hairpin, or insert additional nucleotides next to the hairpin. We designed two additional CDS hairpins with sequences "GCGCGGccattatCCGCGC" and "CGGCGCcattatccGCGCCG". The folding energies were predicted to be -9.41 kcal/mol and -10.51 kcal/mol respectively using ViennaRNA (Gruber *et al*., 2008). These hairpins were to be positioned 7, 8, 10, 11, 13, and 14 nucleotides into the coding section, including the start codon. We also designed two hairpins with one and two A:T base pairs at the bottom of the hairpin to investigate whether the ribosome would expend less free energy unfolding these weaker structures. The sequences of these hairpins are "ACCGCGattatcCGCGGT" and "ATCGCGattatcCGCGAT" with strengths of -9.53 kcal/mol and -7.42 kcal/mol respectively (Gruber *et al*., 2008). One additional sequence we tested was a significantly larger hairpin with three sections of hairpin with 6, 5, and 6 nucleotides in the stem, and bulges (mismatched nucleotides) in between to prevent degradation by RNases: "CGGCGCcaTCGCGaaCGCCGCaattatcaGCGGCGccCGCGAacGCGCCG." The strength of this hairpin was -33.34 kcal/mol (Gruber *et al*., 2008). All the hairpin sequences and positions are shown in Table 2. Constructs labeled $13^+$ and $14^+$ have A:T and 2 A:T base pairs in their hairpin, and construct $17^+$ contains the large hairpin.

**Codon Usage**

A sequence of 44 was appended to the beginning of the RFP. Including the SacI restriction site, the sequence is 51 nucleotides. The leading sequence must be a multiple of three; otherwise it would alter the reading frame of the protein. The original design was to be 35 nucleotides or longer in order to fully validate the effect of downstream structures on the translation initiation rate. The model currently uses 35 nucleotides as the prediction for the area that will be unfolded for the ribosome, so any sequence longer than this was acceptable.

The coding sequence must be single stranded with the exception of the hairpin. In order to accomplish this, codons with mostly adenine and cytosine bases were used to minimize base pairing. Codons that had repeated or adjacent guanine or cytosine nucleotides were generally avoided, as they were more likely to form secondary structures. Codons with A,T, and C nucleotides produced the least secondary structure. We arranged these codons so that the ending position of one codon would not pair with the beginning position of a following codon when the hairpin was inserted between them: there are no adjacent GC or AT nucleotides in the coding section, excluding the hairpin itself. No additional start or stop codons were used in the modified coding section.

There are 64 codons that code for 20 amino acids and three codons that code for the ribosome to stop translation (stop codons). When multiple codons code for the same amino acid, each codon has a different percentage of use in the *E. coli* genome. A prior experiment revealed that adjacent codon repeats and inclusion of rare codons severely inhibited protein expression. No codon was repeated more than three times, and two adjacent codons were never repeats. This was to ensure that the modified RFP reporter gene would not consume too much of one cellular

resource and be selected for mutation. For our purposes, rare codons were defined to be codons with usage percentages of 0-20% of degenerate codons (codons that code for the same amino acid) and genome usage rates below 2.0%. All codons chosen had at least 2.0% usage in *E. coli*, with the exception of the codons in the SacI restriction site*, and were the most common codons for each amino acid. Table 1 shows the percentage usage of the most common codons within the *E. coli* genome (Maloy, 2002). (See APPENDIX for more detailed codon usage information).

**Table 1: Included Codons**

| Amino Acid | Codon | Percent Usage (%) | Number of Use |
|---|---|---|---|
| Ala (A) | GCG | 3.2 | 2 |
| Ala (A) | GCC | 2.3 | 1 |
| Arg (R) | CGC | 2.2 | 1 |
| Asn (N) | AAC | 2.6 | 3 |
| Gln (Q) | CAG | 2.9 | 3 |
| Glu (E) | GAA | 4.4 | 1 |
| Gly (G) | GGC | 3 | 1 |
| Lys (K) | AAA | 3.8 | 3 |
| Ile (I) | ATC | 2.7 | 3 |
| Ile (I) | ATT | 2.7 | 1 |
| Ser (S)* | AGC | 1.5 | 1 |
| Ser (S)* | TCT | 1.1 | 1 |
| Thr (T) | ACC | 2.4 | 1 |

The final design of the leading coding section plus the 18 nucleotide hairpin is shown in Figure 5.

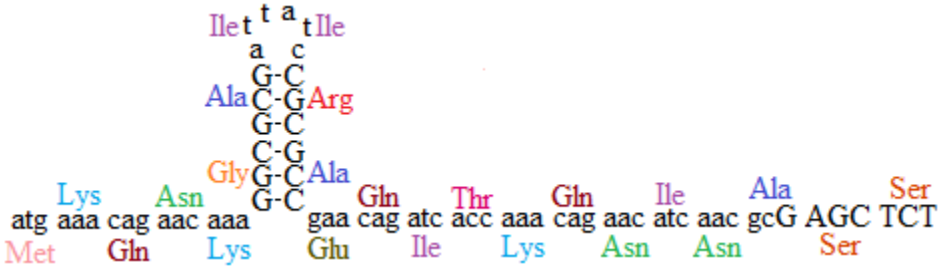



**Figure 5: Codon Usage in Leading Sequence (Footprint Length 15)**

**Final Constructs**

We designed 21 almost identical coding sections of around 51 nucleotides, with the exception of the hairpin, which was moved between every codon from immediately following the start codon (FP=3) to a footprint of 39, at the end of the modified coding sequence. This was to ensure that the reading frame was kept constant with no rare codons introduced, and to provide a preliminary prediction for the ribosome binding footprint. Footprint lengths of 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 21, 24, 30, 33, and 39 were tested. A footprint length greater than 44 nucleotides was the positive control (no hairpin present).

There was to be no secondary structure in the coding section except the prescribed hairpin, and every sequence was linear, with one exception. Moving the coding section hairpin to position 7 caused unwanted base pairing between the coding section and the SD sequence of the RBS. A threonine codon, ACC – the only codon in the coding section with adjacent C nucleotides – had to be switched with an adjacent codon, and the secondary structure was resolved.

The final RBS and CDS sequences are shown in Table 2. The positive control is the sequence without a hairpin. This sequence is expected to produce the most protein because the ribosome footprint region will not be inhibited by secondary structure. As the hairpin position is moved farther away from the start codon, it was expected that the amount of protein expressed would gradually increase.

## Table 2: Final RBS and CDS Sequences

| **RBS** | TCTAGAaCCCGCCatatacGGCGGGacacacacAAGGAGacCAT |
|---|---|
| | |
| **FP** | **Coding Section** |
| 44 (pos. cont.) | ATGaaacagaacaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| 3 | ATGGGCGCGattatcCGCGCCaaacagaacaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| 6 | ATGaaaGGCGCGattatcCGCGCCcagaacaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| 7 | ATGaaacGCGCGGccattatCCGCGCagccaacaaagaacagaccatcaaacagaacatcaacgcGAGCTC |
| 8 | ATGaaaatCGGCGCcattatccGCGCCGcagaaaacaaagaacagaccatcaaacagaacatcaacgcGAGCTC |
| 9 | ATGaaacagGGCGCGattatcCGCGCCaacaaagaacagaccatcaaacagaacatcaacgcGAGCTC |
| 10 | ATGaaacagcGCGCGGccattatCCGCGCcaacaaagaacagaccatcaaacagaacatcaacgcGAGCTC |
| 11 | ATGaaacagatCGGCGCcattatccGCGCCGtaaacaaagaacagaccatcaaacagaacatcaacgcGAGCTC |
| 12 | ATGaaacagaacGGCGCGattatcCGCGCCaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| 13 | ATGaaacagaaccGCGCGGccattatCCGCGCcaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| $13^{+}$ | ATGaaacagaacACCGCGattatcCGCGGTaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| 14 | ATGaaacagaacatCGGCGCcattatccGCGCCGtacagatcaccaaacagaacatcaacgcGAGCTC |
| $14^{+}$ | ATGaaacagaacATCGCGattatcCGCGATaaagaacagatcaccaaacagaacatcaacgcGAGCTC |
| 15 | ATGaaacagaacaaaGGCGCGattatcCGCGCCgaacagatcaccaaacagaacatcaacgcGAGCTC |
| $17^{+}$ | ATGaaacagaaagacaaCGGCGCcaTCGCGaaCGCCGCaattatcaGCGGCGccCGCGAac GCGCCGaacagatcaccaaacagaacatcaacgcGAGCTC |
| 18 | ATGaaacagaacaaagaaGGCGCGattatcCGCGCCcagatcaccaaacagaacatcaacgcGAGCTC |
| 21 | ATGaaacagaacaaagaacagGGCGCGattatcCGCGCCatcaccaaacagaacatcaacgcGAGCTC |
| 24 | ATGaaacagaacaaagaacagatcGGCGCGattatcCGCGCCaccaaacagaacatcaacgcGAGCTC |
| 30 | ATGaaacagaacaaagaacagatcaccaaaGGCGCGattatcCGCGCCcagaacatcaacgcGAGCTC |
| 33 | ATGaaacagaacaaagaacagatcaccaaacagGGCGCGattatcCGCGCCaacatcaacgcGAGCTC |
| 39 | ATGaaacagaacaaagaacagatcaccaaacagaacatcGGCGCGattatcCGCGCCaacgcGAGCTC |

The structure of the RBS is shown in Figure 6. It includes an XbaI restriction site at the 5’ UTR, a standby site, the Shine-Dalgarno sequence, a five nucleotide spacer, and an NdeI site that includes an ATG start codon.

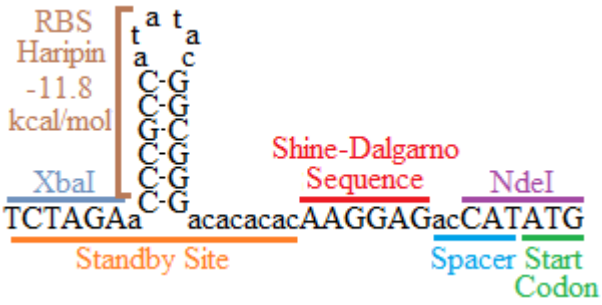



**Figure 6: Final Design of RBS**

Figure 7 shows the minimum free energy structure generated by the Vienna RNA webserver for the RBS and CDS construct of a footprint length of 15. The MFE diagram reflects that both hairpins serve their function to prevent undesired secondary structures from forming.

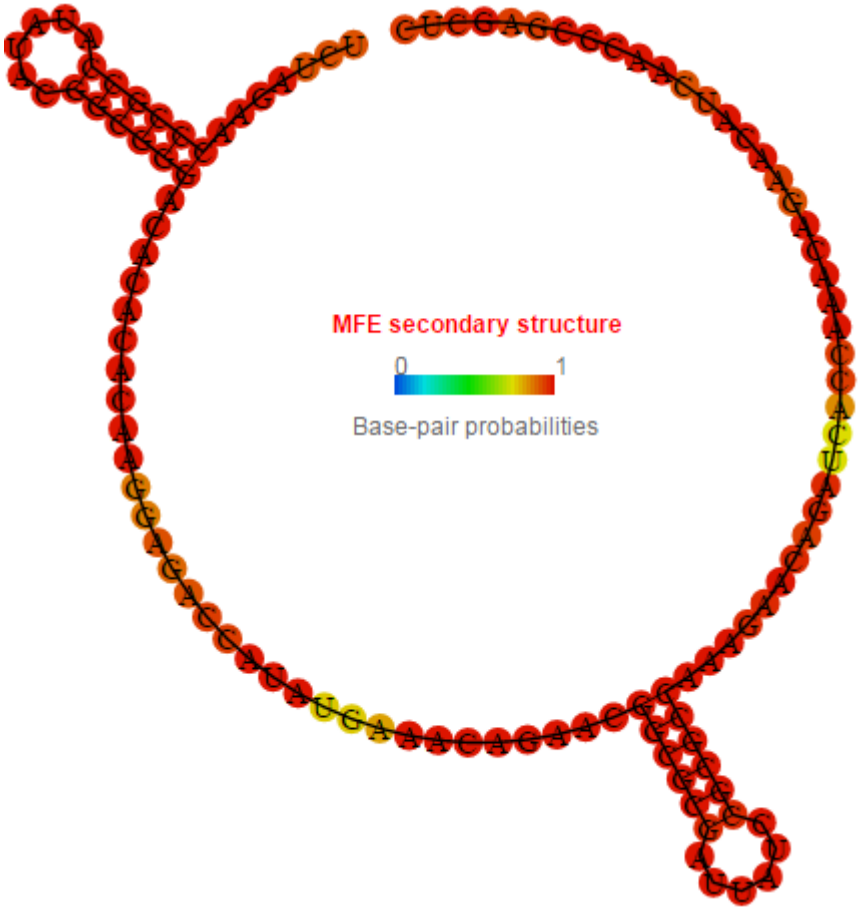



**Figure 7: Minimum Free Energy Structure of Final Construct 12 with RBS and CDS Hairpins (Gruber *et al*., 2008)**

# Chapter 3: Experimental Methods

## Plasmid Construction

The expression system used in this research, the pFTVP vector, was modified from the pFTV1 plasmid (Salis *et al*., 2009) (See APPENDIX for pFTVP sequence). It contains a constitutive $\sigma^{70}$ promoter (BioBrick #J23100), a ColE1 replication origin, a chloramphenicol resistance gene (CmR), and the mRFP1 codon-optimized fluorescent protein. The ribosome binding site and positive control modified coding section sequence were constructed using overlapping oligonucleotide primers (ordered from Integrated DNA Technologies, IDT). These were digested with XbaI and SacI restriction enzymes and were ligated into the digested pFTVP vector using similar restriction enzymes. The final ligation products were transformed into the electro-chemically competent *E. coli* cells (strain DH10B) and plated on Luria-Bertani (LB) agar with 50 μg/ml Chloramphenicol (Cm) antibiotic (Sigma-Aldrich). At least two colonies were picked and grown in LB/Cm media overnight and the plasmids were extracted the next day using common plasmid mini-prep columns (Omega). The sequences of the extracted plasmids were determined using sequencing techniques and the correct plasmid was chosen for the next step.

In the second step of cloning, the coding section constructs containing a hairpin at various positions were PCR-assembled, digested with NdeI and SacI enzymes, and ligated to the newly constructed pFTVP vector (Figure 8), followed by transformation into *E. coli* cells and plating on LB agar. Chloramphenicol antibiotic was used as the selection. The correct clones were sequence-verified. The positive control colonies were very small and dark red, taking approximately three days to grow. The hairpin containing constructs were expected to be white, larger, and produce little fluorescent protein.

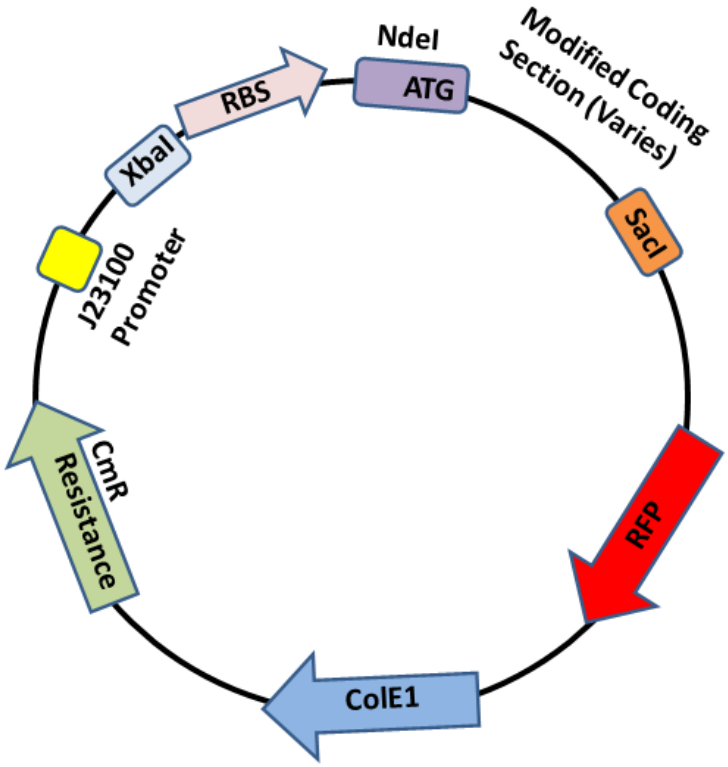



**Figure 8: pFTVP Vector Expression System**

**Storing the Plasmids**

Plasmids with correct sequences were re-transformed into *E. coli* DH10B cells using electroporation and plated on LB/Cm agar. One colony was picked per construct and grown overnight in LB/Cm media. Some of the longer footprint constructs that were expressing the most protein took two or three days to grow. The cultures were centrifuged at 4750 RPM, 37 ºC. The supernatants were removed and the cell pellets were re-suspended into 500 µl LB Miller broth supplemented with 500 µl of 50% glycerol and 50 µg/ml Cm. These 20% glycerol stocks were stored at -80 ºC for long term storage.

**Growth and Fluorescence Measurements**

Two colonies of each construct, together with the positive control, and the parent DH10B strain were incubated overnight at 37 °C, 200 RPM in a 96-well deep well plate containing 700 µl LB Miller media and 50 µg/ml Cm antibiotic. The next day, 10 µl of each sample was transferred to a new 96-well micro-assay plate containing 190 µl LB Miller media and 50 µg/ml Cm antibiotic. The cultures were grown continuously with shaking using a spectrophotometer TECAN machine. The optical density ($OD_{600}$) and fluorescence were measured every 10 minutes. After the $OD_{600}$ reached 0.15 (media corrected), the cultures were diluted 50 times by transferring 10 µl of cultures into 190 µl LB Miller media and 50 µg/ml Cm antibiotic in a fresh 96-well micro-assay plate ($OD_{600}$ of ~0.01). Cultures continued to grow in the second plate until they reached an $OD_{600}$ of 0.15. At this time, 10 µl of culture were transferred to 190 µl phosphate buffered saline (PBS) solution two new 96-well micro-assay plate, sealed, and stored in 4 ºC fridge for flow cytometry measurements. Another 10 µl of cultures were then transferred to a fresh 96-well micro-assay plate with LB Miller media and 50 µg/ml Cm antibiotic and grown similarly until they reached an $OD_{600}$ of 0.15. From this plate, 10 µl of culture were transferred to a fresh 96-well micro-assay plate for flow cytometry measurements. The fluorescence levels from two 96-well micro-assay plates were measured using a Fortessa Flow Cytometer and 100,000 events. The experiments were performed in two separate days (n=4).

# Chapter 4: Results

## Downstream structures greatly affect translation initiation

The positive control (no hairpin) and the coding sections with a movable hairpin were measured for fluorescence using flow cytometry. The data in Figure 9 show a sigmoidal-like curve of protein expression. In this experiment, we demonstrated that protein expression increased 1000-fold through controlling the footprint length. From the data, the ribosomal footprint is approximately 13 or 14 nucleotides, decreasing the footprint from previous estimates (Hüttenhofer *et al.*, 1994). Due to the highly expressing RBS, the constructs with a hairpin beyond the footprint range grow slowly and suppressive mutations are highly favored, which could account for some fluctuation in expression.

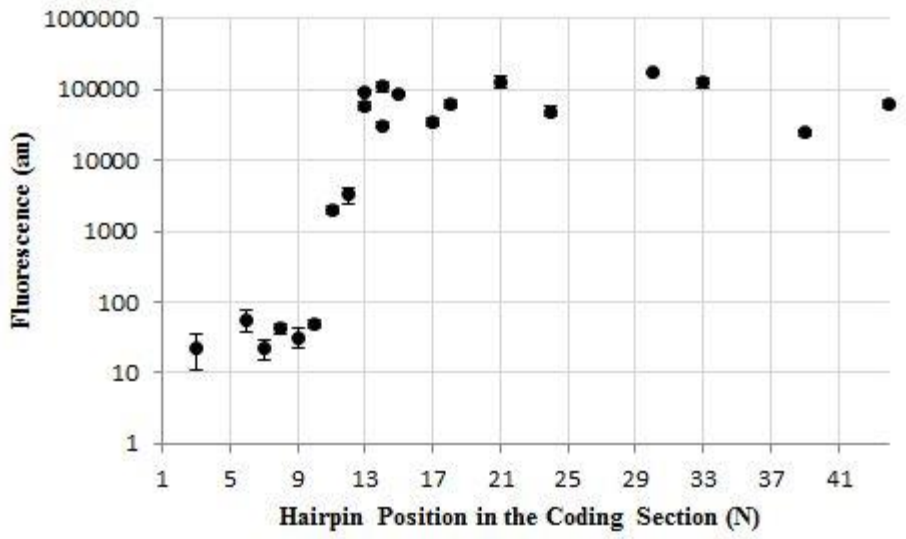



**Figure 9: Fluorescence Data for a Movable Hairpin at Varying Distances Away from the Start Codon**

**Model predictions for different footprint assumptions**

Our goal is to determine the best footprint range so we can update the model with this information. We used the Ribosome Binding Site Calculator version 2.0 to predict the TIR with different footprint lengths ranging from 0 to 35 (Espah Borujeni *et al.*, 2014; Salis *et al.*, 2009). A footprint of 0 means that the ribosome does not unfold any downstream structure. The ribosome must load every hairpin into the channel. A footprint of 35 means that the ribosome unfolds any secondary structure in the first 35 nucleotides of the coding section. In order to determine the footprint, we must match the data with the model's predictions. Figure 10 shows the model predictions (colored lines) overlaid with the fluorescence data. Predictions for footprints of 13 and 14 nucleotides fit the majority of the data in the transition between low and

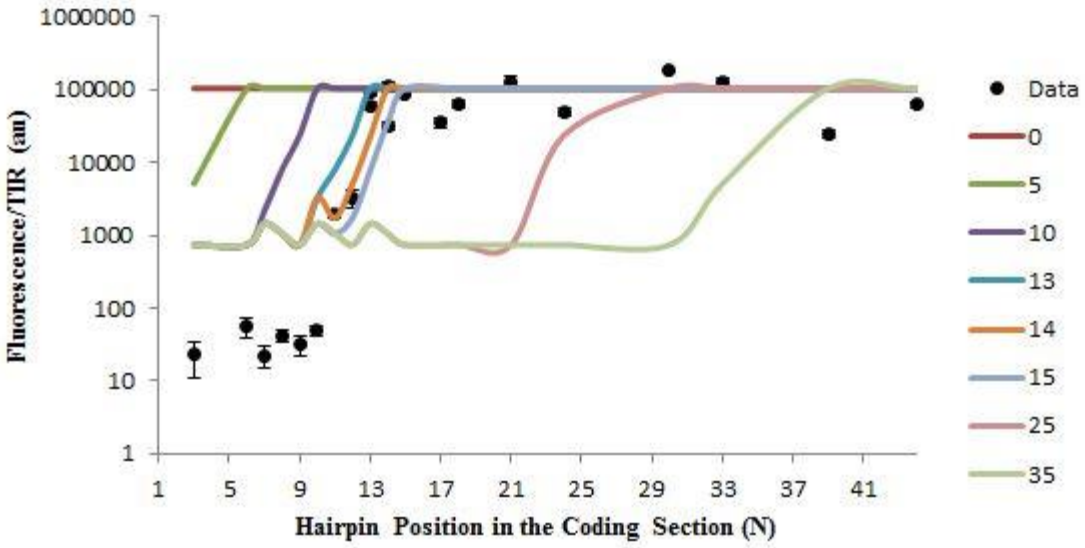



**Figure 10: Comparison of Model Predictions and Experimental Data**

high expression. The discrepancy between the lowest model predictions and the data for secondary structure within 10 nucleotides of the start codon could be due to energetic differences.

**Ribosomal footprint unfolds 13 or 14 nucleotides downstream of the start codon**

From the data above, we see that footprints 13-14 fit the data best. The ribosome must unfold the first 13 or 14 nucleotides of the coding section including the start codon to load the mRNA into the channel. We can determine the most accurate footprint length by determining the average error for each footprint assumption. The footprint assumption that minimizes the error is the value that should be used in future RBS Calculator predictions.

In order to calculate the error, we used equation 1 to calculate the apparent $\Delta G_{total}$ from the measured fluorescence data. Manipulation of equation 1 yields the following equation (3):

$$\Delta G_{total} = -\frac{\ln\left(\frac{TIR}{K}\right)}{\beta} \quad (3)$$

The fluorescence of a protein is a measure of the translation initiation rate. Substituting the fluorescence values, 2500 for K, and 0.45 kcal/mol for $\beta$, we calculated the apparent $\Delta G_{total}$. We compared this apparent total free energy to the total free energy calculated by the RBS Calculator. The error in $\Delta G_{total}$ ($\Delta\Delta G_{total}$) represents the absolute value of the difference between the predicted and experimental total free energies. The model prediction that most resembles the data has the lowest error. Therefore, we found the value of the footprint that minimizes the error by averaging the error values in $\Delta G_{total}$ for each footprint assumption.

Figure 11 shows the average error for each footprint prediction over the range of 0 to 35 footprint lengths. From this data, a footprint length between 13 and 14 nucleotides minimizes the error between the model's predictions and the data. Inputting a footprint parameter of 13 or 14 into the biophysical model will yield the most accurate prediction of the translation initiation rate. More sequences with varying hairpin strength and geometry will need to be tested before we can definitively state the exact footprint length.

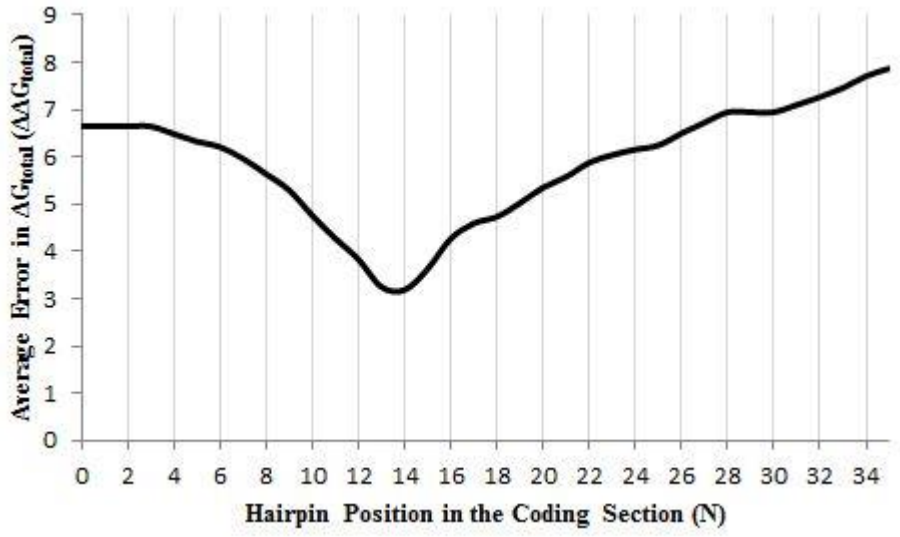



**Figure 11: Average Error of Model Predictions over Varying Footprint Lengths**

**The 30S ribosomal subunit follows thermodynamic minimization**

A secondary structure in the mRNA that is located at the edge of the footprint will be closest to the helicase activity of ribosomal S3, S4, and S5 proteins that surround the exit mouth of the mRNA channel. The probability of the ribosome unfolding this hairpin is greater than the probability of unfolding than a hairpin within the footprint length, which is less energetically favorable. We hypothesized that placing a weaker hairpin close to the footprint would be more favorable for the ribosome to unfold because there will be less of a free energy penalty, and we expected this would correlate with higher expression. The ribosome lacks the helicase activity during initiation and must minimize its energetic work input so any sequence mutation that reduces this work input is favorable. A:T base pair mutations are favorable because there are only two hydrogen bonds between the A and T bases versus three hydrogen bonds for G and C. We modified the coding section hairpin to have one or two A:T base pairs at its base instead of G:C base pairs. We predicted that the ribosome may be able to unfold some or all of the weaker base pairs in the hairpin, allowing it to feed more single-stranded mRNA through the mRNA channel.

We expected that when the hairpin is placed after the $12^{th}$ nucleotide, a ribosome with a footprint of 13 must unfold the first base pair of the hairpin structure. In a secondary structure where the first base pair is G:C, this will be difficult for the ribosome to unfold. However, when the first base pair is mutated to A:T, the ribosome can unfold that hairpin with less of an energetic investment. We expected that a hairpin with two A:T base pairs at its base would also correspond to significantly increased protein expression. The model predicts this; however, it is not shown in the data (Figure 12).

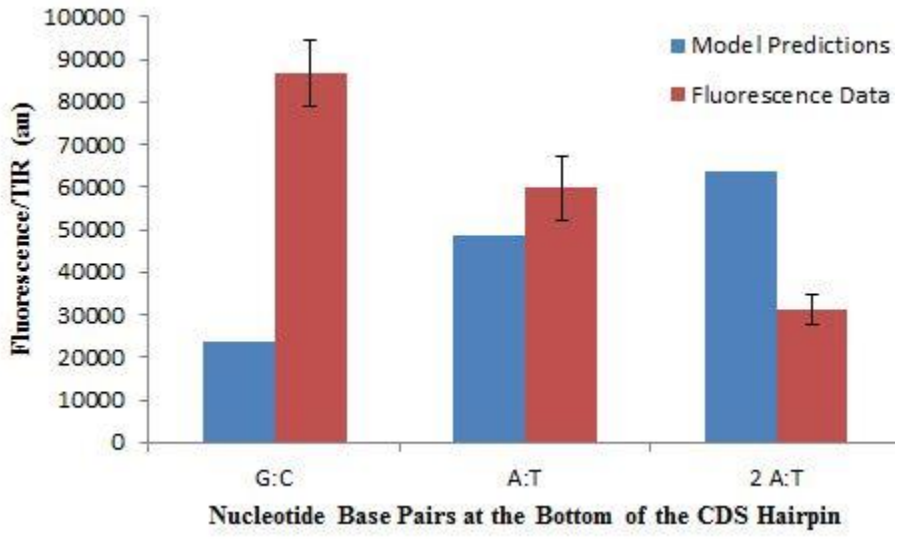



**Figure 12: Comparison of Fluorescence and Model Predictions for Variation of Base Pairs at the Bottom of the CDS Hairpin**

This demonstrates that the interactions occurring to load in the mRNA may not be as simple as just unfolding the secondary structure. The model predicts low translation initiation for a G:C nucleotide pair, but the data show higher protein expression. Perhaps it is more energetically favorable for the ribosome to compress slightly instead of unfolding this hairpin. If this were the case, we would expect to see this borne out for the A:T and 2 A:T pairs, but protein expression decreases. More interactions must be investigated to fully elucidate the mechanism of translation initiation at the ribosomal footprint.

**A correct ribosomal footprint improves the model's predictions**

We expect that there is a linear relationship between an increase in translation initiation rate and an increase in protein expression. We compared the model predictions for a footprint of 13, 14, and 35 with the fluorescence data to see which footprint yields a linear relationship. When a footprint of 13 is applied in the model, there is a linear relationship between TIR and fluorescence. The $R^2$ value is 0.9574 indicating that the linear relationship is well fit by the data (Figure 13).

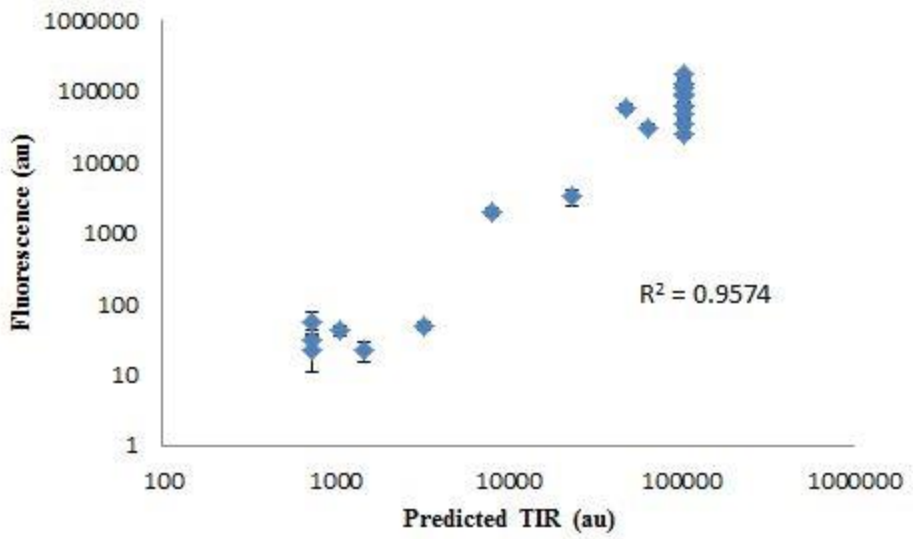



**Figure 13: Linear Relationship between the Predicted TIR at a Footprint of 13and the Fluorescence Data**

Figure 14 shows the model's predictions for a footprint of 14 and the fluorescence data. The data is less linear ($R^2 = 0.8509$).

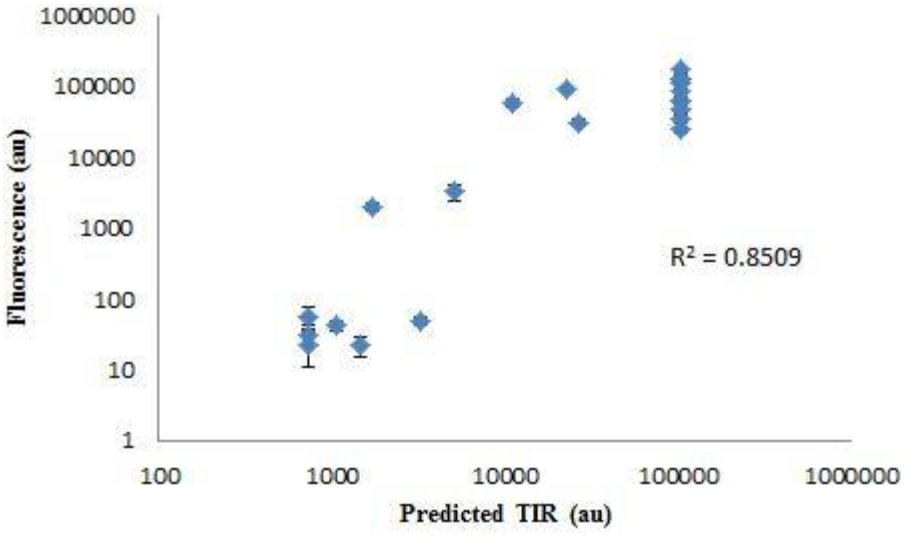



**Figure 14: Linear Relationship between Model Predictions and Fluorescence Data for a Footprint of 14**

The predictions for a footprint of 13 and 14 show a linear trend of TIR with the fluorescence data. For comparison, predictions using a footprint of 35 (the previous model assumption) were plotted in Figure 15. This does not show a linear trend. At this time, a footprint

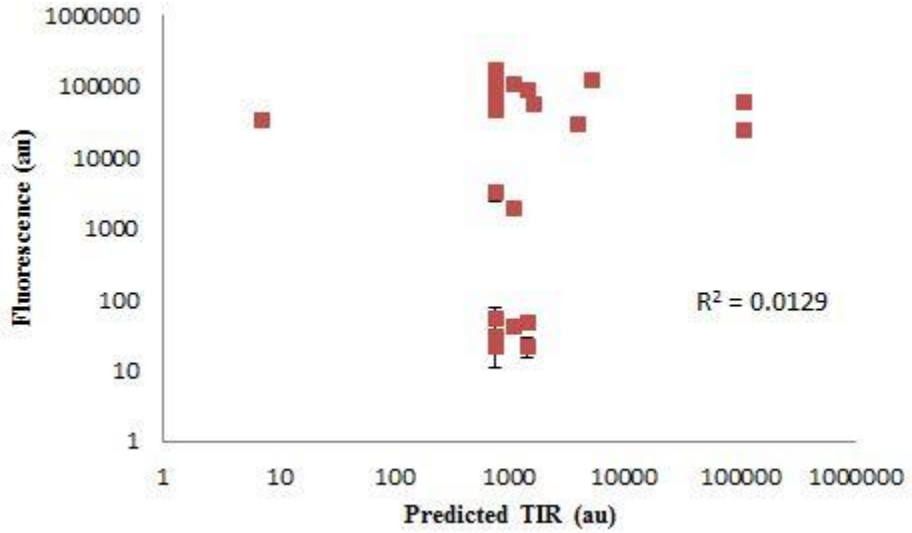



**Figure 15: Nonlinear Relationship between the Model Predictions and Fluorescence Data at a Footprint of 35**

of 13 or 14 will be added to the RBS Calculator parameters to improve the model's predictions. More hairpins and sequences must be evaluated to determine the exact number of the footprint.

**Ribosome requires additional free energy input to accommodate hairpins in the footprint**

In the absence of helicase activity, the ribosome must input more free energy during translation initiation to unfold hairpins near the footprint region. The ribosome can only unfold these hairpins by random, ratchet motion and lacks the catalytic activity of the helicases until elongation is started. The data for hairpins very close to the start codon, which are within the footprint, have a much lower expression level than what the model predicts by using the folding energy of the hairpin (Figure 16). These are highlighted in red.

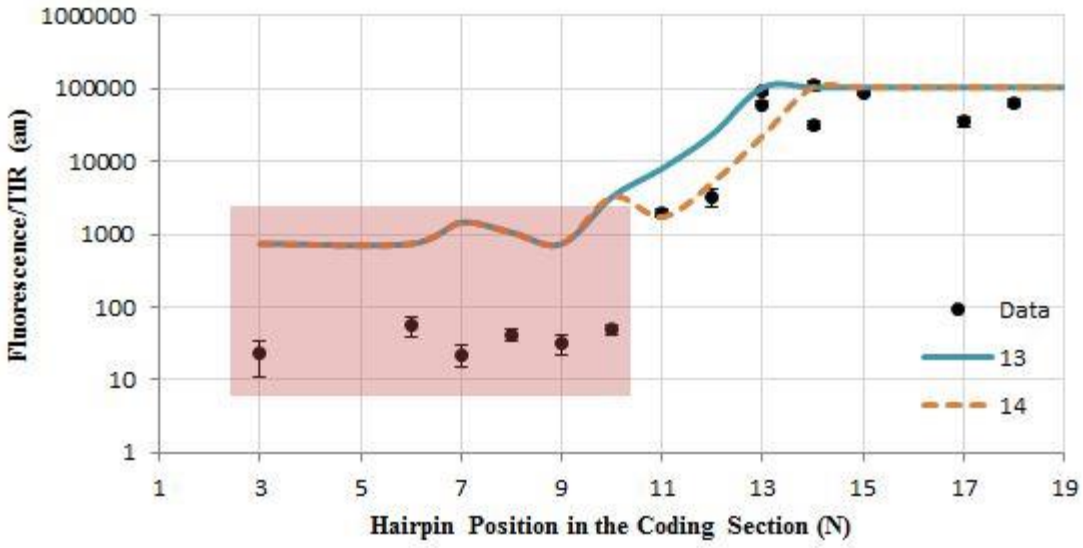



**Figure 16: Disparity between the Model Predictions and the Data for Coding Section Hairpins within the Footprint Region**

The disparity in expression could be the result of mRNA degradation. On freely translated mRNA's, the succession of ribosomes traveling along the mRNA and sequestering the ribosome binding site protect it from RNase degradation. The mRNA's with hairpins in the footprint region would lack ribosomal protection because the ribosome could remain stalled at the footprint region while trying to load the mRNA into its channel. The model assumes the initial and final states of the two-state system have reached equilibrium. However, mRNAs with

secondary structure in the footprint may exhibit intermediate states with a long half-life. In this region, the two-state equilibrium assumption may not be valid (Salis 2011).

The difference in prediction and data suggests there is a non-specific loading energy of approximately 7.7 kcal/mol required for loading the mRNA into the channel and starting translation. This is the average error for the constructs with footprints 3 through 10. From this data, we introduce a new term, $\Delta G_{load}$, which represents the equivalent energy needed to load the mRNAs before being degraded. This could also represent the equivalent energy that is represented by helicase activity after elongation starts.

# Chapter 5: Discussion and Conclusions

The results demonstrate the original hypothesis that protein production will gradually increase as the hairpin moves closer to the footprint edge, and reaches a threshold level at the footprint and beyond. The ribosome is expected to unfold part of the hairpin close to the footprint and bind to the mRNA some of the time, resulting in hindered, but not completely repressed translation, as seen at a footprint length of 10. Translation becomes unrepressed at a footprint length of 13. This data reveals that the footprint is actually smaller than previous predictions, falling into the 13 to 14 nucleotide range. Using this footprint range will improve the RBS Calculator model by making predictions more accurate and synthetic RBS designs that more closely reflect cellular translation.

However, the footprint range can still be narrowed, and more experimentation is needed to pinpoint the footprint length. Hairpins with a larger stem length or stronger folding energy will need to be tested to finalize the length of the footprint, and more experimentation needs to be done to determine how many nucleotides the ribosome unfolds when it binds to the mRNA. Additional RT-qPCR experiments will be helpful to determine the level of mRNAs present in the cell when the hairpin is very close to the start codon. This will help uncouple the mRNA degradation from ribosome stalling. Amin Espah Borujeni is continuing to investigate the interactions during translation initiation.

The data suggest that the ribosome follows the thermodynamic minimization principle. When the ribosome lacks the catalytic helicase activity, it must expend greater free energy to unfold secondary structure by random, ratchet motions. When presented with a strong hairpin, the energetic cost of unfolding the hairpin is too great, and it is unlikely that the hairpin will be

unfolded. Translation remains low. But when the hairpin is weaker, with A:T mutations, the ribosome expends less free energy unfolding it, and protein expression increases. The mechanism of interaction is most likely more complicated, and the ribosome could use other avenues that result in a smaller energetic penalty than unfolding a hairpin, such as compressing to fit a certain amount of nucleotides. More data will be needed to determine these effects.

Due to the lack of helicase activity during initiation, it is apparent that the ribosome must expend additional energy to load mRNA with secondary structure into the channel. This energetic cost greatly reduces protein expression. The model predictions are higher for hairpins within the construct region that what was actually observed. This suggests that the model could be improved by the addition of the free energy of loading the mRNA into the channel, $\Delta G_{load}$. Inclusion of this term in an upcoming version of the calculator would decrease the disparity seen in constructs with secondary structures within the footprint region. From the data, we determined that the $\Delta G_{load}$ is approximately 7.7 kcal/mol, and this could reflect the additional free energy provided by helicase activity when it is present or the energy cost of loading secondary structures into the mRNA channel before the mRNA is degraded. The mRNAs with secondary structure within the footprint region may experience long-lived intermediate states. This data indicates that a more complete model of mRNA existence within the cell is necessary to improve the accuracy of the predictions. The addition of $\Delta G_{load}$ to the calculator will approximate this difference.

Separate from the footprint, the data also reveals that protein expression in *E. coli* can be varied extensively using a specifically designed RBS. Protein expression of RFP was varied 1000-fold between repressed and unrepressed constructs, which is significantly higher than variation found in the literature. This has implications for synthetic biologists looking to fine-tune control of the protein expression level for many industrial and medicinal applications.

# APPENDIX

## Detailed Design Information and Considerations

### Codon Usage in *E. coli*

CODON USAGE IN *E. COLI* GENES[1]

| | Codon | Amino acid[2] | %[3] | Ratio[4] | Codon | Amino acid | % | Ratio | Codon | Amino acid | % | Ratio | Codon | Amino acid | % | Ratio | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| U | UUU | Phe (F) | 1.9 | 0.51 | UCU | Ser (S) | 1.1 | 0.19 | UAU | Tyr (Y) | 1.6 | 0.53 | UGU | Cys (C) | 0.4 | 0.43 | U |
| | UUC | Phe (F) | 1.8 | 0.49 | UCC | Ser (S) | 1.0 | 0.17 | UAC | Tyr (Y) | 1.4 | 0.47 | UGC | Cys (C) | 0.6 | 0.57 | C |
| | UUA | Leu (L) | 1.0 | 0.11 | UCA | Ser (S) | 0.7 | 0.12 | UAA | STOP | 0.2 | 0.62 | UGA | STOP | 0.1 | 0.30 | A |
| | UUG | Leu (L) | 1.1 | 0.11 | UCG | Ser (S) | 0.8 | 0.13 | UAG | STOP | 0.03 | 0.09 | UGG | Trp (W) | 1.4 | 1.00 | G |
| C | CUU | Leu (L) | 1.0 | 0.10 | CCU | Pro (P) | 0.7 | 0.16 | CAU | His (H) | 1.2 | 0.52 | CGU | Arg (R) | 2.4 | 0.42 | U |
| | CUC | Leu (L) | 0.9 | 0.10 | CCC | Pro (P) | 0.4 | 0.10 | CAC | His (H) | 1.1 | 0.48 | CGC | Arg (R) | 2.2 | 0.37 | C |
| | CUA | Leu (L) | 0.3 | 0.03 | CCA | Pro (P) | 0.8 | 0.20 | CAA | Gln (Q) | 1.3 | 0.31 | CGA | Arg (R) | 0.3 | 0.05 | A |
| | CUG | Leu (L) | 5.2 | 0.55 | CCG | Pro (P) | 2.4 | 0.55 | CAG | Gln (Q) | 2.9 | 0.69 | CGG | Arg (R) | 0.5 | 0.08 | G |
| A | AUU | Ile (I) | 2.7 | 0.47 | ACU | Thr (T) | 1.2 | 0.21 | AAU | Asn (N) | 1.6 | 0.39 | AGU | Ser (S) | 0.7 | 0.13 | U |
| | AUC | Ile (I) | 2.7 | 0.46 | ACC | Thr (T) | 2.4 | 0.43 | AAC | Asn (N) | 2.6 | 0.61 | AGC | Ser (S) | 1.5 | 0.27 | C |
| | AUA | Ile (I) | 0.4 | 0.07 | ACA | Thr (T) | 0.1 | 0.30 | AAA | Lys (K) | 3.8 | 0.76 | AGA | Arg (R) | 0.2 | 0.04 | A |
| | AUG | Met (M) | 2.6 | 1.00 | ACG | Thr (T) | 1.3 | 0.23 | AAG | Lys (K) | 1.2 | 0.24 | AGG | Arg (R) | 0.2 | 0.03 | G |
| G | GUU | Val (V) | 2.0 | 0.29 | GCU | Ala (A) | 1.8 | 0.19 | GAU | Asp (D) | 3.3 | 0.59 | GGU | Gly (G) | 2.8 | 0.38 | U |
| | GUC | Val (V) | 1.4 | 0.20 | GCC | Ala (A) | 2.3 | 0.25 | GAC | Asp (D) | 2.3 | 0.41 | GGC | Gly (G) | 3.0 | 0.40 | C |
| | GUA | Val (V) | 1.2 | 0.17 | GCA | Ala (A) | 2.1 | 0.22 | GAA | Glu (E) | 4.4 | 0.70 | GGA | Gly (G) | 0.7 | 0.09 | A |
| | GUG | Val (V) | 2.4 | 0.34 | GCG | Ala (A) | 3.2 | 0.34 | GAG | Glu (E) | 1.9 | 0.30 | GGG | Gly (G) | 0.9 | 0.13 | G |
| | U | | | | C | | | | A | | | | G | | | | |

**Figure 17: Codon Usage in *E. coli***

### Table of Restriction Enzymes

**Table 3: Used Restriction Enzymes and Recognition Sequences**

| Restriction Enzyme | Sequence |
|---|---|
| XbaI | T^CTAGA |
| NdeI | CA^TATG |
| SacI | GAGCT^C |

**pFTVP vector sequence with hairpin at position +3**

TACGTGCCGATCAACGTCTCATTTTCGCCAGATATCCTCGAGCGCGGAATTCC
CTAGGGGATCCGTTGACGGCTAGCTCAGTCCTAGGTACAGTGCTAGCTTCTAGAacccg
ccatatacggcgggacacacacAAGGAGacCATATGGGCGCGattatcCGCGCCaaacagaacaaagaacagatc
accaaacagaacatcaacgcGAGCTCTgaagacgttatcaaagagttcatgcgtttcaaagttcgtatggaaggttccgttaacggtca
cgagttcgaaatcgaaggtgaaggtgaaggtcgtccgtacgaaggtacccagaccgctaaactgaaagttaccaaaggtggtccgctgcc
gttcgcttgggacatcctgtccccgcagttccagtacggttccaaagcttacgttaaacacccggctgacatcccggactacctgaaactgtc
cttcccggaaggtttcaaatgggaacgtgttatgaacttcgaagacggtggtgttgttaccgttacccaggactcctccctgcaagacggtga
gttcatctacaaagttaaactgcgtggtaccaacttcccgtccgacggtccggttatgcagaaaaaaaccatgggttgggaagcttccaccg
aacgtatgtacccggaagacggtgctctgaaaggtgaaatcaaaatgcgtctgaaactgaaagacggtggtcactacgacgctgaagttaa
aaccacctacatggctaaaaaaccggttcagctgccgggtgcttacaaaaccgacatcaaactggacatcacctcccacaacgaagactac
accatcgttgaacagtacgaacgtgctgaaggtcgtcactccaccggtgcttaataaGGCCGGCCgtgctagtgtagatcgctactag
agccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctctactagagtcac
actggctcaccttcgggtgggcctttctgcgtttatacaattgACTAGTCTGCAGaaaaGCGGCCGCGGCATCAAA
TAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGG
TGAACGCTCTCCTGAGTAGGACAAATCCGCCGCCCTAGAGCATGCGCGTTCGGCTGC
GGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGG
GATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTA
AAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACA
AAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAG
GCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCG
GATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTG
TAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACC

CCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCC

GGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAG

CGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACA

CTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAA

GAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTG

TTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATC

TTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTC

ATGCCCGGGTGCTTGGATTCTCACCAATAAAAAACGCCCGGCGGCAACCGAGCGTT

CTGAACAAATCCAGATGGAGTTCTGAGGTCATTACTGGATCTATCAACAGGAGTCCA

AGCGGTACCGATATCAAATTACGCCCCGCCCTGCCACTCATCGCAGTACTGTTGTAA

TTCATTAAGCATTCTGCCGACATGGAAGCCATCACAAACGGCATGATGAACCTGAAT

CGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAAAAC

GGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCA

CCCAGGGATTGGCTGAGACGAAAAACATATTCTCAATAAACCCTTTAGGGAAATAG

GCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAAACTGCCGG

AAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAA

ACGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCTCACCGTCTTTCATTGCC

ATACGAAATTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGAATAAAGGCCGG

ATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAGCTGA

ACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATGTTCTTTA

CGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAG

CTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTA

TTTCATTATGGTGAAAGTTGGAACCTCT

# REFERENCES


Carlson, R. (2009). The changing economics of DNA synthesis. *Nature biotechnology*. **27** (12), 1091-1094.

Chen, H., Bjerknes, M., Kumar, R. and Jay, E. (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res.* **22**, 4953–4957.

Espah Borujeni, A., Channarasappa, A.S., and Salis, H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.***42** (4), 2646-2659.

Farasat, I., Kushwaha M., Collens J., Easterbrook M., Guido M., and Salis H.M., (2014) Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Mol. Syst. Biol.* **v10** (6).

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R. and Hofacker, I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res*. **36**, W70–W74.

Hüttenhofer, A., Noller, H. (1994) Footprinting mRNA-ribosome complexes with chemical probes. *Embo. J*. **13** (16), 3892-3901.

Kozak, M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* **234** (2), 187-208.

Malik, V.S. (1981) Recombinant DNA technology. *Adv. Appl. Microbiol.* **27**, 1-84.

Maloy, S. (2002, July 15). Codon Usage in E. coli. Retrieved November 17, 2014, from http://www.sci.sdsu.edu/~smaloy/MicrobialGenetics/topics/in-vitro-genetics/codon-usage.html

Mironova, R. S., Xu, J., AbouHaidar, M. G., and Ivanov, I. G. (1999) Efficiency of a novel non-Shine-Dalgarno and a Shine-Dalgarno consensus sequence to initiate translation in *Escherichia coli* of genes with different downstream box composition. *Microbiol. Res.* **154** (1), 35-41.

Pertzev, A.V. and Nicholson, A.W. (2006) Characterization of RNA sequence determinants and antideterminants of processing reactivity for a minimal substrate of *Escherichia coli* ribonuclease III. *Nucleic Acids Res*. **34**, 3708-3721.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D., and Gold, L. (1992). Translation initiation in Escherichia coli: sequences within the ribosome-binding site. *Mol.Microbiol.* **6** (9), 1219-1229.

Simonetti, A., Marzi, S., Jenner, L., Myasnikov, A., Romby, P., Yusupova, G., Klaholz, B.P. and Yusupov, M. (2009) A structural view of translation initiation in bacteria. *Cell. Mol. Life Sci.* **66** (3), 423-436.

Salis, H.M. (2011) The Ribosome Binding Site Calculator. *Method Enzymol.* **498**, 19-42.

Salis, H.M., et al. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotech.* **27**, 946-950.

Schuwirth, B. S., Borovinskaya, M. A., Hau, C. W., Zhang, W., Vila-Sanjurjo, A., Holton, J. M., and Cate, J. H. D. (2005). Structures of the bacterial ribosome at 3.5 Å resolution. *Science*. **310** (5749), 827-834.

Takyar, S., Hickerson, R. P., and Noller, H. F. (2005). mRNA helicase activity of the ribosome. *Cell*. **120** (1), 49-58.

Yusupova, G. Z., Yusupov, M. M., Cate, J. H. D., and Noller, H. F. (2001). The path of messenger RNA through the ribosome. *Cell*. **106** (2), 233-241.

# ACADEMIC VITA

Ashlee Smith
Yardley, PA - ashleesmith77@gmail.com

---

## EDUCATION

**B.S. in Chemical Engineering, Bioprocess/Biomolecular Option, Honors**
**Biochemistry and Molecular Biology, Minor**
Graduation date: December 2014
The Schreyer Honors College, Pennsylvania State University, University Park, PA

## WORK EXPERIENCE

**Project Lead, Mentor: Synthetic Biology –** *Dr. Howard Salis, Penn State (January 2014 – Present)*
Leading team for the International Genetically Engineered Machine (iGEM) Competition 2014

- Instructing new undergraduate student in synthetic biology and molecular cloning techniques, including PCR, plasmid purification, electro-competent cell preparation, transformation
- Validating the dCas9 gene knockdown system in *P. putida* to identify genes responsible for furfural toxin catabolism in the production of biofuels from lignocellulose feedstock

**Pharmaceutical Development and Manufacturing Sciences Co-op, API Large Molecule: Janssen Research & Development, LLC., Johnson & Johnson** *(January 2013 – August 2013)*
Worked on cell culture fed-batch process development for monoclonal antibody in PER.C6 human cell line

- Initiated and maintained pre-culture from thaw, scaled-up with shake flasks, continued into wave bags, seed bioreactor within aseptic environment
- Executed bioreactor assembly, sterilization, routine sampling with offline and online measurements for growth and metabolic profile monitoring (BGA, Cedex, NOVA*flex*)
- Performed feed optimization and robustness studies in 5L bioreactors
- Optimized cell counting method using a dissociation reagent to disperse cell culture sample

**Professional Development Chair: J&J Intern and Co-op Association (ICA)**
Organized ICA Career Development Fair for co-ops to interact with senior leadership and maximize career potential through resume reviews, interview tips, and presentations on J&J opportunities

**Laboratory Assistant: Synthetic Biology –** *Dr. Howard Salis, Penn State (May 2012 – January 2013)*

- Determined the ribosome binding site footprint of DH10B *E. coli* cells through recombinant DNA and molecular cloning techniques
- Improved knowledge of ribosome binding for more simplified and complete predictions of protein expression levels

## AWARDS

The Pennsylvania State University's Dean's List *(7/8 semesters)*
The Schreyer Honors College Academic Excellence Scholarship *(2010 – Present)*

## ACTIVITIES

**Shift Leader, Trainer, Volunteer: Centre County PAWS Animal Shelter** *(March 2014 – Present)*
**Exam Proctor: Chemistry Department, Penn State University** *(September 2013 – Present)*
**OXE National Honor Society for Chemical Engineering** *(January 2012 – Present)*
**Secretary: PSU CrossFit Club** *(September 2011 – May 2012)*