THE DEVELOPMENT OF EQUILIBRIUM AND NON-EQUILIBRIUM MODELS OF TRANSLATION AND RIBOSWITCH REGULATION: TOWARDS THE AUTOMATED DESIGN OF CELLULAR SENSORS

A Dissertation in
Chemical Engineering

by
Amin Espah Borujeni

© 2015 Amin Espah Borujeni

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

December 2015
The dissertation of Amin Espah Borujeni was reviewed and approved* by the following:

Howard M. Salis  
Assistant Professor of Chemical Engineering  
Dissertation Advisor  
Chair of Committee

Manish Kumar  
Assistant Professor of Chemical Engineering

Kyle Bishop  
Assistant Professor of Chemical Engineering

Philip C. Bevilacqua  
Professor of Chemistry  
Professor of Biochemistry and Molecular Biology

Janna Maranas  
Professor of Chemical Engineering  
Professor of Materials Science and Engineering  
Graduate Program Coordinator of the Department of Chemical Engineering

*Signatures are on file in the Graduate School
ABSTRACT

Tunable control of protein production inside cells is one of the main challenges in metabolic engineering and pharmaceutical industries. There are many cellular processes that are responsible for altering protein expression levels, including transcription initiation and termination, translation initiation and elongation, and co-translational protein folding. Successful manipulation of each of these processes requires the understanding of their mechanism, and how macromolecular machineries such as RNA polymerases and ribosomes interact with DNAs and mRNAs. In particular, the ribosome’s interactions with mRNA govern its translation rate and the effects of post-transcriptional regulation. It has been shown that translation initiation is a rate limiting step in protein production, in which ribosome must efficiently make the first contact with mRNA. Although translation initiation is one of the well-studied processes inside cells, there are still many unanswered questions that puzzled researchers for several decades. For example, what is the role of 5’ untranslated regions on protein synthesis rate? How the N-terminal coding section controls ribosome binding to mRNA? How the regulatory components such as small RNAs and ligand-responsive riboswitches alter the ribosome binding rate? Can RNA folding kinetics compete with ribosome binding rate to alter the binding equilibrium between mRNA and ribosome?

In this dissertation, we aim to answer these questions by taking a “learn-by-design” approach, in which we first propose a plausible mechanism, and then we conduct specific experiments to test our hypothesis while eliminating other confounding interactions, followed by developing mechanistic models that can explain the experimental observations. Our ultimate goal is to develop simplified yet accurate models that can predict the rate of translation initiation and riboswitch regulation directly from DNA sequence, using kinetics and thermodynamics first
principles. These biophysics-based models can be further improved by identifying new interactions, characterizing their effects, and incorporating them into the existing models.

Overall, through our biophysical modeling approach, we developed an accurate and comprehensive model of bacterial translation initiation that predicts the rate of translation under both equilibrium and non-equilibrium conditions. In addition, we also developed the first sequence-structure-function relationship for translation-regulating riboswitches that predicts the function of riboswitches from the energetic interactions between mRNA, ligand, and ribosome. Importantly, the incorporation of this biophysics-based model into a web-interface called the Riboswitch Calculator, allows the scientific community to design new generation of high-performance biosensors for diverse applications.
TABLE OF CONTENTS

List of Figures .................................................................................................................. x

Acknowledgements .......................................................................................................... xiii

Chapter 1 Introduction ....................................................................................................... 1

Chapter 2 Translation Rate is Controlled by Coupled Trade-offs Between Site
Accessibility, Selective RNA Unfolding, and Sliding at Upstream Standby Sites .... 9
  2.1. Introduction .................................................................................................................. 9
  2.2. Materials and methods .............................................................................................. 12
    2.2.1. Strains, media, and plasmid construction ............................................................... 12
    2.2.2. Fluorescent protein and mRNA measurements .................................................... 13
    2.2.3. Biophysical model calculations ............................................................................ 14
    2.2.4. Measurements of translation rate and binding free energy penalties ............... 17
    2.2.5. Statistical analysis ............................................................................................... 18
  2.3. Results ....................................................................................................................... 18
    2.3.1. Standby site module surface area controls the ribosome’s translation rate ....... 18
    2.3.2. Coupled energetic trade-offs when unfolding structures in standby site
        modules ....................................................................................................................... 23
    2.3.3. Non-cooperative binding to multiple standby site modules .............................. 26
    2.3.4. Distant standby site modules can support high translation rates .................... 27
    2.3.5. The ribosomal platform can slide across upstream structures ......................... 29
    2.3.6. A biophysical model predicts translation rates from diversely structured
        5’ UTRs ..................................................................................................................... 31
    2.3.7. Genome-wide predictions of 5’ UTR translation rates ...................................... 33
    2.3.8. Biophysical modeling to calculate translation rates from split ribosome
        binding sites ........................................................................................................ 36
  2.4. Discussion and conclusion ....................................................................................... 37

Chapter 3 Temperature Regulates Translation Initiation Rate by Altering Both the
Stability of RNA Base-Pairings and the Flexibility of RNA Hairpins ....................... 41
  3.1. Introduction ................................................................................................................ 41
  3.2. Materials and methods ............................................................................................. 43
    3.2.1. Plasmids and strains ........................................................................................... 43
    3.2.2. Growth and characterization .............................................................................. 44
    3.2.3. Biophysical model calculations at different temperatures ............................. 44
    3.2.4. Measurements of translation rate and binding free energy penalties ............ 46
  3.3. Results ....................................................................................................................... 47
    3.3.1. Structured standby sites are less inhibitory at lower temperatures ............... 47
    3.3.2. Lowering temperature reduces the RNA hairpin flexibility and increases
        the single-stranded surface area ........................................................................... 50
    3.3.3. Ribosome follows the thermodynamics minimization principle at lower
        temperatures ....................................................................................................... 52
    3.3.4. Reduced hairpin flexibility makes the sliding process energetically more
        favorable ......................................................................................................... 55
Chapter 4 Reparametrizing the Biophysical Model of Translation Initiation for *Bacillus subtilis* .................................................60

4.1. Introduction ........................................................................60
4.2. Materials and methods ..........................................................61
  4.2.1. Growth, transformation and selection ................................61
  4.2.2. Plasmid constructions .....................................................62
  4.2.3. Fluorescence measurement ..............................................63
  4.2.4. Quantification of spacing and standby free energy penalties ..........64
4.3. Results ................................................................................65
  4.3.1. 30S ribosomes in both *B. subtilis* and *E. coli* experience similar energetic penalty at non-optimal spacing lengths ........................................65
  4.3.2. Investigating the anti-SD sequence of 30S ribosome in *B. subtilis* ......67
  4.3.3. Characterizing mRNAs with structured 5\' UTR in *B. subtilis* ............69
  4.3.4. Ribosome platform in *B. subtilis* is more flexible to accommodate mRNAs with inaccessible standby sites ........................................71
  4.3.5. Modest distortion energy penalty in *B. subtilis* can eliminate the need for partial RNA unfolding ....................................................75
  4.3.6. Ribosome platform in *B. subtilis* slides on mRNA surface with less energy penalty .................................................................77
  4.3.7. Biophysical model predicts the mRNA’s standby energy and translation initiation rate in *B. subtilis* .................................................................79
4.4. Discussion and conclusion .....................................................80

Chapter 5 Precise Quantification of Translation Inhibition by RNA Structures that Overlap with the Ribosome Footprint at N-terminal Coding Sections ..............83

5.1. Introduction ........................................................................83
5.2. Materials and methods ..........................................................86
  5.2.1. Plasmids and strains .........................................................86
  5.2.2. Cloning strategy ...............................................................87
  5.2.3. Growth and characterization ............................................87
  5.2.4. Biophysical model of translation initiation .............................88
5.3. Results ................................................................................90
  5.3.1. Design of expression systems to investigate ribosome footprint ..........90
  5.3.2. RNA hairpins that overlap with ribosomal footprint substantially inhibit translation initiation .........................................................91
  5.3.3. Ribosome footprint unfolds RNA hairpins regardless of their size ..........93
  5.3.4. Determination of ribosome footprint length using biophysical model calculations .................................................................................94
  5.3.5. Overlapping RNA hairpins inhibit translation initiation according to their folding free energy .........................................................97
5.4. Discussion and conclusion .....................................................100
Chapter 6 Automated Physics-Based Design of Synthetic Riboswitches from Diverse RNA Aptamers ................................................................. 103

6.1. Introduction ................................................................................. 103
6.2. Materials and methods ................................................................. 105
   6.2.1. Plasmids and strains ............................................................... 105
   6.2.2. Ligand, media, and buffer conditions ........................................ 107
   6.2.3. Growth, fluorescence, and luminescence measurements ................. 107
   6.2.4. Cell-free transcription-translation assays and luminescence measurements ................................................................. 109
   6.2.5. Transcription-translation decoupling experiment for TMR riboswitches .................................................................................. 109
6.3. Results .......................................................................................... 110
   6.3.1. A biophysical model of translation-regulating riboswitches .................. 110
   6.3.2. Initial validation of the biophysical model ....................................... 112
   6.3.3. Riboswitch stability and concentration-dependence .......................... 114
   6.3.4. Automated design of synthetic riboswitches ................................... 116
   6.3.5. Validating model-predicted requirements for riboswitch activation ......... 119
   6.3.6. Design of dopamine, thyroxine, and DNT-binding riboswitches .......... 122
   6.3.7. Co-transcriptional folding is required for riboswitch activation ........... 125
   6.3.8. Macromolecular crowding controls riboswitch activation .................... 125
   6.3.9. Model-predicted limits of riboswitch-based sensors ............................ 129
6.4. Discussion and conclusion ............................................................... 129

Chapter 7 Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism .................................................. 133

7.1. Introduction ................................................................................. 133
7.2. Materials and methods ................................................................. 134
   7.2.1. Plasmids and strains ............................................................... 134
   7.2.2. Growth and characterization ..................................................... 134
   7.2.3. Design of the expression system ............................................... 135
   7.2.4. Predicting the RNA hairpin folding time using Kinfold calculations .... 137
   7.2.5. Biophysical model of translation initiation ..................................... 138
   7.2.6. A Markov model for Ribosome Drafting mechanism ....................... 140
7.3. Results .......................................................................................... 141
   7.3.1. A Ribosome Drafting mechanism for controlling translation initiation ...... 141
   7.3.2. Validating the Ribosome Drafting mechanism .................................. 147
   7.3.3. The effect of 30S binding rate on the Ribosome Drafting ..................... 149
   7.3.4. The effect of RNA hairpin’s folding times and folding free energies on the Ribosome Drafting .......................................................... 153
7.4. Discussion and conclusion ............................................................... 158

Chapter 8 An Integrated Non-Equilibrium Model of Translation Initiation .......... 160

8.1. Introduction ................................................................................. 160
8.2. Materials and methods ................................................................. 163
   8.2.1. Plasmids and strains ............................................................... 163
   8.2.2. Growth and characterization ..................................................... 164
   8.2.3. Equilibrium-based biophysical model of translation initiation ............. 164
8.2.4. A Genetic Algorithm optimization for model identification ..................165
8.2.5. *in silico* determination of RNA folding kinetics using Kinfold program ......166
8.2.6. Coarse-graining the Kinfold trajectories using a four-state RNA folding model .................................................................167
8.2.7. An integrated non-equilibrium model of translation initiation ..................168
8.3. Results ..................................................................................................174
8.3.1. Ribosome Drafting controls the rate of translation initiation when RNA hairpins are located within N-terminal coding section ......................174
8.3.2. The extent of Ribosome Drafting is reduced for N-terminal RNA hairpins ..179
8.3.3. Effect of Ribosome Drafting on RNA hairpins with two folding transitions .............................................................................182
8.3.4. A four-state model of RNA folding .....................................................184
8.3.5. An integrated non-equilibrium model of translation initiation ..................186
8.3.6. Non-equilibrium model more accurately predicts the rate of translation initiation than the equilibrium model ........................................192
8.4. Discussion and conclusion ..................................................................195

References ..................................................................................................200

Appendix A Supplementary Information for Chapter 2 ..................................219
A.1. Supplementary Figures .........................................................................219
A.2. Supplementary notes .........................................................................225
A.2.1. Ribosomal platform interactions with structured standby sites control translation initiation rate .................................................225
A.2.2. The relationship between ribosomal platform binding and mRNA stability ...........................................................................226
A.2.3. Branched structured 5’ UTRs control the ribosomal platform’s binding according to their effective hairpin height ................................227

Appendix B Supplementary Information for Chapter 6 ..................................229
B.1. Supplementary Figures .........................................................................229
B.2. Supplementary notes .........................................................................247
B.2.1. Biophysical model of riboswitches to control translation initiation ..........247
B.2.2. Thermodynamic minimization model to capture the effect of ligand and mRNA concentration on riboswitch’s function .........................253
B.2.3. Riboswitch Calculator: automated design of synthetic riboswitches .....255
B.2.4. Incorporation of pseudoknot structural constraint into Riboswitch Calculator ....................................................................................257
B.2.5. Combinatorial library and high throughput screening method ..............259
B.2.6. Volume and intracellular concentration of mRNA and ligands ..........259
B.2.7. The mRNA and ligand level comparison between *cell-free* and *in vivo* systems for mRNA dose response experiments ......................261
B.2.8. Designing a “perfect” riboswitch .......................................................263
B.2.9. Examples of data analysis (DNT, dopamine, and thyroxine riboswitches) ...264
LIST OF FIGURES

Figure 2-1. Structured 5’ UTRs control mRNA’s translation initiation rate..........................11
Figure 2-2. Shape and geometry of 5’ UTR structures control ΔG_distortion..........................20
Figure 2-3. 30S platform partially unfolds 5’ UTR structures only to minimize ΔG_standby ....25
Figure 2-4. 30S platform can bind distant standby sites and slide on mRNA surface toward the initiation region.................................................................28
Figure 2-5. Biophysical model predictions for diversely structured 5’ UTRs. .........................32
Figure 2-6. Predicted ΔG_standby across Escherichia coli MG1655 genome .......................34
Figure 2-7. ΔG_standby for split ribosome binding sites ....................................................37
Figure 3-1. The geometry of standby site controls translation initiation rate. .......................48
Figure 3-2. RNA hairpin flexibility at different temperatures controls the distortion energy penalty .........................................................................................................52
Figure 3-3. Ribosome follows a thermodynamic minimization principle at all temperatures. ..................................................................................................................55
Figure 3-4. Reduced RNA flexibility at lower temperatures facilitates the ribosome’s sliding motion. ..........................................................................................................57
Figure 3-5. Biophysical model predicts the standby energy penalty at different temperatures. ..................................................................................................................58
Figure 4-1. Characterization of the spacing energy penalty in B. subtilis............................67
Figure 4-2. 30S ribosomes in both E. coli and B. subtilis have similar anti-SD sequence. ....69
Figure 4-3. Comparison of mRFP1 expression levels in two bacterial hosts.......................71
Figure 4-4. Available single-stranded surface area controls the distortion energy penalty in B. subtilis .........................................................................................................73
Figure 4-5. Investigating the partial RNA unfolding in B. subtilis.........................................77
Figure 4-6. Ribosome platform in B. subtilis efficiently slides on mRNA surface. ...............78
Figure 4-7. Biophysical model predictions in B. subtilis......................................................79
Figure 5-1. N-terminal secondary structures inhibit translation initiation by overlapping with 30S ribosome footprint.................................................................84
Figure 5-2. Position of N-terminal secondary structures regulates the extent of translation inhibition. .......................................................... 92

Figure 5-3. Determination of ribosome footprint length .......................................................................................... 95

Figure 5-4. RNA hairpin’s folding free energy controls translation inhibition .................................................. 99

Figure 6-1. A biophysical model of riboswitch regulation .................................................................................. 111

Figure 6-2. Biophysical model calculations .................................................................................................. 113

Figure 6-3. Automated design of synthetic riboswitches .............................................................................. 118

Figure 6-4. *in vivo* and cell-free characterization of designed riboswitches ........................................... 123

Figure 6-5. Volume exclusion controls riboswitch function inside cells .................................................. 128

Figure 6-6. A “perfect” riboswitch to test the limits of sensing ................................................................. 130

Figure 7-1. A Ribosome Drafting mechanism for controlling translation initiation .................................. 142

Figure 7-2. A large sub-class of long and slow-folding mRNAs shows higher than predicted translation rates. .................................................................................. 144

Figure 7-3. A slow-folding mRNA is largely under-predicted by the thermodynamic model .......................................................... 145

Figure 7-4. Demonstrating the Ribosome Drafting mechanism .................................................................. 148

Figure 7-5. Designed RNA hairpins follow a single folding transition ................................................. 149

Figure 7-6. The extent of Ribosome Drafting is controlled by the 30S binding rate .................................. 151

Figure 7-7. Sequence and structure of designed RNA hairpins to test the effect of 30S binding rate .................................................................................. 152

Figure 7-8. Thermodynamic model predictions for mRNAs with fast-folding RNA hairpins .......................................................... 153

Figure 7-9. Sequence and structure of designed RNA hairpins to test the effect of folding free energy (the case of fast-binding ribosome) ................................................................................ 155

Figure 7-10. Sequence and structure of designed RNA hairpins to test the effect of folding free energy (the case of slow-binding ribosome) ................................................................................ 156

Figure 7-11. Ribosome Drafting’s impact on translation rate is controlled by the RNA hairpins’ folding times and folding free energies ................................................................................ 157

Figure 7-12. Non-equilibrium Markov model solutions are close to thermodynamic model predictions when 30S binding rate is slow ................................................................................ 158
Figure 8-1. Design and characterization of N-terminal RNA hairpins with diverse folding kinetics. ................................................................. 176

Figure 8-2. Sequence and structure of designed RNA hairpins to demonstrate Ribosome Drafting for N-terminal RNA hairpins. ......................................................... 177

Figure 8-3. Ribosome Drafting’s impact on translation rate is controlled by the folding free energy of N-terminal RNA hairpins. ......................................................... 178

Figure 8-4. Extent of Ribosome Drafting is reduced for N-terminal RNA hairpins. .......... 181

Figure 8-5. Translation initiation rates for RNA hairpins with two folding transitions......... 183

Figure 8-6. A four-state, coarse-grained model of RNA folding ........................................ 185

Figure 8-7. A non-equilibrium model of translation initiation ............................................. 190

Figure 8-8. Monitoring the elongation of 70S ribosomes during translation initiation. ........ 192

Figure 8-9. Non-equilibrium Markov model accurately predicts the rate of translation initiation ................................................................. 194

Figure A-1. Translation initiation begins with ribosomal platform’s binding to structured 5’ UTRs.................................................................................................................. 219

Figure A-2. Effect of structured 5’ UTR on mRNA stability and translation initiation rate .................................................................................................................. 220

Figure A-3. Biophysical model predicts the ribosomal platform’s binding free energy penalty to branched structured 5’ UTRs. ................................................................. 221

Figure A-4. Ribosomal platform unfolds mRNA structures based on hairpin sequence and energy ........................................................................................................ 222

Figure A-5. Examples of synthetic structured 5’ UTRs designed to characterize the interactions between ribosomal platform and structured 5’ UTR. ................................. 223

Figure A-6. Analysis of distribution of error in biophysical model predictions ............... 224

Figure B-1. Thermodynamic equilibrium in a simplified intracellular environment ........ 229

Figure B-2. Schematic of optimization algorithm ................................................................. 230

Figure B-3. The biophysical model can predict the output of a combinatorial library ........ 231

Figure B-4. The switching free energy of 52 designed riboswitches .................................... 232

Figure B-5. The biophysical model generates riboswitches with diverse sequences .......... 233

Figure B-6. Secondary structure of a fluoride riboswitch in the OFF and ON states .......... 234
Figure B-7. Secondary structure of a TMR riboswitch in the OFF and ON states. .............235

Figure B-8. The designed riboswitches showed continuous activation during growth.........236

Figure B-9. Predicted AR_{max} versus measured AR. ..........................................................237

Figure B-10. Predicted AR_{conc} versus measured AR for stable and metastable riboswitches. .............................................................................................................238

Figure B-11. Error distribution analysis for model predictions. .............................................239

Figure B-12. Predicted versus measured translation initiation rates and activation ratios for individual aptamers. .................................................................................240

Figure B-13. Correlation between errors in predicted translation initiation rate ON and translation initiation rate OFF. .................................................................241

Figure B-14. Model errors versus individual Gibbs free energy terms.................................242

Figure B-15. The measured activities of all 62 designed riboswitches.................................243

Figure B-16. Secondary structure of a dopamine binding aptamer.........................................245

Figure B-17. Relative gene expression from different promoters and plasmid copy numbers. .........................................................................................................................245

Figure B-18. Comparison of riboswitch growth rate and expression levels at different mRNA levels..............................................................................................................246

Figure B-19. Accepted and unaccepted types of pseudoknot base pairings in the biophysical model. ..................................................................................................................247
ACKNOWLEDGEMENTS

My utmost gratitude goes to my advisor, Dr. Howard M. Salis, for introducing me to the Synthetic Biology world and giving me the opportunity to work in this cutting-edge field of science. I greatly appreciate his continuous support, guidance, and encouragement from the beginning of my PhD studies to the completion of this dissertation. His deep scientific knowledge, excellent mentorship, detailed emphasis on improving the technical writing, data analysis, presentations, and communication skills were central toward achieving my academic goals. I am grateful to him for teaching me how to conduct high quality research and I will be honored to work with him again in the future.

I would like to extend my special thanks to Drs. Kyle Bishop, Manish Kumar, Costas D. Maranas, and Philip C. Bevilacqua for agreeing to serve on my doctoral committee. In particular, I would like to thank Dr. Maranas and Dr. Bevilacqua for their support and advice during the completion of my degree. My sincere appreciation goes to Dr. Ali Borhan, for being both, a mentor and a friend, and giving me constant support and advice during this journey.

I am grateful to the past and present members of Salis group, including Dr. Iman Farasat (whom I started my PhD with and learned a lot from), and also Dr. Manish Kushwaha, and Dr. Jason Collens for their invaluable advices and pleasant discussions. Thanks to Tian Tian, Chiam Yu Ng, and Long Chen for keeping the lab environment a learning experience but also fun. I also thank Emily Dong and Walker Huso for their hard work, detailed attention, and assistance in my research.

This journey has been very enjoyable, thanks to my many friends including Ali, Pouria, Mahsa, Bardia, Sam, and Bijan. I also give special thanks to my dearest friend and partner Lymaris, for her constant support and love, and all the fun times and trips that we had together.
I wish to express my special thanks to my dearest uncle Abbas, for always showing me the right path toward the success. His advise, wisdom, vision, and support were crucial for me to go through difficult times in my life. I also thank my dear brother Ehsan, who is also my best friend, for being my support and helping me going through this process together; having him here made being-away-from-family less difficult. I wish to thank my brother Arash and my sister Nasim for their support and love. Even being thousands of miles away, a phone call always brought joy and happiness. Finally, I express my ultimate gratitude to my parents Nasrin and Houshang, for their unconditional love and support. Without their helps and sacrifices I simply couldn’t become the person who I am. There are no other two people in this world that I can love and respect more.
Chapter 1

Introduction

Translation initiation is a rate controlling step in protein synthesis, in which 30S ribosome subunit binds to the initiation region of an mRNA molecule and forms the 30S initiation complex, followed by recruitment of 50S ribosome subunit and peptide elongation by 70S. The mechanisms by which a ribosome interacts with mRNA have been extensively studied for the past several decades (1–31), making the translation initiation step one of the most studied processes inside the cell. Many interactions contribute to the efficiency of ribosome binding, including secondary structure formation around the initiation site, codon rarity at the N-terminal coding section, kinetics of tRNA binding, assembly and release of initiation factors, and ribosome translocation rate (1, 8, 14, 17). In particular, the most important interaction between mRNA and 30S ribosome is the formation of hydrogen bonds between the last nine nucleotides of 16S ribosomal RNA and a colloquially known region of mRNA called Shine-Dalgarno (SD) (8). After mRNA has been stabilized, 30S ribosome anchors it to its peptide-elongation pocket by base-pairing the mRNA’s start codon (mostly AUG) to the anti-codon loop of tRNA^Met (22, 23, 31). However, if the nucleotide distance between SD-antiSD hybridization region and the start codon is too long or too short, 30S ribosome undergoes an unfavorable stretching or compression, respectively (4, 21), which reducing the ribosome binding affinity to mRNA. mRNA secondary structure formation is another crucial interaction during translation initiation, which inhibits the 30S ribosome from hybridizing to the SD region (8, 24). Ribosome has an elaborate helicase center that unfolds any secondary structure that it encounters during elongation step (27).
However, since 30S ribosome lacks this helicase activity during initiation, unfolding RNA structures becomes energetically costly for 30S ribosome. Therefore, 30S ribosome tends to minimize its unfolding activity during initiation step, and if necessary only partially unfolds RNA structures (12).

Although translation initiation is considered the most studied cellular process, there are still many unanswered questions that have puzzled researchers for several decades. For example, what is the role of 5’ untranslated regions on translation initiation? Most bacterial genomic operons have long 5’ UTRs with complex secondary structures. How ribosome distinguishes between these 5’ UTR structures and structures that form around the initiation site? Are these 5’ UTR structures as inhibitory as the structures around the initiation site? Similar question can be asked for the structures that form downstream from the start codon and inside the coding section. Although it has been shown that the N-terminal coding sections of most bacterial genes have evolved to minimize secondary structures (1), it is still unclear how ribosome deals with these structures inside the coding section, especially since the function of N-terminal structures is strongly coupled with the effect of N-terminal codon sequences. The folding kinetics of RNA structures is another essential but forgotten components that can significantly control translation initiation process. Although it was shown, about two decades ago, that the rate of RNA folding can largely alter the ribosome binding affinity to mRNA (32, 33), the role of RNA folding kinetics has remained mostly unexplored. This is particularly important, because similar to other cellular processes, translation initiation is assumed to take place under equilibrium conditions. This assumption is inherently incorrect, because the initiation process operates under continuous cycles of mRNA unfolding, 30S ribosome binding, and mRNA refolding, which can potentially partition toward non-equilibrium conditions. Finally, not only mRNA itself can control the rate of translation initiation, other regulatory components such as small RNAs and ligand-responsive riboswitches can alter mRNA-ribosome interactions as well. However, their direct role is not
quite understood yet. In particular, it is not clear how the ligand binding to mRNA can alter the energetic cost of ribosome to unfold mRNA structures.

This long list of challenges and unanswered questions are the inspirations for the works in this dissertation. Although there are different ways to approach these problems, we feel that a “learn-by-design” approach combined with biophysical modeling is an efficient method to study biological processes. One main reason is that biological systems are complex and without a careful design of experiment and elimination of confounding interactions, one cannot elucidate the true mechanisms governing a system of interest. Altogether, in the following eight chapters, we will explain how we use biophysical modelling and systematic design of experiment to discover some of the important and exciting interactions that govern translation initiation process in a wide range of bacterial species.

In the next chapter of this dissertation, we study the role of 5′ untranslated regions of mRNAs (also called 5′ UTRs) on translation initiation rate. Long, structured 5′ UTRs are commonly found in bacterial mRNAs, though the physical mechanisms that determine how the ribosome binds these upstream regions remain poorly defined. Here, we systematically investigate the ribosome’s interactions with structured standby sites, upstream of Shine–Dalgarno sequences, and show that these interactions can modulate translation initiation rates by over 100-fold. Our extensive data analysis elucidated that an mRNA’s translation initiation rate is controlled by the amount of single-stranded surface area, the partial unfolding of RNA structures to minimize the ribosome’s binding free energy penalty, the absence of cooperative binding, and the potential for ribosomal sliding. We report the development of a biophysical model employing thermodynamic first principles and a four-parameter free energy model to accurately predict the ribosome’s translation initiation rates for 136 synthetic 5′ UTRs that contain large structures, diverse shapes and multiple standby site modules.
In the third chapter, we investigate how temperature can exclusively alter the ribosome interactions with the structured standby sites. Microorganisms have several mechanisms that can sense and response to temperature changes in their environment. RNA-based temperature regulation is one of the wide-spread mechanisms that control translation initiation by either stabilizing or destabilizing the inhibitory mRNA structures that sequester Shine-Dalgarno (SD) and start codon. However, it is not clear how the non-inhibitory structures upstream of SD are regulated by temperature. In this work, we characterized the function of 46 synthetic 5’ UTRs that have diverse sequences and geometries inside E. coli cells and under a wide range of temperatures (25, 30, and 37 °C). We discovered that in addition to stabilizing the RNA base-pairings, temperature controls translation initiation rate by altering the flexibility of RNA hairpins. At lower temperatures, RNA hairpins at standby sites have limited flexibility and can only partially bend over the neighboring single-stranded regions, increasing the available surface area for binding to the ribosome platform. This reduced RNA hairpin flexibility facilitates the ribosome’s sliding motion, and also eliminates the necessity for unfolding RNA hairpins to increase available surface area. Importantly, we showed that Ribosome still follows the thermodynamic minimization principle even at lower temperatures. Overall, our findings introduce a new mechanistic role for temperature in controlling gene expression, which can be applied to other cellular processes whereby RNA hairpin’s geometry controls its function.

In the fourth chapter, we investigate whether or not the same biophysical rules that govern the function of 5’ UTRs in E. coli (best-studied gram-negative bacteria) can be applied to the gram-positive bacteria as well. Bacillus subtilis is one of the best-studied gram-positive bacterial hosts, widely used in industrial applications for its ability to secret proteins. Although the efficiency of translation initiation in B. subtilis has been extensively compared with its gram-negative counterpart Escherichia coli, a detailed quantitative study is needed to identify the exact differences between their translation initiation mechanisms. In this work, by characterizing 47
mRNA sequences with diverse structured 5’ UTRs and spacing lengths, and by using biophysical model calculations, we discovered that the flexibility of 30S ribosome plays a major role in distinguishing the translation initiation mechanisms within each organism. Although *E. coli* 30S ribosome is more flexible in binding mRNAs with non-optimal spacing lengths, *B. subtilis* 30S ribosome is highly efficient in accommodating structured 5’ UTRs on its platform surface. These findings could explain why *B. subtilis* is a model organism for studying riboswitch regulation, as most of the natural riboswitches have been discovered in its genome. Moreover, our biophysical model calculations can be used to predict the potential consequences of horizontal gene transfer between these two microorganisms.

In the fifth chapter, we study the direct effect of N-terminal coding section on translation initiation. In contrast to the 5’ UTR section, the mRNA’s N-terminal coding sequence controls protein synthesis rate through two coupled interactions: inhibition of translation initiation by N-terminal secondary structure, and reducing translation elongation by N-terminal rare codons. Recently, it has been shown that secondary structures have more profound effect on protein synthesis than codon rarity. However, it is still unclear which RNA structure is inhibitory for 30S initiation, and at what distance from start codon RNA structures need to be unfolded. Here, we systematically investigate the direct effect of N-terminal structures, by introducing exclusive RNA hairpins at varying distant from start codon, while selecting only fast codons to maintain efficient elongation. We report an up to 6000-fold sigmoidal increase in translation as RNA hairpins were shifted away from start codon. Only hairpins that overlapped with the ribosome footprint inhibited translation rate, according to their folding free energies. By comparing our experimental data with the predictions of a biophysical model, we determined that the *in vivo* ribosome footprint length is between 14 and 15 nucleotides, which is consistent with previously reported *in vitro* studies. Overall, our results help to understand how coding section regulates
translation rate, enabling the rational design of post-transcriptional regulatory elements that target N-terminal coding section.

In the sixth chapter, we introduce riboswitches as the shape-changing regulatory RNAs that bind chemicals and regulate translation initiation, directly coupling sensing to cellular actuation. Although many natural riboswitches have been found in the genomic 5’ UTRs of bacterial species such as Bacillus subtilis, it remains unclear how their sequence controls the physics of riboswitch switching and activation. We report the development of a statistical thermodynamic model that predicts the sequence-structure-function relationship for translation-regulating riboswitches, characterized inside cells and within cell-free transcription-translation assays. Using the model, we carried out automated computational design of 62 synthetic riboswitches that sense diverse chemicals (theophylline, tetramethylrosamine, fluoride, dopamine, thyroxine, 2,4-dinitrotoluene) and activated gene expression by up to 383-fold. The model explains how aptamer structure, ligand affinity, switching free energy, and macromolecular crowding collectively control riboswitch activation. Our model-based approach for engineering riboswitches quantitatively confirms several physical mechanisms governing ligand-induced RNA shape-change, and enables the development of cell-free and bacterial sensors for diverse applications.

In the seventh chapter, we show how the RNA folding kinetics and non-equilibrium conditions can significantly control translation initiation rate. RNA folding plays an important role in controlling protein synthesis as well as other cellular processes. Existing models have focused on how RNA folding energetics control translation initiation rate under equilibrium conditions, but have largely ignored the effects of RNA folding kinetics. Therefore, we introduce a new mechanism, called Ribosome Drafting, which explains how the ribosome’s binding rate and the mRNA’s folding kinetics collectively control its translation initiation rate. During cycles of translation, Ribosome Drafting takes place whenever fast-binding ribosomes bind to mRNAs
with slow-folding RNA structures, maintaining the mRNA in non-equilibrium states, with an acceleration of protein synthesis. By designing RNA structures with disparate folding kinetics, but equivalent folding energetics, we show that Ribosome Drafting increases protein synthesis rates by over 1000-fold. Overall, we characterized protein synthesis rates from 36 mRNA variants with rationally designed ribosome binding rates, folding kinetics, and folding energetics to validate a non-equilibrium Markov model of translation and to confirm the necessary conditions for Ribosome Drafting. Our computational design, non-equilibrium modeling, and systematic experiments demonstrate a versatile approach for studying how RNA folding kinetics control cellular processes.

Finally, in the eighth chapter, we introduce a comprehensive framework that can predict the rate of translation initiation under both equilibrium and non-equilibrium conditions. Translation initiation has been assumed to function under equilibrium conditions, whereby the thermodynamics of mRNA-ribosome interactions controls its rate. However, we recently showed that RNA folding kinetics can also control translation initiation at non-equilibrium conditions via a Ribosome Drafting mechanism. Therefore, to predict translation initiation rate under both equilibrium and non-equilibrium conditions, we developed an integrated model, in which several molecular interactions such as mRNA folding kinetics, mRNA-ribosome thermodynamics, mRNA loading, ribosome elongation, mRNA degradation, and mRNA transcription are incorporated to predict translation initiation directly from DNA sequence. In particular, we predict RNA folding kinetics from its sequence by coarse-graining the Kinfold predictions into a four-state RNA folding pathway. By enumerating all possible states of mRNA refolding, 30S initiation, and 70S elongation within the first 22 codons of mRNA’s coding section, we developed a non-equilibrium Markov model of translation initiation with 443 distinct states that are connected using 29 specific transition rates. We parameterize the Markov model by identifying 12 unknown variables that minimize the relative error between predicted and
measured mRFP1 expression levels at steady-states, across 119 synthetic mRNA sequences. Overall, our model provides the first steps for building a predictive non-equilibrium model of protein synthesis rate from DNA sequence.
Chapter 2

**Translation Rate is Controlled by Coupled Trade-offs Between Site Accessibility, Selective RNA Unfolding, and Sliding at Upstream Standby Sites**

This chapter has been previously published in *Nucleic Acids Research* (12), and won the 2014 Best Graduate Student Paper Award in Chemical Engineering Department at The Pennsylvania State University. Anirudh S Channarasappa served as a collaborator in this work. All the measurements and model predictions for this work are presented in *Supplementary_Data_2* and can be found online at: [http://salislab.net/files/Espah_Borujeni_ThesisSupplementaryData.zip](http://salislab.net/files/Espah_Borujeni_ThesisSupplementaryData.zip).

2.1. Introduction

Long 5′ untranslated regions (5′ UTRs) are commonly found in bacterial mRNAs, and their interactions with the ribosome play a central role in controlling translation initiation and post-transcriptional regulation (22, 34, 35). Changes in the 5′ UTR sequence, structure or overall shape have been found to alter an mRNA’s translation rate, including when conformational changes are triggered by the binding of proteins, small RNAs or when cis-acting riboswitches bind their ligand (36–40). However, our understanding of the physical mechanisms that determines the ribosome’s ability to bind long 5′ UTRs remains incomplete. While the ribosome’s interactions with Shine–Dalgarno (SD) sequences and nearby inhibitory mRNA structures have been extensively characterized (3, 8, 9, 41) and are well-predicted by biophysical models (21, 42, 43), many natural 5′ UTRs possess further upstream sequences and structures whose interactions
with the ribosome are still poorly defined (Figure 2-1A). In this chapter, we use a learn-by-design approach to systematically investigate the ribosome’s interactions with long, structured 5’ UTRs, and to identify new physical mechanisms that determine the ribosome’s ability to bind and initiate translation.

According to crystal and cryo-EM structures, the 30S ribosomal complex appears to bind upstream 5’ UTRs through its platform domain, which is a positively charged surface that binds non-specifically to mRNA’s sugar phosphate backbone (23, 31, 35, 44). The ribosomal platform first binds an initial landing pad, known as a ‘standby site’ (10) (Figure A-1), followed by coordinated insertion of the mRNA into the Entry channel and the 16S rRNA’s base pairing with SD like sequences to position the start codon over the P-site (31). Ribosomal S1 protein likely plays a role in remodeling mRNA structure, prior to the formation of the 30S-initiation complex, by binding to single-stranded RNA and promoting the unfolding of RNA hairpins and duplexes; however, free energy must be inputted to unfold RNA structures (45). The mRNA and ribosomal platform bind together independently of other participants, including tRNA^{fMet} and the initiation factors (18, 34). Interactions that delay binding to the 5’ UTR can slow the mRNA’s translation initiation rate, including the presence of mRNA structures (24, 44), competitive binding by small RNAs (46) or the absence of a standby site altogether (24). We consider a 5’ UTR as composed of one or more standby site modules, followed by a 16S rRNA binding site (a SD sequence), a spacer region and ended with a protein coding sequence (Figure 2-1B). An mRNA that folds into several hairpins, all located upstream of the SD sequence, will contain several standby site modules.

Thermodynamic modeling has enabled a comprehensive dissection of the ribosome’s interactions with mRNA with clearly defined contributions, quantitatively described by their free energies (Figure 2-1C). The most stringent test of our current understanding is to combine thermodynamic modeling with in silico optimization to design completely novel, non-natural
Figure 2-1. Structured 5' UTRs control mRNA’s translation initiation rate. (A) Natural *E. coli* 5' UTRs have diverse lengths and structures with varying binding free energy penalties to the ribosomal platform. Green and light blue regions represent SD and start codon, respectively. (B) 5' UTRs are separated into multiple standby site modules, followed by a Shine-Dalgarno sequence, spacer region, and a protein coding sequence. (C) A schematic shows the mRNA regions that contact the ribosome in its initial and final states. The ribosomal platform binds to 5' UTRs with a binding free energy penalty $\Delta G_{\text{standby}}$. The sum of all binding free energies, $\Delta G_{\text{total}}$, determines how likely the ribosome binds to a mRNA and initiates translation.
RNA sequences with predicted ribosomal interactions and translation initiation rates. The differences between model predictions and experimental measurements objectively draw the boundaries between our current understanding of RNA interactions and its limitations. A growing body of evidence, from our work and others, indicates the presence of unpredicted interactions between the ribosome and long 5′ UTRs, highlighted as context effects (Figure A-1C, and supplementary notes in Appendix A.2.1) (21, 26, 47–50).

In this chapter, we decipher the biophysical rules governing the ribosome’s interactions with long, structured 5′ UTRs. We systematically alter 5′ UTR sequence, structure and shape to measure the ribosome’s interactions using in vivo reporter assays. We use thermodynamic modeling to quantify how the ribosome binds less accessible 5′ UTRs, accommodates or partially unfolds RNA structure, selects from multiple binding sites, slides across the mRNA and initiates translation. We develop a predictive biophysical model that combines thermodynamic minimization with a four-parameter free energy model to accurately predict the translation initiation rates of 136 long, diversely structured 5′ UTRs. Finally, we demonstrate that the ribosome can bind to distant sites and engage in long-range RNA interactions, providing a mechanism for the observed context effects and new ways for riboswitches and small RNAs to regulate translation.

2.2. Materials and methods

2.2.1. Strains, media, and plasmid construction

Luria-Bertani (LB) media (10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl) was obtained from BD and supplemented with 50 μg/ml chloramphenicol (Sigma-Aldrich). The expression
system is derived from the pFTV1 plasmid (ColEI, Cm$^R$) (21) and encodes a σ70 constitutive promoter (BioBrick #J23100), a synthetic 5′ UTR and a codon-optimized mRFP1 fluorescent protein. Synthetic 5′ UTRs were inserted by standard cloning techniques, utilizing either BamHI and SacI, or XbaI and SacI sites, depending on 5′ UTR length. 5′ UTRs with desired sequences were created by either annealing of complementary oligonucleotides with overhangs, or by PCR assembly of oligonucleotides with corresponding restriction sites. Plasmid variants were cloned by restriction digest of DNA backbone, purification and ligation to DNA fragments with 5′ UTR sequences and transformation of chemically competent Escherichia coli DH10B cells via heat shock. Correct clones were identified by DNA sequencing. All the designed 5′ UTR sequences are presented in the Supplementary Data_2.

2.2.2. Fluorescent protein and mRNA measurements

Fluorescent protein reporter measurements were performed in 96-well format. An initial 96 deep-well plate containing 700 μl LB and 50 μg/ml chloramphenicol was inoculated, from single colonies, with up to 30 different E. coli DH10B cultures in an alternating pattern that excluded the outer wells. Cultures were incubated overnight at 37 °C with 200 rpm orbital shaking. A fresh 96-well plate containing 200 μl LB and chloramphenicol media was inoculated by overnight cultures using a 1:100 dilution. Plates were incubated at 37 °C in a plate-based spectrophotometer (TECAN M1000) with high orbital shaking. OD$_{600}$ measurements were recorded every 10 min. Once a culture reached an OD$_{600}$ of ~0.15, a sample of each culture was transferred to a new plate containing 200 μl phosphate buffered saline [NaCl 137 mM, KCl 2.7 mM, Phosphate buffer 10 mM (Na₂HPO₄/KH₂PO₄ pH 7.4), purchased from OmniPur] and 2 mg/ml kanamycin (Sigma Aldrich) for flow cytometry measurements. Periodic serial dilutions were repeated two more times, maintaining cultures within the exponential growth phase. Three
samples were taken from each culture. The single-cell fluorescence distributions of samples were measured using a Fortessa flow cytometer (BD Biosciences). All distributions were unimodal. The auto-fluorescence distribution of *E. coli* DH10B cells was also measured. The arithmetic mean of each distribution was taken, and the mean auto-fluorescence was subtracted from each sample. The reported fluorescence values in this study are the average of three measurements (*n* = 3).

For mRNA level measurements, single colonies were first grown overnight in 5 ml LB media with chloramphenicol, then diluted to 0.01 OD<sub>600</sub> in fresh media and grown to mid-exponential phase until reaching 1.0 OD<sub>600</sub>, as measured by a cuvette-based spectrophotometer (NanoDrop 2000C). Total RNA was extracted using the Total RNA Purification kit (Norgen Biotek). Treatment with Turbo DNase (Ambion) eliminated genomic DNA contamination. Reverse transcription was carried out with a High Capacity cDNA Reverse Transcription kit (Applied Biosystems), followed by quantitative PCR with a TaqMan probe (5′-ACCTTCCATACGAACTTT-3′), targeting the internal coding sequence for mRFP1 (forward primer 5′-ACGTTATCAAAGAGTTCATGCGTTTC-3′ and reverse primer 5′-CGATTTCGAACTCGTGACCGTTAA-3′) and using a ABI 7300 real-time cycler. A TaqMan probe for 16S rRNA was used as an endogenous control. Triplicate measurements were each performed on two separate RNA extractions. ΔΔC<sub>t</sub> calculations were then averaged for each sample.

### 2.2.3. Biophysical model calculations

The complete biophysical model employs a free energy model to calculate the total binding free energy between the ribosome and mRNA, according to the equation (Figure 2-1C):

\[
\Delta G_{\text{total}} = \Delta G_{\text{mRNA-rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} + \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}}
\]  
(2.1)
Using statistical thermodynamics, we may relate this total Gibbs free energy change to the mRNA’s translation initiation rate, \( r \), according to:

\[
 r \propto \exp\left(-\beta \Delta G_{\text{total}}\right)
\]  

(2.2)

This relationship has been previously validated on 132 mRNA sequences where the \( \Delta G_{\text{total}} \) varied from \(-10\) to 16 kcal/mol, resulting in well-predicted translation rates that varied by over 100 000-fold (21). The apparent Boltzmann constant, \( \beta \), has been measured as \( 0.45 \pm 0.05 \) mol/kcal, which was confirmed in a second study (51).

In the initial state, the mRNA exists in a structured conformation, where its free energy of folding is \( \Delta G_{\text{mRNA}} \) (\( \Delta G_{\text{mRNA}} \) is negative). After assembly of the 30S ribosomal subunit, the last nine nucleotides of its 16S rRNA have hybridized to the mRNA, while all non-clashing mRNA structures are allowed to fold. The free energy of folding for this mRNA–rRNA complex is \( \Delta G_{\text{mRNA–rRNA}} \) (\( \Delta G_{\text{mRNA–rRNA}} \) is negative). mRNA structures that impede 16S rRNA hybridization or overlap with the ribosome footprint remain unfolded in the final state. These Gibbs free energies are calculated using a semi-empirical free energy model of RNA and RNA–RNA interactions (52, 53) and the minimization algorithms available in the Vienna RNA suite version 1.8.5 (54).

Three additional interactions will alter the translation initiation rate. The tRNA\textsuperscript{OMET} anti-codon loop hybridizes to the start codon (\( \Delta G_{\text{start}} \) is most negative for AUG and GUG). The 30S ribosomal subunit also prefers a five nucleotide aligned spacing (4). Non-optimal distances between the 16S rRNA binding site and the start codon lead to an energetic binding penalty. This relationship between the ribosome’s binding penalty (\( \Delta G_{\text{spacing}} > 0 \)) and aligned spacing distance was measured in our previous study (21). Finally, the 5′ UTR binds to the ribosomal platform with a free energy penalty \( \Delta G_{\text{standby}} \), which is calculated according to Equation 2.5 and described in the Results section. A maximally accessible standby site module will bind to the ribosomal
platform with a zero $\Delta G_{\text{standby}}$. Gibbs free energy calculations for all mRNA sequences are listed in the Supplementary Data 2.

The biophysical model automatically decomposes a 5′ UTR into one or more standby site modules using the following definitions. A standby site module contains a single mRNA hairpin, surrounded by two single-stranded regions: an upstream distal binding site and a downstream proximal binding site. The lengths of the distal and proximal binding sites are denoted by D and P, respectively. The hairpin’s height, H, is determined by the number of nucleotides from the hairpin’s base to the mid-point of the hairpin’s loop, including bulges, internal loops and base pairs. The heights of branched hairpins are determined by averaging the branches’ heights (Figure A-3). Hairpins are only considered as components of standby site modules if they are mutually exclusive with the 16S rRNA binding site. A 5′ UTR with multiple upstream hairpins will contain multiple standby site modules. We consider a 5′ UTR without upstream hairpins to contain a single standby site module with zero hairpin height and zero distal site length. Using these definitions, the biophysical model calculates a $\Delta G_{\text{standby}}$ for each standby site module.

The available single-stranded surface area, $A_S$, of a standby site module is determined according to $A_S = 15 + P + D – H$. The key aspect of this equation is that proximal length, distal length and hairpin height have equal and interchangeable effects on the amount of single-stranded surface area. The constant 15 was selected to create a positive metric for single-stranded surface area. Based on data shown in Figure 2-2, increasing the hairpin height beyond 15 nt did not reduce the mRNA translation rate, regardless of the proximal or distal site lengths. Accordingly, the value of H in the equation is not allowed to exceed 15 nt.

The ribosomal platform’s distortion energy penalty is a quadratic function of the available RNA surface area, $A_S$, (Figure 2-2), according to:

$$\Delta G_{\text{distortion}} = C_1 (A_S)^2 + C_2 (A_S) + C_3$$  \hspace{1cm} (2.3)
where $C_1$, $C_2$ and $C_3$ are the fitted parameter values and equal to 0.038 kcal/mol/nt$^2$, $-1.629$ kcal/mol/nt and 17.359 kcal/mol, respectively. An alternative constant in the definition of surface area would alter these coefficients, but the overall relationship between the surface area and the distortion free energy penalty would remain the same.

### 2.2.4. Measurements of translation rate and binding free energy penalties

In our steady-state, long-time growth conditions, translation initiation is a rate-limiting step in codon-optimized mRFP1 fluorescent protein expression; fluorescence levels are proportional to translation initiation rates under our controlled growth conditions (42), and translation initiation rate is related to total binding free energies according to a log-linear formula (Equation 2.2). We measure the change in apparent binding free energy penalty by comparing the fluorescence of a structured 5′ UTR with the fluorescence of a maximally accessible 5′ UTR, according to:

$$\Delta G_{\text{standby}} = -\frac{1}{\beta} \ln \left( \frac{X}{X_{\text{ref}}} \right) - (\Delta \Delta G_{\text{other}}) \quad (2.4)$$

where $X$ is the fluorescence from a sample synthetic 5′ UTR, $X_{\text{ref}}$ is the fluorescence of the unstructured 5′ UTR, and $\beta$ is 0.45 mol/kcal. $\Delta \Delta G_{\text{other}}$ is the difference in free energies, between the sample and reference 5′ UTRs, that depends on other ribosomal interactions and not on the standby site. By using identical and carefully designed 16S rRNA binding sites, spacer regions and coding sequences in the 136 synthetic mRNA sequences, the value for $\Delta \Delta G_{\text{other}}$ was zero for 123 mRNA sequences, and a maximum of 0.90 kcal/mol for the remaining sequences.
2.2.5. Statistical analysis

Two-sample T-test, at two-tailed 5% statistical significance, was performed to compare the error distribution of different sets of model predictions. The null hypothesis assumed that the errors in two different sets of model predictions have same mean and variance. The alternative hypothesis indicates that the errors in two sets follow different distributions. The error in model prediction, ΔΔG_{standby} was calculated according to the formula: \( \Delta \Delta G_{\text{standby}} = \text{measured } \Delta G_{\text{standby}} - \text{predicted } \Delta G_{\text{standby}} \). The low probabilities (P < 0.05) reject the null hypothesis, showing that the differences in two sets of model predictions were statistically significant.

2.3. Results

2.3.1. Standby site module surface area controls the ribosome’s translation rate

We first employed a learn-by-design approach to determine the ribosomal platform’s interactions with standby sites that contain a single hairpin, and therefore contain a single standby site module. Standby site modules are defined by their hairpin height (H), upstream distal binding site length (D) and downstream proximal binding site length (P) (Figure 2-2A; Materials and methods section). We designed, constructed and characterized 62 synthetic 5’ UTRs containing single standby site modules with hairpins from 9 to 18 bp long, distal binding sites from 1 to 15 nt long and proximal binding sites from 1 to 24 nt long. Long mRNA hairpins contain internal loops to avoid potential RNAse activity (55). All mRNA sequences are transcribed by the J23100 constitutive promoter, encode the mRFP1 fluorescent protein, and are designed with identical SD and spacer sequences (Materials and methods section). Flow cytometry measurements from long-time, log-phase cultures with periodic serial dilutions revealed that mRFP1 fluorescence varied
by 221-fold (Figure 2-2CDE). In contrast, mRNA levels varied by at most 5.8-fold, according to RT-qPCR measurements (Figure A-2, and supplementary notes in Appendix A.2.2). Changes in fluorescence are primarily due to changes in translation rate and, in these growth conditions, fluorescence and translation initiation rates are proportional. All sequences and measurement data are listed in the Supplementary_Data_2.

According to the data, standby site module accessibility plays an important role in modulating the ribosome’s interactions with 5’ UTRs and controlling translation rate across a large range. In particular, lengthening the distal binding site by four nucleotides increased the mRNA’s translation rate by 4.2-fold, whereas lengthening the proximal binding site by the same amount increased it by 4.0-fold. Shortening the hairpin’s height by 4 bp increased the mRNA’s translation rate similarly by 3.1-fold. More generally, the slopes (first derivatives) of the log-transformed translation rates all have similar magnitudes (Supplementary_Data_2, column E). Highly accessible standby site modules with long proximal binding sites or short hairpins also cause the mRNA’s translation rate to reach a plateau with the same fluorescence level as an mRNA with an unstructured 5’ UTR (Figure 2-2C). These observations led to the hypothesis that the geometric characteristics of the standby site module (D, P, H) are interchangeable, and that its available single-stranded surface area, A_S, determines the ribosomal platform’s ability to bind to it.
Figure 2-2. Shape and geometry of 5' UTR structures control $\Delta G_{\text{distortion}}$. (A) Standby site modules are defined by a proximal and distal binding site, separated by an RNA structure. (B) Standby site module surface area is calculated by summing the proximal P and distal D binding site lengths, and subtracting the hairpin height H. The surface area is kept positive by adding a constant 15. (C, D, E) Translation rate and ribosome binding free energy penalties are measured as proximal binding site length, distal binding site length, and hairpin height are increased. The $\Delta G_{\text{standby}}$ numbers shown in secondary y-axis are directly related to the data according to Equation 2.4. (C) Hairpin heights are 9 nt (green stars), 12 nt (black asterisks), 15 nt (red squares), or 18 nt (blue circles). (D) Hairpin heights are 15 nt (red squares) or 18 nt (blue circles). (E) Hairpin height varies from 9 to 12 nt by adding to loop length of a 6 bp stem (blue circles) or from 14 to 18 nt by adding to loop length of a 12 bp stem (red squares). (F) 56 characterized 5'
UTRs ($A_S < 24$) from parts CDE are combined to quantitatively relate the standby site module’s surface area to the ribosome’s translation rate and binding free energy penalty. Dashed line is a best-fit quadratic equation. (G) The validity of the quadratic equation is tested by measuring translation rates from 15 additional 5’ UTRs and comparing to model predictions ($R^2 = 0.78$, average error is 0.66 kcal/mol). Data points are the result of three measurements in one day. In parts CDEFG, the horizontal dashed line is the translation rate and ribosome binding free energy penalty of the reference unstructured 5’ UTR.

A standby site module’s single-stranded surface area is the number of nucleotides capable of single-stranded contact with the ribosomal platform. Longer proximal or distal binding sites increase the standby site module’s surface area, while longer hairpin structures reduce the standby site module’s accessibility. mRNA structures can potentially reduce site accessibility by folding over the neighboring single-stranded sites and sequestering binding regions. A standby site module’s available surface area is calculated by summing its proximal and distal binding site lengths and subtracting the hairpin’s height (Figure 2-2B; Materials and methods section). A highly accessible standby site module has a large $A_S$, whereas a less accessible standby site module has a small $A_S$. We then evaluated whether the standby site module’s surface area is the key determinant by comparing $A_S$ to the measured log-transformed translation rates. For all 62 5’ UTRs with diverse geometric features, we found a single quantitative relationship that described the effect of surface area on the mRNA’s translation rate (Figure 2-2F).

We now turn to thermodynamic modeling to quantify the ribosome’s interactions with standby site modules. The competitive binding of free ribosomes to the pool of intracellular mRNAs allows one to use statistical thermodynamics to relate a ribosome’s binding free energy to its translation initiation rate, according to a constant log-linear relationship (Equation 2.2) (21, 42). We calculate the interaction energy between the ribosome and an mRNA’s standby site modules by comparing its translation to the translation rate of a reference, unstructured 5’ UTR,
both with an identical SD sequence, spacer region and protein coding sequence (Equation 2.4).

Using the data in Figure 2-2CDE, we found that the standby site binding free energy penalty, $\Delta G_{\text{standby}}$, varied from 0 to 11.8 kcal/mol when changing its geometric characteristics, and that a simple quadratic function succinctly described the effect of site accessibility on the ribosome’s binding free energy penalty ($R^2 = 0.97$, average error is 0.51 kcal/mol) (Figure 2-2F). A maximally accessible standby site module has a zero binding free energy penalty ($\Delta G_{\text{standby}} = 0$), indicating that the mRNA’s translation initiation rate is not reduced by the ribosomal platform’s interactions. A standby site module with a positive binding free energy penalty ($\Delta G_{\text{standby}} > 0$) has reduced the mRNA’s translation initiation rate by decreasing the ribosomal platform’s ability to initially bind the mRNA.

Next, we critically tested the parameterized model by constructing 15 new 5’ UTRs and measuring their translation initiation rates and ribosome binding free energy penalties. The 5’ UTRs each contain a standby site module with diverse structures, between 43 and 91 nt long, including variable-length distal and proximal binding sites, hairpin heights, and including two- and three-branched structures (Supplementary Data_2). For branched structures, we show that the relevant hairpin height is the average across the multiple branches (Figure A-3, and supplementary notes in Appendix A.2.3). The quadratic relationship accurately predicted the ribosome’s translation rate and binding free energy penalty ($R^2 = 0.78$, average error is 0.66 kcal/mol), confirming that standby site module’s accessibility controls the ribosome’s ability to bind and initiate translation (Figure 2-2G). The quadratic relationship contains three parameters, which are now treated as constants (Equation 2.3; Materials and methods section).
2.3.2. Coupled energetic trade-offs when unfolding structures in standby site modules

Prior to initiating translation, the 30S ribosome does not have an external source of energy. Initiation factor 2 does not hydrolyze GTP until after translation has been initiated, during 50S recruitment (24). Consequently, the 30S complex expends its own free energy to unfold mRNA structures. Previous studies have shown that mRNA structures that sequester the 16S rRNA binding site are unfolded prior to initiating translation (9, 21, 24). The mRNA’s translation initiation rate is reduced according to the free energy penalty for unfolding these structures, as described in Equations 2.1 and 2.2 (Materials and methods section). However, it remains unclear whether mRNA structures located upstream of SD sequences, within standby site modules, are unfolded or accommodated. We next investigated how the ribosome unfolds hairpins in long 5’ UTRs.

By definition, there is a coupled thermodynamic relationship between the unfolding of an mRNA hairpin and increasing the accessibility of the corresponding standby site module, where the accessibility of standby site module controls the ribosomal platform’s distortion energy penalty $\Delta G_{\text{distortion}}$ (see Materials and methods section). Unfolding a single base pair in a module’s hairpin will increase the standby site module’s surface area by three nucleotides, the hairpin’s height decreases by one and the lengths of both the proximal and distal binding sites increase by one. Therefore, unfolding a module’s hairpin will lead to higher accessibility, a lower $\Delta G_{\text{distortion}}$ and a higher translation rate. However, unfolding base pairs requires an input of free energy ($\Delta G_{\text{unfolding}} > 0$), which will lead to a lower translation rate. Consequently, there is an energetic trade-off between unfolding mRNA structures, and increasing the surface accessibility of standby site modules. The ribosome may completely unfold a structured standby site module to maximize its surface area, but the energetic cost may be high (high $\Delta G_{\text{unfolding}}$ penalty with $\Delta G_{\text{distortion}} = 0$). In contrast, the ribosome may fully accommodate a standby site module’s
hairpin, without unfolding it, even though its low surface accessibility will penalize the ribosome’s ability to bind to it (high $\Delta G_{\text{distortion}}$ penalty with $\Delta G_{\text{unfolding}} = 0$). Based on thermodynamic principles, we hypothesize that the ribosome will ‘selectively’ and ‘partially’ unfold structured standby site modules to minimize its total binding free energy penalty ($\Delta G_{\text{standby}} = \Delta G_{\text{distortion}} + \Delta G_{\text{unfolding}}$). Figure 2-3AB presents an illustrative example of this principle and our use of minimization to calculate the ribosome’s binding free energy penalty to a standby site module. A 5’ UTR with a single standby site module initially contains a short proximal binding site ($P = 2$ nt), a short distal binding site ($D = 2$ nt) and a long hairpin ($H = 15$ nt), resulting in a largely inaccessible 5’ UTR ($A_S = 4$). Based on the standby site module’s surface area, the ribosome has an initial distortion energy penalty of 11.5 kcal/mol. However, the ribosome can selectively unfold the hairpin’s closing A:U base pairs, increasing the standby site module’s surface area and lowering its $\Delta G_{\text{distortion}}$ at the expense of increasing its $\Delta G_{\text{unfolding}}$. Unfolding 1 bp increases the standby site module’s surface area by 3 nt, decreases the distortion energy penalty to 7.8 kcal/mol and increases the unfolding energy penalty to 0.90 kcal/mol, leading to a total binding free energy penalty of 8.7 kcal/mol. The ribosome can continue to unfold base pairs to minimize its total binding free energy penalty; unfolding 3 bp at a cost of 2.7 kcal/mol leads to a distortion energy penalty of 2.6 kcal/mol, and a total binding free energy penalty of 5.3 kcal/mol (Figure 2-3B). Two additional examples are shown in Figure A-4. The unfolding free energies are calculated using the Vienna RNA suite with the Turner 1999 RNA parameters (see Materials and methods section). Importantly, thermodynamic minimization does not add any new parameters to the model.

We tested the proposed mechanism by characterizing the translation rate of 22 synthetic 5’ UTRs (Figure 2-3CDE) that contain single standby site modules with varied geometries ($D = 0$ to 7 nt and $P = 1$ to 20 nt) and differing hairpin sequences, sizes and folding energetics (Figure
We employed the minimization principle to calculate the number of unfolded base pairs and to predict the ribosome’s binding free energy penalty, $\Delta G_{\text{standby}}$ (Figure 2-3D). As comparisons, we performed the same calculation while asserting that either no additional energy was inputted to unfold hairpins (high $\Delta G_{\text{distortion}}$ and $\Delta G_{\text{unfolding}} = 0$) (Figure 2-3C) or that hairpins were completely unfolded by the ribosome (high $\Delta G_{\text{unfolding}}$ and $\Delta G_{\text{distortion}} = 0$) (Figure 2-3E).

Figure 2-3. 30S platform partially unfolds 5' UTR structures only to minimize $\Delta G_{\text{standby}}$. (A, B) Competition between ribosomal distortion and selective RNA unfolding is illustrated on a structured 5’ UTR with a weak hairpin. The ribosome's distortion penalty $\Delta G_{\text{distortion}}$ (dashed red line), hairpin unfolding energy (solid black line), standby site module surface area (A$_s$, green line), and fluorescence (red and blue dots) are shown as a function of the predicted $\Delta G_{\text{standby}}$. (C, D, E) The effect of RNA unfolding on fluorescence is shown for no RNA unfolding, partial RNA unfolding, and complete RNA unfolding.
penalty $\Delta G_{\text{unfolding}}$ (dotted-dashed blue line), and their sum $\Delta G_{\text{standby}}$ (green solid line) are shown as the hairpin's closing base pairs are unfolded. The total binding free energy penalty has a local minima when three base pairs are unfolded. See also Figure A-4. (C,D,E) Scatter plots comparing measured translation rates from 22 synthetic 5' UTRs and model predictions to test three mechanisms for the ribosome's unfolding of RNA structures. The average difference between model predictions and measurements is (C) 2.42 kcal/mol when not unfolding structures; (D) 0.90 kcal/mol when partially unfolding structures; and (E) 10.24 kcal/mol when fully unfolding structures. Comparative p-values from two-sample T-tests are 0.004 (C vs. D) and $10^{-13}$ (D vs. E). Data points are the result of three measurements in one day. In parts CDE, the diagonal dashed line is the predicted translation rate according to Equation 2.2.

The data supports our hypothesis that the ribosome partially and selectively unfolds structured standby site modules to minimize its total binding free energy penalty ($\Delta G_{\text{standby}}$), rather than separately minimize its distortion or unfolding penalties (average error is 0.90 kcal/mol when minimizing $\Delta G_{\text{standby}}$, 2.42 kcal/mol when $\Delta G_{\text{unfolding}} = 0$ and 10.24 kcal/mol when $\Delta G_{\text{distortion}} = 0$; comparative p-values are 0.004 and $10^{-13}$, respectively). Remarkably, in all 22 cases, only 1 to 3 bp of each hairpin are predicted by the model to be unfolded, while accommodating the remaining large RNA structures.

### 2.3.3. Non-cooperative binding to multiple standby site modules

We have shown that the ribosome can accommodate large structures of standby site modules and engage in frequent translation, but it is unclear how the presence of multiple standby site modules in a 5’ UTR will affect its translation rate. The presence of multiple standby site modules could allow the ribosome to cooperatively bind mRNA and increase its translation rate. We next characterized nine synthetic 5’ UTRs (76–164 nt long) with up to four standby site
modules to investigate the potential for cooperative binding (Figure 2-4A). If the ribosome can engage in cooperative binding, then the additional standby site modules will increase the overall available surface area, reduce the ribosome’s binding free energy penalty and increase the translation initiation rate. In particular, we expect the greatest cooperativity when the available surface area of a single standby site module is limited.

We found that 5′ UTRs with two, three or four standby site modules were all translated at the same rate, indicating an absence of cooperative binding (Figure 2-4B), including when the available surface areas for individual standby site modules were significantly lowered. Therefore, even when the combined surface areas of multiple standby site modules could potentially increase contact with the ribosomal platform and increase binding, a cooperative increase in translation rate was not observed. Consequently, the data supports a mechanism whereby the ribosome binds to only a single standby site module prior to initiating translation.

2.3.4. Distant standby site modules can support high translation rates

We next investigated whether any standby site module could be bound by the ribosome, even if located far upstream of the start codon. Using the model’s predictions, we forward-engineered multi-standby site module synthetic 5′ UTRs where the most distantly upstream site is the only one capable of supporting high translation rates (Figure 2-4C). Internal standby site modules have long hairpins and limited surface areas (model predictions: $\Delta G_{\text{standby}}$ between 9.0 and 17.4 kcal/mol) compared to the maximally accessible upstream standby site modules. Therefore, if the ribosome binds the farthest upstream standby site module, its translation rate will be at least 12.7-fold higher than binding to the internal standby site modules. We measured fluorescence levels with flow cytometry and mRNA levels with RT-qPCR. Compared to an unstructured 5′ UTR, the mRNA levels decreased by only 1.3- to 5.5-fold, while the fluorescence
levels were correspondingly reduced between 2.5- and 18.3-fold, showing that changes in fluorescence were mainly due to changes in translation rate and ribosome binding free energy penalties (Figure 2-4D).

Figure 2-4. 30S platform can bind distant standby sites and slide on mRNA surface toward the initiation region. (A) Synthetic 5' UTRs with multiple, evenly spaced standby site modules. Spacing N is varied from 4 to 12 nt. (B) Measured translation rates are compared to standby site module multiplicity, showing the absence of cooperative binding to multiple standby site modules, regardless of limiting surface areas. The ΔG_{standby} numbers shown in secondary y-axis are directly related to the data according to Equation 2.4. (C) Synthetic 5' UTRs with one upstream and multiple internal standby site modules, labeled with Roman numbers. (D) The reduction in fluorescence (red bars) and mRNA levels (blue bars) of 5' UTRs M₁ to M₄, compared to an unstructured 5' UTR (see also Figure A-2). (E) Fluorescence measurements of 5' UTRs M₁ to M₄ show that upstream standby site modules can support high translation rates. (F) Model predictions are shown for 5' UTRs M₁ to M₄ using either ΔG_{distortion} (grey bars) or
ΔG\text{distortion} + ΔG\text{sliding} (blue bars) for each standby site module (I to IV, corresponding to labels in part C), compared to the ribosome's apparent binding free energy penalties (dashed lines, yellow region) (average error with and without the sliding penalty is 1.15 and 3.24 kcal/mol, respectively; p-value is 0.013). ΔG\text{distortion} for each standby site module was calculated according Equation 2.3. In parts B and E, the horizontal dashed line is the translation rate and ribosome binding free energy penalty of the unstructured 5′ UTR.

We found that all the mRNAs’ translation rates remained very high, even when the accessible standby site module was located over 100 nt upstream of the start codon (M4 5′ UTR) (Figure 2-4E). In particular, the M2 5′ UTR is translated at a 33.8-fold higher rate when compared to the translation rate that could be supported by binding to the standby site module closest to the start codon. Moving the accessible standby site module upstream, away from the start codon, does not significantly reduce the translation rate. Based on these measurements, the apparent standby site binding penalties are 3.6, 6.2 and 6.5 kcal/mol for M2 to M4 5′ UTRs (Figure 2-4F), respectively, which is much lower than the internal standby site modules’ predicted binding free energy penalties (9.0–17.4 kcal/mol). These measurements suggest that the ribosome will bind to the standby site module that has the lowest binding free energy penalty, regardless of its distance from the start codon.

2.3.5. The ribosomal platform can slide across upstream structures

Several mechanisms would allow the ribosomal platform to bind distant, upstream standby site modules and transition to a pre-initiation complex at the start codon. After the ribosome has bound to the upstream standby site module, the 30S complex could engage in diffusive hopping that would increase its local concentration and assembly rate at the start codon.
The rate of assembly would largely depend on the physical distance between the standby site module and start codon, and not on the presence of RNA structures. Alternatively, the ribosome could employ a full contact, sliding mechanism that reorients the mRNA until a stable pre-initiation complex has formed. The positively charged surface of the ribosomal platform makes it ideal for maintaining contact with a 5′ UTR until 16S rRNA hybridization has anchored the mRNA. Unlike a diffusive hopping mechanism, the presence of mRNA structures would impede this sliding motion and reduce the pre-initiation complex’s assembly rate.

Further analysis of our measurements shows that RNA structures in between the upstream standby site module and start codon partly attenuate the mRNA’s translation rate (Figure 2-4E). The apparent binding free energy penalty of the M2 5′ UTR (3.6 kcal/mol) is higher than the upstream standby site module’s predicted binding free energy penalty (0.4 kcal/mol). This energetic difference cannot be explained by binding to the internal standby site module (11.4 kcal/mol), or by the unfolding of the downstream RNA structure (17.5 kcal/mol). Thus, the free energy difference (3.6 – 0.4 = 3.2 kcal/mol) is significant, but much lower than predicted by alternative binding to other standby site modules or from unfolding of RNA structures.

Our data supports the presence of a sliding mechanism for the ribosomal platform. Intervening mRNA structures are not unfolded, but are pushed aside with an energetic cost. Using our measurements of the M2 5′ UTR, we estimate that 3.2 ± 0.5 kcal/mol free energy is required to push aside the single RNA structure with a height of 15 nt, yielding a coefficient of 0.20 ± 0.04 kcal/mol per hairpin height. We then determined whether including this sliding penalty could improve the model’s ability to predict the translation rate of long 5′ UTRs with multiple standby site modules with ‘hairpins of differing heights’. For each standby site module, we calculate the heights of all downstream RNA structures, and multiplied the sum by the coefficient (c = 0.2 kcal/mol/nt) to yield a free energy penalty of sliding, \( \Delta G_{\text{sliding}} \). The sliding free energy penalties
were 0, 3.0, 4.8 and 7.8 kcal/mol for M1 to M4 5′ UTRs, respectively. We add together the sliding, distortion and unfolding free energy penalties to calculate the ribosome’s binding free energy penalty to each standby site module.

**Figure 2-4E** shows the measured translation initiation rates for M1 to M4, alongside the predicted binding free energy penalties (**Figure 2-4F**), both with and without the sliding free energy penalty. The data shown in **Figure 2-4EF** are on the same scale, according to the log-linear relationship between translation initiation rate and binding free energy differences (Materials and methods section), enabling a comparison between the measurements and the two predictions. The introduction of the sliding free energy penalty enabled the biophysical model to more accurately predict the translation rate from longer 5′ UTRs with several standby site modules (average error is 1.86 kcal/mol with the sliding energy penalty, compared to 4.45 kcal/mol without the sliding energy penalty; *p*-value is 0.013).

### 2.3.6. A biophysical model predicts translation rates from diversely structured 5′ UTRs

Altogether, the ribosome binds to 5′ UTRs by selecting the most geometrically accessible standby site module that requires the least amount of RNA unfolding. Standby site module selection is not limited by distance to the start codon, and upstream standby site modules can support high rates of translation. We identified that ribosomal distortion, selective RNA unfolding and ribosomal sliding control standby site module selection and binding energetics, and combine them together to create a single biophysical model. The ribosome’s binding free energy penalty to the standby site modules in a structured 5′ UTR is calculated according to:

\[
\Delta G_{\text{standby}} = \Delta G_{\text{distortion}} + \Delta G_{\text{unfolding}} + \Delta G_{\text{sliding}}
\]

(1.5)

All three Gibbs free energies are positive, and a long, unstructured 5′ UTR would bind to the ribosomal platform without an energetic penalty (\(\Delta G_{\text{standby}} = 0\) kcal/mol). All RNA energetics
are calculated using a semi-empirical free energy model of RNA and RNA–RNA interactions (52, 53) and the minimization algorithms available in the Vienna RNA suite version 1.8.5 (54). This model extends our previous biophysical model that focused on downstream mRNA interactions, including the 16S rRNA binding site, spacer region and downstream RNA structures (21, 42).

We further tested the accuracy of the biophysical model by characterizing an additional 28 diversely structured 5’ UTRs (68–164 nt long) with two to four standby site modules of varying geometries, RNA structure energetics and hairpin heights (Figure A-5B). The biophysical model was able to accurately predict the relative translation rates and ribosome binding free energy penalties (average error is 0.79 kcal/mol and \(R^2 = 0.83\)) (Figure 2-5A). Overall, the biophysical model contains four, empirically determined parameter values, but can accurately predict the binding free energy penalties and translation rates of the 136 synthetic 5’ UTRs characterized in this study (average error is 0.75 kcal/mol and \(R^2 = 0.89\)) (Figure 2-5B).

**Figure 2-5. Biophysical model predictions for diversely structured 5’ UTRs.** (A and B) A scatter plot showing fluorescence measurements and apparent \(\Delta G_{\text{standby}}\) energy penalties compared to predicted \(\Delta G_{\text{standby}}\) energy penalties for 28 synthetic 5’ UTRs with diverse structures (average error is 0.79 kcal/mol and \(R^2 = 0.83\)). (B) The same comparison for the 136 synthetic 5’ UTRs characterized in this study (average
error is 0.75 kcal/mol, $R^2 = 0.89$). In parts AB, the apparent $\Delta G_{\text{standby}}$ numbers shown in secondary y-axis are directly related to the data according to Equation 2.4. An x=y diagonal line is shown (dashed).

### 2.3.7. Genome-wide predictions of 5’ UTR translation rates

We next employed the biophysical model to examine how standby sites in genomic 5’ UTRs control their translation initiation rates. From the annotated *E. coli* MG1655 genome, provided by EcoCyc (56), we collected 3430 5’ UTRs that varied from short leaders to gene-sized regulatory hot spots with average 5’ UTR length of 100 nt. The 5’ UTR sequences from the +1 of the mRNA transcript to +100 after the protein coding sequence’s start codon were inputted into the biophysical model. For individual 5’ UTRs, a list of standby site modules was automatically generated. For each standby site module, the surface area $A_S$, distortion energy penalty $\Delta G_{\text{distortion}}$, unfolding energy penalty $\Delta G_{\text{unfolding}}$ and sliding energy penalty $\Delta G_{\text{sliding}}$ were calculated. These calculations are combined to determine the $\Delta G_{\text{standby}}$ for each standby site module. Because of the absence of cooperative binding, as shown, the standby site module with the lowest binding free energy penalty is selected as the one controlling the ribosome’s translation initiation rate.

According to the model calculations, we found that the lengths of the natural 5’ UTRs do not correlate with their ability to bind the ribosomal platform (Figure 2-6A). Short 5’ UTRs with low accessibility can inhibit translation equally well, compared to long 5’ UTRs where accessibility is limited by RNA structures. The model calculations indicate that the ribosomal platform can bind to natural 5’ UTRs with a wide range of energetic penalties (up to 17.4 kcal/mol), repressing translation by up to 2500-fold (Figure 2-6B). Long 5’ UTRs with low $\Delta G_{\text{standby}}$ can be subjected to translation repression when trans-acting small RNAs bind and reduce their single-stranded accessibility (46, 51, 57–59), or when cis-acting riboswitches reduce
accessibility when bound to a chemical ligand (60–64). In one example, the standby site in the tisB 5’ UTR is occluded when bound by the IstR-1 small RNA, and its translation is repressed (46). In contrast, long 5’ UTRs with high $\Delta G_{\text{standby}}$ can be targets for translation activation when structural remodeling by small RNAs or riboswitches increases their surface accessibility. RprA small RNA has been shown to bind and up-regulate the translation of rpoS mRNA with the help of Hfq-protein (65).

Figure 2-6. Predicted $\Delta G_{\text{standby}}$ across Escherichia coli MG1655 genome. (A) The ribosomal platform binding free energy penalties, $\Delta G_{\text{standby}}$, for 3430 transcribed 5’ UTRs from the Escherichia coli MG1655 genome (EcoCyc release 17.1) are compared to their lengths. The inset shows the $\Delta G_{\text{standby}}$ of short 5’ UTRs. (B) The number of genomic 5’ UTRs with different values of $\Delta G_{\text{standby}}$ is shown. (C) 5’ UTR
isoforms within transcripts across the genome affect their standby site module characteristics, as quantified by the maximum calculated change in $\Delta G_{\text{standby}}$.

The model predicts that both types of translation regulation occur in natural 5′ UTRs. In particular, 39% of genomic 5′ UTRs have standby site modules whose distal or proximal binding sites are longer than 15 nt, which are potent targets for translational repression; in contrast, 9% of 5′ UTRs contain standby site modules with low accessibilities ($A_S < 10$), making them targets for translation activation. Moreover, 79 diverse 5′ UTRs with an average length of 175 nt have positive sliding energy penalties ($\Delta G_{\text{sliding}}$), indicating that distant standby site modules are controlling translation, and making them potential targets for long-range regulation.

Further, there are 769 operons where variability in their transcriptional start sites, due to adjacent or overlapping promoters, result in 5′ UTRs with different lengths and mRNA structures. For example, the mmuPM operon has five different transcriptional start sites, leading to 5′ UTR isoforms with lengths from 9 to 205 nt. The biophysical model predicts a 1088-fold change in the translation initiation rate of mmuP, as a result of changes in the structure and surface area of the different standby site modules. Overall, across all 5′ UTR isoforms within the genome, model calculations show large changes in translation initiation rates through alterations in the standby site modules’ characteristics (Figure 2-6C). These differences imply that switching promoter usage, due to changing transcription factor or sigma factor levels, can modulate an mRNA’s translation rate, and potentially create standby site modules whose accessibility can be additionally regulated by translation factors. However, further investigation is necessary to examine the 5′ UTRs of isoforms and the regulation of their translation rates.
2.3.8. Biophysical modeling to calculate translation rates from split ribosome binding sites

Sacerdot et al. (66) studied the translation of the 5′ UTR of *E. coli* thrS mRNA that uses a split ribosome binding site. The biophysical model classifies the 5′ UTR as containing three standby site modules with large RNA structures and varying distal and proximal binding site lengths (Figure 2-7). The authors concluded that a 24 nt long single stranded region (domain 3) is essential for ribosome binding and that the ribosome can accommodate a long hairpin (domain 2) without unfolding it, which was later supported by crystallographic data (35). The biophysical model’s calculations are strongly consistent with their conclusions. In particular, a mutation that deleted the entire domain 3 and its upstream region (ILOΔ4) repressed the translation rate by over 50-fold, where it significantly reduced the standby site accessibility (from $A_S = 24$ and $\Delta G_{\text{distortion}} = 0$ to $A_S = 0$ and $\Delta G_{\text{distortion}} = 17.4$ kcal/mol). To offset this large energetic burden, model calculations indicate that the ribosomal platform partially unfolded 6 bp from the domain 2 hairpin, which minimized its binding free energy penalty to a $\Delta G_{\text{standby}}$ of 5.32 kcal/mol ($A_S = 20.5$, $\Delta G_{\text{distortion}} = 0.02$ and $\Delta G_{\text{unfolding}} = 5.3$ kcal/mol). Destabilization of the domain 2 hairpin (ILOΔ5) restored the mRNA’s translation to its wild-type rate by increasing the surface area of the standby site to 20.5 nt, resulting in an almost zero $\Delta G_{\text{standby}}$. The biophysical model calculations indicate that other mutations to the thrS 5′ UTR (L6, M1, N1, BS4-9, BS4-9/CS29, L19, L7, CS30Δ2, ILOΔ1 and ILOΔ2; see Figure 2-7) did not alter its standby site module surface area or RNA structure; the maximum observed change in translation was 2.5-fold for these mutations.
Figure 2-7. $\Delta G_{\text{standby}}$ for split ribosome binding sites. The sequence and structure of E. coli thrS 5' UTR are shown. Similar to Ref (66), nucleotide numbering is negative for the upstream 5' UTR with respect to the start codon. The list of mutations is: BS4-9: G(-32)>A; BS4-9/CS29: G(-32)>A and G(-46)>A; L19: A(-13)>C; L7: U(-49)>G; CS30Δ2: deletion of domain 2 from A(-14) to U(-48); ILOΔ1: transcription begins at G(-159); ILOΔ2: transcription begins at G(-68); ILOΔ4: transcription begins at G(-49), U(-49)>G, and A(-13)>C; ILOΔ5: transcription begins at G(-49), U(-49)>G, and G(-46)>A. Bolded nucleotides are the positions of new transcriptional start sites. SD: Shine-Dalgarno sequence.

2.4. Discussion and conclusion

We designed and characterized synthetic mRNA sequences and applied biophysical modeling to decipher the rules for the ribosome’s interactions at structured 5' UTRs. The 30S ribosome searches for a single standby site module that supports the highest translation rate, regardless of its distance from the start codon. The ribosome’s binding free energy penalty to individual standby site modules is governed by a competing trade-off between the accessibility of standby site modules (Figure 2-2) and the unfolding of RNA structures to increase accessibility.
(Figure 2-3). We demonstrate that thermodynamic minimization predicts the extent of the ribosome’s selective and partial unfolding of RNA structures. Our data supports a sliding mechanism whereby downstream RNA structures are not unfolded, but attenuate translation rate in a height-dependent manner (Figure 2-4).

Importantly, initially designed sequences have well-defined features that perturb the ribosome’s individual interactions with mRNA, allowing us to precisely measure their effect on the ribosome’s binding free energy penalty. The forward design of synthetic sequences, according to model predictions, quantitatively tests our knowledge of these interactions and the model’s ability to predict their strengths. Overall, the biophysical model employs first-principles calculations with four empirically measured constants to accurately predict the translation initiation rates of 136 mRNAs with diversely structured, long 5’ UTRs (Figure 2-5).

Based on our findings, there is also the potential for long-range regulation of translation. Long, structured 5’ UTRs can feature several upstream standby site modules extending over 100 nt away from an open reading frame. Any of these standby site modules are potential binding sites for the ribosome, according to their binding free energy penalties, but only a single standby site module is ultimately bound. Both cis-acting and trans-acting factors can modulate individual standby site module surface accessibilities to control standby site module selection and the resulting translation rate (35, 44, 46). Importantly, the regulation of standby site module selection will produce discrete step changes in translation rate that can further regulate downstream processes according to non-Boolean, multi-state logic.

In this study, we introduce a new metric to quantify the surface area of a standby site module, which was validated using diverse, but canonical, RNA secondary structures. The metric relates the geometric characteristics of the standby site module (P, D, H) to its available single-stranded surface area, implicitly using coefficients of one to indicate interchangeability. Additional measurements will be needed to precisely measure these coefficients, and relate them
to the helical shape of A-form RNA. In addition, tertiary structures, such as G-quadruplexes and pseudoknots, will likely have a differently defined metric of surface area. Further investigation will be necessary to determine the effect of tertiary RNA structures on standby site accessibility, ribosome binding and translation initiation.

As suggested by the biophysical model, the ribosomal platform’s ability to bind standby site modules with low surface area relies on its structural flexibility. Previously published cryo-EM data shows that the platform domain in the 30S subunit, in both the free and 50S-bound states, can occupy variable conformational states (67). The two halves of the platform, which are stabilized by ribosomal proteins at its surface, switch conformations during translation (7). Restraining this flexibility will reduce the ribosome’s translation initiation rate; for example, when the antibiotic Edeine binds 16S ribosomal RNA helices H23 and H24 (68). Consequently, the ribosomal platform’s structural flexibility appears to be an essential aspect of translating mRNAs with structured 5’ UTRs.

Similarly, the rigidity of single-stranded RNA likely plays an important role in modulating the ribosome’s distortion penalty, due to the entropic differences of binding flexible or rigid macromolecules (69). Similar to DNA, polyA RNA sequences are less flexible than mixed or polyU RNA sequences (70, 71). Differences in RNA rigidity have been shown to alter the ribosome’s binding affinity to spacer regions separating the SD and start codon sequences (72); binding free energy differences are ~2.5 kcal/mol between polyA and polyAU spacer sequences, and ~5 kcal/mol between polyA and polyU spacer sequences. Incorporating a sequence-dependent model for RNA rigidity into the distortion penalty calculations could potentially increase their accuracy.

Analysis of the model’s error can lead to further insight into the ribosome’s interactions. The model’s predictions have a normally distributed error for 90% of sequences (Figure A-6). Thirteen outliers have errors between 2 and 4 kcal/mol (Supplementary_Data_2). Some outliers
are due to the effects of increased mRNA degradation of 5′ UTRs with very long proximal or distal binding sites (P, D > 20 nt) (Figure A-2). In others, extremely short proximal or distal binding sites can result in non-canonical base pairings that alter the ribosome’s binding affinity. For example, in the absence of a proximal binding site (P = 0 nt), co-axial stacking will take place between the nucleotides in the distal binding site and the mRNA–rRNA duplex that anchors the ribosome to the mRNA. Similarly, co-axial stacking will play a larger role in predicting the effects of small RNAs that bind to sites nearby RNA structures.

As the list of interactions that control gene expression continues to grow, we will stretch the limits of human pattern recognition and the use of observational correlations. Biophysical modeling offers a comprehensive approach to account for all known interactions with the gene expression machinery, identify gaps, design experiments, precisely measure strengths and make testable predictions. The accuracy of these models will go hand-in-hand with our ability to engineer genetic systems without trial-and-error. We have incorporated the ribosome’s interactions with structured standby site modules into an improved biophysical model, called the RBS Calculator v2.0. A software implementation and user-friendly web interface is available at http://salislab.net/software.
Chapter 3

Temperature Regulates Translation Initiation Rate by Altering Both the Stability of RNA Base-Pairings and the Flexibility of RNA Hairpins

This work was done in collaboration with Tian Tian. All the measurements and model predictions for this work are presented in Supplementary_Data_3 and can be found online at: http://salislab.net/files/Espah_Borujeni_ThesisSupplementaryData.zip.

3.1. Introduction

Microorganisms have specific preference on their optimum growing temperature. Multiple cellular mechanisms that can sense and repose to the change in temperature allow the microorganism to maintain their efficient growth in rapidly changing environmental conditions. There are two distinct temperature dependent mechanisms inside cells: protein-based regulations and RNA-based regulations. In the protein-based regulations, the cold- and heat-shock chaperon proteins and their associated sigma-factors control the gene expressions through both transcriptional and post-translational interactions (73–78). The RNA-based regulations, however, control the gene expression during the translation level by altering either translation rate or RNA stability (79–85). For example, most natural RNA thermosensors, such as ROSE (Repression Of heat-Shock gene Expression (86)), control the translation rate by perturbing structures that sequester mRNA’s Shine-Dalgarno region, resulting in an acceleration of translation rate (79). Several efforts have been done previously to develop synthetic RNA thermosensors that can control gene expression by regulating RNase E activity (87), ribozyme activity (88), and
ribosome-mRNA interactions (89–92) at diverse temperatures. However, the successful generation of such temperature-based RNA devices requires an in-depth knowledge of how temperature alters the stability and flexibility of mRNA folding, as well as mRNA interactions with ribosome.

Here we specifically investigate how temperature controls the interactions that take place between mRNA and ribosome during the translation initiation step. As explained in Chapter 2, ribosome accommodates the mRNA on its platform surface, and remains at standby site until the mRNA’s inhibitory structures are unfolded, allowing the ribosome’s anti-SD region to base pair with mRNA at SD-like sequences. The total energy change from the unfolding and hybridization events mainly controls the rate of translation initiation. Importantly, the Gibbs free energy change associated with both RNA unfolding and mRNA-rRNA hybridization is largely affected by the temperature, providing a predictive basis for tunable control of translation initiation rate at different temperatures. However, what is not clear is that how temperature controls the binding of ribosome to mRNA’s standby site. This is particularly important because it has been shown that the standby site structures upstream of SD sequence undergo no to partial unfolding step, and therefore their thermodynamics has a minimum effect on ribosome binding strength (12). As we showed in Chapter 2, it is the standby site’s geometry (and not its energy) that mainly controls the ribosome binding efficiency. Therefore, we sought to investigate how temperature can affect the geometry-to-function relationship between ribosome and structured standby sites.

In this chapter, we characterized the in vivo translation initiation rate of 46 mRNAs under a wide range of temperatures (25, 30, and 37 °C). All these synthetic mRNAs encode an identical mRFP1 fluorescent protein reporter, but have dissimilar 5’ UTRs with diverse sequences, structures, and geometries. By analyzing the measured intracellular fluorescence levels at different temperatures, we discovered that lowering the temperature reduces the flexibility of RNA hairpins that bend over the neighboring mRNA regions and sequester the single-stranded
surface area for binding to the ribosome platform. This reduction in RNA flexibility increases the mRNA’s available single-stranded surface area for binding to ribosomal platform surface, which in turn enhances the translation initiation rate. More interestingly, we found that, at lower temperatures, ribosome still follows the same thermodynamic minimization principle that was observed at 37 °C. According to this principle, ribosome platform partially unfolds the standby site structures only to minimize its standby energetic penalty. Finally, the reduced RNA hairpin flexibility assists the ribosome for an efficient sliding process, in which ribosome expends less energy penalty to push aside hairpins that impede its sliding motion. Overall, our results provide new insights on how ribosome initiates translation at different temperatures, which can be used in rational design of RNA thermosensors that have diverse translation initiation rates under wide range of temperatures.

3.2. Materials and methods

3.2.1. Plasmids and strains

46 expression plasmids that were designed and characterized at 37 °C (reported in Chapter 2 and Ref. (12)) were re-characterized in this work at temperatures 25 °C and 30 °C. These plasmids are pFTV1 vectors with ColE1 origin of replication, Chloramphenicol (Cm) resistance gene, and mRFP1 fluorescent protein reporter. All measurements are performed using *E.coli* DH10B cells.
3.2.2. Growth and characterization

The growth and characterization procedures were similar to Chapter 2, except that cultures were grown at different temperatures (25 °C and 30 °C), and four colonies were characterized per each construct (n = 4). During each characterization, the temperature of both the overnight culture and the “serial dilution” growth were maintained the same to avoid any undesired effect of temperature change on cellular conditions. The single-cell protein expression level measurements were carried out using BD Fortessa flow cytometer. The average fluorescence levels were corrected for the background *E. coli* DH10B fluorescence. All the fluorescence values are listed in Supplementary_Data_3.

3.2.3. Biophysical model calculations at different temperatures

As explained in Chapter 2, our biophysical model calculates the total binding free energy change between the ribosome and mRNA using Equation 3.1, and relates it to the mRNA’s translation initiation rate, \( r \), using Equation 3.2.

\[
\Delta G_{\text{total}} = \Delta G_{\text{mRNA-rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} + \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}}
\]

\[
r \propto \exp \left( - \beta \Delta G_{\text{total}} \right)
\]

From these five energetic interactions, we assume that only \( \Delta G_{\text{mRNA}} \), \( \Delta G_{\text{mRNA-rRNA}} \), and \( \Delta G_{\text{standby}} \) can change at different temperatures, because they represent the temperature-dependent RNA base-pairing stability and RNA hairpin flexibility. We should mention that, since \( \Delta G_{\text{spacing}} \) is only a function of ribosome’s conformational flexibility and is not sequence-specific, we assume that it does not change by temperature.

\( \Delta G_{\text{mRNA}} \) and \( \Delta G_{\text{mRNA-rRNA}} \) represent the Gibbs free energy content of RNA folding and RNA-RNA hybridization, and are calculated using empirically determined RNA energy models.
(52, 53) and the minimization algorithms available in the Vienna RNA suite version 1.8.5 (54). Therefore, at each temperature, we calculate the RNA energy contents using these models. In general, the stability of RNA base-pairing increases as temperature decreases. For example, the folding Gibbs free energy of an RNA hairpin sequence (5'-GGCCCGAUACGCGGGCC-3') at 25, 30, and 37 °C varies between -13.9, -12.9, and -11.6 kcal/mol, respectively. Similarly, the Gibbs free energy change for the hybridization of canonical SD sequence (5'-UAAGGAGGU-3') and the ribosome’s anti-SD sequence (5'-ACCUCUUA-3') varies between -17.6, -16.5, and -15.0 kcal/mol at 25, 30, and 37 °C, respectively. Consequently, a more energetically favorable mRNA-rRNA hybridization at lower temperature can compensate a higher energy penalty for the mRNA unfolding at similar temperature.

We also assume that the apparent Boltzmann constant, $\beta$, which has been previously measured at 37 °C (0.45 ± 0.05 mol/kcal (21)), remains constant at different temperatures.

Finally, the temperature regulates $\Delta G_{\text{standby}}$ via two mechanisms: first, altering the RNA base-pairing’s stability, and second, altering the RNA hairpin’s flexibility. Below, we explain how the model calculates $\Delta G_{\text{standby}}$ at different temperatures.

Similar to previously explained in Chapter 2, the biophysical model automatically decomposes a 5' UTR into one or more standby site modules, where a standby site module contains a single mRNA hairpin (with height H), surrounded by two single-stranded regions: an upstream distal binding site (with length D) and a downstream proximal binding site (with length P). We also consider a 5' UTR without upstream hairpin as the no-hairpin control mRNA.

We assume that mRNA hairpin is flexible enough that can bend over the surrounding mRNA region, and sequester the single-stranded regions. However, the flexibility depends on the temperature and it drops as temperature decreases. The detailed analysis of our fluorescence data revealed that the flexibility of RNA hairpin drops from 100% at 37 °C to ~82% and ~70% at 30 °C and 25 °C, respectively. This is equivalent of a 36 and 45 degree reduction in the bending
rotation of RNA hairpin over adjacent mRNA region. Therefore, we incorporate the effect of temperature in our model by calculating the available single-stranded surface area, $A_S$, according to:

\[ A_S = 15 + P + D - (\alpha \cdot H) \quad (3.3) \]

Similar to Chapter 2, the value of $(\alpha \cdot H)$ in Equation 3.3 is not allowed to exceed 15 nt.

In Equation 3.3, $\alpha$ is related to temperature according to the following formula, where $T$ is temperature in Celsius, and $\alpha$ reaches maximum of 1.0 at $T = 37^\circ C$.

\[ \alpha = 0.0242 (T) + 0.1046 \quad (3.4) \]

Having the $A_S$ calculated using the effective hairpin height values (Equation 3.3), the ribosomal platform’s distortion energy penalty $\Delta G_{\text{distortion}}$ is calculated according to:

\[ \Delta G_{\text{distortion}} = C_1 (A_S)^2 + C_2 (A_S) + C_3 \quad (3.5) \]

where $C_1$, $C_2$ and $C_3$ are the fitted parameter values and equal to 0.038 kcal/mol/nt$^2$, −1.629 kcal/mol/nt and 17.359 kcal/mol, respectively. Importantly, the effective hairpin height is applied when calculating the $\Delta G_{\text{unfolding}}$ and $\Delta G_{\text{sliding}}$ as well.

### 3.2.4. Measurements of translation rate and binding free energy penalties

Changing the temperature can alter protein expression level through many processes including translation, transcription, and mRNA degradation. To remove the non-specific effects of temperature, and relate the changes in protein expression levels to the translation initiation rates, we measured the steady-state fluorescence level of a no-hairpin control mRNA under the identical conditions as the mRNAs with structured 5’ UTRs. This no-hairpin control mRNA has maximally accessible 5’ UTR with similar promoter, coding section, terminator, and plasmid’s origin of replication as other structured mRNAs.
Next, we calculate the change in the apparent binding free energy penalty of a mRNA with structured 5’ UTR by comparing its fluorescence with the fluorescence of the no-hairpin control mRNA, according to:

$$
\Delta G_{\text{standby}} = -\ln \left( \frac{X}{X_{\text{ref}}} \right) - \Delta \Delta G_{\text{other}}
$$

(3.6)

where $X$ is the fluorescence from a sample structured mRNA, $X_{\text{ref}}$ is the fluorescence of the no-hairpin control mRNA, and $\beta$ is 0.45 mol/kcal. $\Delta \Delta G_{\text{other}}$ is the difference in free energies, between the sample and control mRNAs, that depends on other ribosomal interactions and not on the standby site. The value for $\Delta \Delta G_{\text{other}}$ varied between 0.0 and 2.5 kcal/mol at 25 0C, between 0.0 and 1.7 kcal/mol at 30 0C, and between 0.0 and 0.6 kcal/mol at 37 0C, across all 46 mRNAs characterized in this work. All the Gibbs free energy calculations for all mRNA sequences are listed in the Supplementary_Data_3.

3.3. Results

3.3.1. Structured standby sites are less inhibitory at lower temperatures

We selected 28 mRNA sequences from our previously characterized mRNA sequences (see Chapter 2), that have varying distal binding sites (D) (between 2 and 12 nt), varying proximal binding sites (P) (between 1 and 15 nt), and different hairpin height (H) (between 9 and 18 nt). All these mRNA sequences have GC-rich hairpin stems and therefore remain folded when binding to the ribosome platform ($\Delta G_{\text{standby}} = \Delta G_{\text{distortion}}$). We characterized the steady-state fluorescence level of these mRNAs, along with a no-hairpin control mRNA, at both 25 and 30 0C (Figure 3-1). Overall, in comparison with the fluorescence level at 37 0C, we observed a 20% reduction, and a 10% increase in the fluorescence level of the no-hairpin control mRNA at 25 0C.
and 30 °C, respectively. These changes are the indication of non-specific effects of temperature on the protein expression levels, and therefore were removed from our data analysis using Equation 3.6.

**Figure 3-1. The geometry of standby site controls translation initiation rate.** 28 synthetic mRNA sequences that encode an mRFP1 fluorescent protein, and have diverse geometry (varying hairpin height H, distal binding site D, and proximal binding site P) were characterized at 25 and 30 °C. The measurements at
were taken from Chapter 2. The symbols are similar to Figure 2-2. Briefly: mRNAs with varying proximal length (P) all have 6 nt at distal site and hairpin heights of 9 nt (green stars), 12 nt (black asterisks), 15 nt (red squares), or 18 nt (blue circles); mRNAs with varying distal length (D) have 1 nt at proximal site and hairpin heights of 18 nt; and mRNAs with varying hairpin height (H) have 6 nt at distal site and 5 nt at proximal site. The hairpin with H = 11 nt (blue circle) has 6 bp in the stem and 10 nt on the loop. The hairpins with H = 14 to 18 nt (red squares) have 12 bp in the stem and varying loop length. The dashed lines are the fluorescence levels of no-hairpin control mRNA at each temperature. The data and standard deviations are the average of 2 measurements in 2 separate days (n = 4). All the data are listed in Supplementary_Data_3.

Our measurements showed that the geometry of structured standby sites control translation initiation rate similarly in different temperatures. A longer distal and proximal binding sites increase the ribosome binding rate, while a longer hairpin height is not favorable for ribosome binding. However, we also observed that the structured 5’ UTRs have less inhibitory effects on translation initiation rates at lower temperatures. For example, the fluorescence level of an mRNA with hairpin height of 11 nt reached the fluorescence level of no-hairpin control mRNA at 25 °C (only 1.3-fold apart), while the same hairpin caused 2.2-fold reduction in mRNA’s fluorescence level at 37 °C. Similarly, the maximum fold change in translation initiation rates with respect to the no-hairpin control mRNA dropped from 200-fold to 150-fold and 125-fold, respectively, when lowering the temperature from 37 °C to 30 °C and 25 °C. All the mRNA sequences, fluorescence levels, and Gibbs free energy values are listed in Supplementary_Data_3.
3.3.2. Lowering temperature reduces the RNA hairpin flexibility and increases the single-stranded surface area

We calculated the apparent standby free energy penalty ($\Delta G_{\text{standby}}$) for each structured mRNA according to Equation 3.6. Importantly, the effect of temperature on the energy content of RNA-RNA interactions was already considered in $\Delta\Delta G_{\text{other}}$ (Equation 3.6). The maximum $\Delta\Delta G_{\text{other}}$ values across these 28 mRNA sequences increased from 0.6 to 1.7 and 2.5 kcal/mol when decreasing the temperature from 37 to 30 and 25 °C. In general, $\Delta\Delta G_{\text{other}}$ depends on energetic interactions that involve RNA base-pairing and not standby site interactions. We then calculated the available single-stranded surface area ($A_S$) of all 28 mRNA sequences, using formula: $A_S = 15 + D + P - H$ (Figure 3-2A), and plotted them against the apparent $\Delta G_{\text{standby}}$ values (Figure 3-2B). The apparent $\Delta G_{\text{standby}}$ values follow a similar trend with respect to the $A_S$ at three different temperatures. However, the apparent $\Delta G_{\text{standby}}$ at lower temperatures increasingly deviated from the model predictions at 37 °C (quadratic dashed lines). The average error between apparent $\Delta G_{\text{standby}}$ and model prediction (at $T = 37$ °C) was 1.35, 0.77, and 0.5 kcal/mol at $T = 25$, 30, and 37 °C, respectively. In addition, the observed maximum surface accessibility (where the apparent $\Delta G_{\text{standby}} = 0$ kcal/mol) dropped dramatically when decreasing the temperature ($A_S = 13$, 17, and 21 nt at $T = 25$, 30, and 37 °C, respectively). This large difference indicated that a similar standby site region with constant D, P, and H is more accessible at lower temperature than higher temperature. We hypothesize that when temperature is low, the rotational flexibility of RNA hairpin decreases, and therefore the RNA hairpins cannot fully bend over the surrounding mRNA regions and only a portion of the hairpin covers the single stranded regions (Figure 3-2C). The less flexible RNA hairpin becomes, the more single-stranded surface area is available for binding to the ribosome platform, resulting in lesser
apparent $\Delta G_{\text{standby}}$. We demonstrated our hypothesis by introducing a temperature-dependent coefficient for the hairpin height $H$ to calculate the new $A_s$, according to Equation 3.3. We found that the coefficients 0.71 and 0.83 can minimize the error between the apparent $\Delta G_{\text{standby}}$ and the model predicted $\Delta G_{\text{distortion}}$ using Equation 3.5 at temperatures 25 and 30 °C, respectively (Figure 3-2D).

\[ A_s = 15 + P + D - H \]
Figure 3-2. RNA hairpin flexibility at different temperatures controls the distortion energy penalty.

(A) The available single-stranded surface area $A_S$ is calculated assuming that the hairpin can fully bend over the adjacent mRNA region. (B) The model predictions using maximum RNA flexibility assumption become increasingly inaccurate at lower temperatures. The average error between apparent $\Delta G_{\text{standby}}$ and the model predictions are 1.35, 0.77, and 0.5 kcal/mol at $T = 25, 30, \text{and} 37 \, ^\circ \text{C}$, respectively. (C) A cartoon showing how lowering temperature reduces the RNA hairpin flexibility, and prevent the hairpin from fully bending over the surrounding mRNA regions. (D) When calculating the surface area $A_S$ using the effective hairpin height ($H_R$) based on formula: $H_R = H \cos(\theta)$, the apparent $\Delta G_{\text{standby}}$ values at three temperatures are well-predicted by the model. The average error between apparent $\Delta G_{\text{standby}}$ and the new model predictions are 0.71, 0.61, and 0.5 kcal/mol at $T = 25, 30, \text{and} 37 \, ^\circ \text{C}$, respectively. The symbols are similar to part B.

The average error between apparent $\Delta G_{\text{standby}}$ and newly calculated model prediction at each temperature is now 0.71, 0.61, and 0.5 kcal/mol at $T = 25, 30, \text{and} 37 \, ^\circ \text{C}$, respectively. Interestingly, these coefficients are equivalent of $\pi/4$ and $\pi/5$ degree reduction in rotational movement of RNA hairpin, respectively ($\cos(\pi/4) = 0.71$ and $\cos(\pi/5) = 0.81$) (Figure 3-2C). These values showed a linear relationship between the temperature and the hairpin coefficient that can be used to extrapolate the coefficient for a wide range of temperatures (Equation 3.4).

3.3.3. Ribosome follows the thermodynamics minimization principle at lower temperatures

In Chapter 2, we showed that, due to the lack of external energy resources during translation initiation process, ribosome platform only partially unfolds the standby site structures to minimize the energy penalty for binding to standby site. There is a coupled trade-off between site accessibility and partial RNA unfolding. Here, we sought to investigate that whether the same
thermodynamic minimization principle is applied at lower temperatures. In particular, since at lower temperatures, the unfolding of RNA hairpins becomes energetically more costly while the site accessibility becomes more available, we expect to see a lower extent of RNA hairpin unfolding at lower temperatures. We illustrate the effect of temperature on the extent of RNA hairpin unfolding in Figure 3-3A. At lower temperature, the reduction in RNA hairpin flexibility leads to higher site accessibility, which in turn makes the unfolding of RNA hairpins more unfavorable. Therefore, for the hairpin sequence shown in Figure 3-3A, no unfolding is predicted at 25 °C, while the ribosome platform unfolds 2 and 3 base pairs of RNA hairpins at 30 and 37 °C, respectively. We demonstrated this concept in more details by selecting 14 structured mRNAs that have diverse sequences and structures, and were previously characterized at 37 °C. These mRNAs are predicted to have up to 3 unfolded base-pairs at 37 °C (Chapter 2). In this work, we re-characterized these mRNAs at 25 and 30 °C. We then calculated the apparent $\Delta G_{\text{standby}}$ values using Equation 3.6, and plotted them against the model predictions under three different scenarios: 1- ribosome platform does not unfold any RNA hairpins, 2- ribosome platform only partially unfolds hairpins to minimize its energy penalty, and 3- ribosome platform fully unfolds the RNA hairpins (Figure 3-3C). We found that at each temperature, the thermodynamic minimization principle accurately predicts the standby free energy penalty ($R^2 = 0.96, 0.95, 0.94, p = 1.0 \times 10^{-9}, 5.1 \times 10^{-9}, 1.4 \times 10^{-8}$, and average $\Delta \Delta G_{\text{standby}} = 0.48, 0.51, 0.78$ kcal/mol when $T = 25, 30, \text{ and } 37 \, ^{\circ}C$, respectively), while the fully unfolding scenario poorly predicts the standby free energy penalty ($R^2 = 0.22, 0.23, 0.33, p = 0.09, 0.08, 0.03$, and average $\Delta \Delta G_{\text{standby}} = 16.28, 14.14, 11.33$ kcal/mol when $T = 25, 30, \text{ and } 37 \, ^{\circ}C$, respectively). Interestingly, the no-unfolding scenario becomes increasingly more accurate when temperate decreases, indicating that reduction in the RNA flexibility at lower temperature naturally minimizes the energy penalty such that ribosome does not need to further unfold any RNA
hairpin ($R^2 = 0.95, 0.95, 0.92$, $p = 3.8 \times 10^{-9}, 6.0 \times 10^{-9}, 6.2 \times 10^{-8}$, and average $\Delta G_{\text{standby}} = 0.68, 1.37, 2.89$ kcal/mol when $T = 25, 30, \text{and } 37 \, ^\circ C$, respectively). In addition, we found that, for most of these mRNAs, the number of unfolded base-pairs decreased at lower temperature, in accordance with the thermodynamic minimization principle (Figure 3-3B).
Figure 3-3. Ribosome follows a thermodynamic minimization principle at all temperatures. (A) The sequence of a structured standby site is shown to illustrate the thermodynamic minimization principle at different temperatures. (B) Ribosome platform only partially unfolds the standby site structures to minimize its binding free energy penalty. At lower temperatures, the reduced RNA flexibility results in a higher surface area, and a lower distortion energy, which ultimately eliminates the need for partial RNA unfolding. The ribosome's distortion penalty $\Delta G_{\text{distortion}}$ (dashed red line), hairpin unfolding penalty $\Delta G_{\text{unfolding}}$ (dotted-dashed blue line), and their sum $\Delta G_{\text{standby}}$ (green solid line) are shown as the hairpin's closing base pairs are unfolded. The total binding free energy penalties have local minima (shown with asterisk symbol) when 0, 2 and 3 base pairs are unfolded at 25, 30, and 37 °C, respectively. (C) Scatter plots comparing the apparent and model predicted $\Delta G_{\text{standby}}$ from 14 synthetic 5' UTRs to test three mechanisms for the ribosome's unfolding of RNA structures. The thermodynamic minimization principle accurately predicts the standby energy penalty at all three temperatures ($R^2 = 0.96, 0.95, 0.94$, $p = 1.0 \times 10^{-9}, 5.1 \times 10^{-9}, 1.4 \times 10^{-8}$, and average $\Delta \Delta G_{\text{standby}} = 0.48, 0.51, 0.78$ kcal/mol when $T = 25, 30, \text{and } 37 \, ^\circ\text{C}$, respectively). Data points are the average of 2 measurements in 2 separate days ($n = 4$). In part C, the diagonal dashed lines are the $x = y$ line.

3.3.4. Reduced hairpin flexibility makes the sliding process energetically more favorable

In Chapter 2, we showed that the ribosome platform can bind the mRNA’s distant standby sites, and then slide on the mRNA surface toward the initiation region. However, RNA hairpins that impede the sliding motion should be pushed aside, which penalizes the ribosome with an energy penalty proportional to their hairpin height. In this work, we were interested in testing whether lower temperature is beneficial for the sliding motion since the effective hairpin height is reduced at lower temperatures. For this purpose, we selected three structured mRNAs that have more than one standby site modules with the most accessible standby site located far upstream from initiation site (shown with roman number I) (Figure 3-4A). These mRNAs were
previously characterized at 37 °C demonstrating the effect of sliding free energy penalty. In this work, we re-characterized these mRNAs at 25 and 30 °C, and then calculated their apparent ΔG_{standby} values (Figure 3-4B). We found that the apparent ΔG_{standby} (white bars) for each mRNA decreases at lower temperatures, indicating that ribosome more efficiently slides on these mRNAs at lower temperature. The model predicts the sliding energy penalty (blue bars) using the formula: ΔG_{sliding} = 0.2 (Σ H_i), where Σ H_i is the summation of the heights of all hairpins that impede the sliding motion (hairpins shown with asterisk symbol). We found that a lower effective hairpin height at lower temperature allowed the model to more accurately predict the standby energy penalties (Figure 3-4B). The reduced hairpin flexibility is beneficial for the sliding motion, especially when there are multiple hairpins to be bypassed by the ribosome.
Figure 3-4. Reduced RNA flexibility at lower temperatures facilitates the ribosome’s sliding motion.

(A) Synthetic 5’ UTRs with one upstream (labeled with roman number I) and multiple internal standby site modules (labeled with asterisk symbols) are shown. (B) Each mRNA was characterized at three temperatures. The apparent $\Delta G_{\text{standby}}$ at each temperature (white bars) are compared with the predicted $\Delta G_{\text{standby}}$ (blue bars) at the same temperature. Model predictions include the sliding energy penalty to bypass the internal hairpins. The reduced hairpin heights to calculate the sliding energy penalty at lower temperatures are calculated according to $\alpha.H$, where $\alpha$ is obtained from Equation 3.3.
3.3.5. Biophysical model accurately predicts the standby energy at different temperatures

We incorporated the effect of temperature in the biophysical model calculations by introducing a temperature-dependent coefficient for the hairpin heights (see Materials and methods section). Overall, the biophysical model accurately predicted the ΔG_{standby} for all 46 mRNA sequences, characterized at different temperature (R^2 = 0.92, 0.95, 0.93, p = 6.6x10^{-26}, 2.0x10^{-29}, 1.3x10^{-27} when T = 25, 30, and 37 °C, respectively) (Figure 3-5). In particular, the average error in predicted ΔG_{standby} varied between 0.67, 0.62, and 0.67 kcal/mol at T = 25, 30, and 37 °C, respectively. These results demonstrate that the model is accurate in predicting the ribosome binding free energy at different temperature by less than 0.7 kcal/mol.

![Figure 3-5. Biophysical model predicts the standby energy penalty at different temperatures](image)

3.4. Discussion and conclusion

In summary, we showed that temperature can regulate translation initiation rate via two separate mechanisms: first by altering the Gibbs free energy of RNA base-pairing and second by
altering the RNA hairpin’s effective height when binding to the ribosome platform. We showed that at lower temperature, the flexibility of RNA hairpins may drop, increasing the available single-stranded surface area for binding to the ribosome platform. For example, at 25 °C, the reduced RNA flexibility allows an RNA hairpin to bend over the surrounding sequences by only 45 degree, resulting in a higher site accessibility and larger single-stranded surface area. In contrast, a maximum flexibility at 37 °C allows the same RNA hairpin to fully bend over its adjacent nucleotide regions. This reduced RNA flexibility decreases the net standby energy penalty by: 1- decreasing the distortion energy penalty, 2- lowering the extent of RNA unfolding, and 3- enhancing the sliding motion by lowering the sliding energy penalty.

The reduced RNA rigidity at lower temperature can be further demonstrated in a molecular basis by designing an RNA molecule with a specified distal length (D), hairpin height (H), and proximal length (P), such that a fluorophore and a quencher is attached on the hairpin loop and the single stranded proximal site, respectively. The RNA produces a minimum fluorescence level when the hairpin fully bends over the single stranded region, bringing the fluorophore to the closest distance from the quencher. Therefore, by measuring the fluorescence level at a wide range of temperatures, we can test whether or not the RNA rigidity is affected by temperature. Overall, our findings illustrate the diverse role of temperature on cellular functions, and provide a new design approach for developing temperature-controlled RNA devices.
Chapter 4

Reparametrizing the Biophysical Model of Translation Initiation for *Bacillus subtilis*

This work was done in collaboration with Manish Kushwaha, Walker Huso, and Kevin M. Thyne. All the measurements and model predictions are presented in Supplementary_Data_4 and can be found online at: http://salislab.net/files/Espah_Borujeni_ThesisSupplementaryData.zip.

4.1. Introduction

*Bacillus subtilis* is one of the most well-studied gram-positive bacteria (93), with many interesting biological properties, enabling its diverse lab-scale and industrial applications, including the ability for sporulation (94, 95), the natural competence (96–98), the large-scale production of antibiotic and anti-microbial active compounds (99), and the efficient protein secretion mechanism (100–104). These excellent capabilities make *B. subtilis* an efficient enzyme-producing alternative for its gram-negative counterpart, *E. coli*. As the rate limiting step in protein production, translation initiation in *B. subtilis* has been largely studied to determine how it differs from translation initiation in *E. coli* (105–111). Although, several mechanistic similarities and differences between translation initiation in *B. subtilis* and *E. coli* have been reported, no quantitative study has been done to predict the effect of mRNA’s sequence and structure on protein synthesis rate in *B. subtilis*.

Here, we characterized 47 mRNA sequences, encoding an mRFP1 fluorescent protein reporter, and compared their measured protein expression levels in both *B. subtilis* and *E. coli*. 
Using our biophysical model calculations, previously developed for translation initiation in *E. coli*, we identified the detailed differences between *B. subtilis* and *E. coli* translation initiation mechanisms. We discovered that although mRNA-ribosome binding in both bacterial species follows the same biophysical rules, the execution of these rules largely depends on the conformational flexibility of ribosomes. The 30S ribosome in *E. coli* shows more flexibility than *B. subtilis*’ 30S ribosome, when accommodating mRNAs with non-optimal spacing distance between SD and start codon. In contrast, highly flexible 30S ribosome platform in *B. subtilis* allows the ribosome to support efficient translation initiation for mRNAs with highly structured 5’ UTRs, an ability that is lacking for 30S ribosome platform in *E. coli*. Importantly, this function can explain why most of highly structured natural riboswitches have been discovered in *B. subtilis* (112–124). Finally, by comparing measurements with model predictions, we parameterized our biophysical model for translation initiation in *B. subtilis*. Development of such predictive model would allow researchers to rationally fine-tune their protein production rate in *B. subtilis*, and can lead to design of synthetic riboswitches functioning inside *B. subtilis*.

### 4.2. Materials and methods

#### 4.2.1. Growth, transformation and selection

*Bacillus subtilis* 168 (BGSCID# 1A1) used in this study were cultured at 30 °C in LB broth Miller (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) obtained from BD (#244610). Solid media for plates was prepared by adding 1.5% w/v agar (BD# 214010) and 1% w/v potato starch (Sigma-Aldrich # S5651). Naturally competent *B. subtilis* cells were transformed using a standard protocol (125), and screened for correct genomic integration into the *amyE* locus by the iodide starch plate assay, where absence of a zone of starch hydrolysis around the colony
indicates successful construct integration. B. subtilis cells with genomically integrated construct were selected at 5 µg/mL chloramphenicol (Alfa Aesar# B20841).

4.2.2. Plasmid constructions

Plasmids were constructed using standard cloning techniques involving PCR, primer annealing, restriction digestion, gel-extraction, ligation, as well as chew-back anneal-repair (CBAR) (126). Cloning was performed in E. coli DH10B, and constructs were confirmed by sequencing before transformation into Bacillus. Bacillus genomic integration vector pDG1661 (127) (BGSCID# ECE112: ampR, cmR, spcR, ColE1 replication origin) was used to clone all constructs after removing the spoVG-lacZ region. Each construct contained an expression cassette and a Chloramphenicol resistance (CamR) gene between the amyE homologous recombination arms. The cassette consisted of a constitutive pVeg promoter (BBa_K143012) (5’-TTTAAATTTATTTGACAAAAATGGGCTCGTGTTACAATAATGT-3’) with a previously identified transcription start site (128), a 5’UTR containing the ribosome binding site (RBS), an mRFP1 gene, and a T1 terminator (BBa_B0015). Construct variants tested here varied only in their 5’UTR sequence.

For constructing the 5’ UTRs with varying spacing lengths, unique SpeI and EcoRI restriction sites were introduced to pDG1661 vector, immediately after the promoter and the ATG start codon, respectively. Oligonucleotides containing 5’-UTR with varying spacing lengths, flanked by SpeI and EcoRI sites, were orders from IDT (Integrated DNA Technologies), and cloning was performed as mentioned above. To construct the 5’ UTRs with diverse standby site structures, the 5’ UTR sequences from previously constructed pFTV1 vectors (Chapter 2) were PCR amplified and directly transferred to pDG1661 vector. However, these 5’ UTR inserts had two main issues. First, they contained XbaI and SacI restriction sites, which were not suitable for
cloning into pDG1661 vector, since multiple XbaI and SacI sites were present in pDG1661 backbone. Second, they contained the *E.coli* canonical SD sequence (5’-TAAGGAGGT-3’), which must be replaced with the colloquially known SD sequence of *B. subtilis* (5’-AAAGGAGGT-3’). Therefore, the inserts were PCR amplified using two primers, in which EcoRI restriction site and the *B. subtilis* SD sequence were introduced in the reverse primer, while the forward primer still contained the XbaI restriction site. We kept the XbaI site in the forward primer because elimination of XbaI site could disrupt the RNA structures for most of the 5’ UTRs. Next, the PCR products were digested with XbaI and EcoRI and ligated to the pDG1661 backbone that was digested with SpeI and EcoRI. Importantly, since XbaI and SpeI restriction enzymes produce compatible overhangs, they can be ligated together, although at the expense of eliminating both XbaI and SpeI sites in the final ligation product. All the constructs in this study were sequence-verified, and all listed in the Supplementary_Data_4.

4.2.3. Fluorescence measurement

The growth and steady-state fluorescence measurements were performed as described in Chapter 2, except that *B. subtilis* cells containing the cloned 5’ UTRs, together with a control transformant (Vector-Only), were inoculated in triplicate into 500 μL LB broth (with 5 μg/mL chloramphenicol) each, in a 96 deep-well plate, and incubated overnight at 30 °C with 200 rpm orbital shaking. The “serial dilution” experiment was carried out at 30 °C as well. The mean auto-fluorescence of the control mRNA measured using Fortessa flow cytometer (BD Biosciences) was subtracted from each sample mRNA. The experimental measurements are reported in Supplementary_Data_4.
4.2.4. Quantification of spacing and standby free energy penalties

We calculate the change in the apparent standby energy penalty of a mRNA with structured 5’ UTR by comparing its fluorescence with the fluorescence of the no-hairpin control mRNA measured under identical conditions, according to Equation 4.1. This no-hairpin control mRNA had maximally accessible 5′ UTR (14 nt single-stranded standby site).

\[
\Delta G_{\text{standby}} = -\frac{1}{\beta} \ln \left( \frac{X}{X_{\text{ref}}} \right) - (\Delta \Delta G_{\text{other}})
\]  

(4.1)

In above equation, \(X\) is the fluorescence from a sample structured mRNA, \(X_{\text{ref}}\) is the fluorescence of the no-hairpin control mRNA, and \(\beta\) is 0.45 mol/kcal. \(\Delta \Delta G_{\text{other}}\) is the difference in free energies, between the sample and control mRNAs, that depends on other ribosomal interactions and not on the standby site. The value for \(\Delta \Delta G_{\text{other}}\) varied between 0.0 and 1.5 kcal/mol across all 39 mRNAs characterized in this work.

The spacing energy penalty was calculated by comparing the fluorescence level of a sample 5’ UTR with non-optimal spacing, with the fluorescence level of the 5’ UTR with optimal spacing (with maximum fluorescence level), according to Equation 4.2.

\[
\Delta G_{\text{spacing}} = -\frac{1}{\beta} \ln \left( \frac{X_S}{X_{S_{\text{opt}}}} \right) - (\Delta \Delta G_{\text{other}})
\]  

(4.2)

In above equation, \(X_S\) is the fluorescence from a sample mRNA with spacing length \(S\), \(X_{S_{\text{opt}}}\) is the fluorescence from a mRNA with optimal spacing length \(S_{\text{opt}}\) (maximum fluorescence level), and \(\beta\) is 0.45 mol/kcal. \(\Delta \Delta G_{\text{other}}\) is the difference in free energies, between the sample mRNA and mRNA with maximum fluorescence, that depends on other ribosomal interactions such as \(\Delta G_{\text{mRNA}}\). The value for \(\Delta \Delta G_{\text{other}}\) varied between -0.3 and 2.0 kcal/mol across
all 7 mRNAs with varying spacing length. All the Gibbs free energy calculations for all mRNA sequences are listed in the Supplementary_Data_4.

4.3. Results

4.3.1. 30S ribosomes in both B. subtilis and E. coli experience similar energetic penalty at non-optimal spacing lengths

It has been shown that the 30S ribosomes in E.coli undergoes an energetically unfavorable conformational change to accommodate mRNA sequences with non-optimal spacing distance between SD and start codon (21). In this work, we sought to investigate and precisely characterize the spacing energy penalty for mRNA-ribosome binding in B. subtilis. We designed and characterized 7 synthetic 5’ UTRs that are identical in their sequences, except that their spacing length varied between 0 and 15 nt. These 5’ UTRs all encode the same mRFP1 fluorescent protein reporter, and have the colloquially known SD sequence of B. subtilis (5’-AAAGGAGGT-3’) (106, 129), with a 12 nt single stranded standby site upstream of SD. The measured steady-state fluorescence levels showed a maximum at spacing of 6nt (Figure 4-1A), indicating that the optimal spacing for B. subtilis is 6 nt, which is in agreement with the previous reports (130). However, this is one nucleotide longer than the optimal spacing in E. coli (4, 21). The reduction in fluorescence level at non-optimal spacing length in B. subtilis follows the same trend as in E. coli. A shorter spacing (S < 6 nt) causes a steep reduction in fluorescence, whereas a longer spacing (S > 6nt) has a modest decrease in fluorescence (Figure 4-1A). The maximum reductions in fluorescence level is 410-fold at 0 nt spacing.
We calculated the apparent spacing energy penalty using Equation 4.2, and plotted the results against their corresponding spacing lengths (Figure 4-1B). We fitted the apparent spacing energies to a set of compression and stretching energy penalty formulas that were used previously for spacing energy calculation in *E.coli* (21). The fitted values are obtained by minimizing the error between the apparent and calculated energies. When S < 6 nt, the ribosome undergoes compression to accommodate the mRNA, with a sigmoidal energy formula:

\[
\Delta G_{\text{spacing}} = \frac{C_1}{(1 + \exp\left(C_2 \left(S - S_{\text{opt}} + C_3\right)\right)^3}
\]

(4.3)

where \(C_1\), \(C_2\), and \(C_3\) are 11.40 kcal/mol, 2.67 nt\(^{-1}\), and 0.79 nt, respectively (Figure 4-1B). These values are relatively similar to the parameters identified for *E.coli* (12.2 kcal/mol, 2.5 nt\(^{-1}\), and 2 nt). Conversely, when S ≥ 6 nt, the ribosome stretches to place the widely spaced SD and ATG in their right positions, with the following quadratic energy function:

\[
\Delta G_{\text{spacing}} = C_1 \left(S - S_{\text{opt}}\right)^2 + C_2 \left(S - S_{\text{opt}}\right)
\]

(4.4)

where \(C_1\) and \(C_2\) are 0.12 kcal/mol/nt\(^2\) and 0.22 kcal/mol/nt, respectively (Figure 4-1B). The corresponding parameters for *E. coli* are 0.05 kcal/mol/nt\(^2\) and 0.24 kcal/mol/nt, respectively. Overall, although the compression and stretching mechanisms exist in both *B. subtilis* and *E. coli*, the differences in the spacing energy penalties (Figure 4-1B) suggest that the 30S ribosome in *B. subtilis* is conformationally less flexible than 30S ribosome in *E. coli*. For example, *E. coli* 30S ribosome can efficiently accommodate a 3 nt non-optimal spacing (\(\Delta G_{\text{spacing}} = 1.50\) kcal/mol), whereas this short spacing is very unfavorable for *B. subtilis* (\(\Delta G_{\text{spacing}} = 11.30\) kcal/mol). Moreover, in contrast to the *E. coli*, in *B. subtilis* a 15 nt spacing still causes the same energetic penalty as a 0 or 3 nt spacing, indicating that similar to compression, stretching too is unfavorable for 30S ribosomes in *B. subtilis*. 
Figure 4-1. Characterization of the spacing energy penalty in *B. subtilis*. (A) The measured fluorescence levels are plotted against the spacing length. Data points are the average of 3 measurements in one day (n=3). (B) The apparent $\Delta G_{\text{spacing}}$ values quantified in *B. subtilis* (yellow circles) are fitted to a sigmoidal ($S < 6$) or quadratic ($S > 6$) energy function (blue line). For comparison, the spacing energy function in *E. coli*, as determined in Ref. (21), is also depicted (red dashed line). The optimal aligned spacing in *E. coli* is 5 nt. The difference between the spacing energy functions in *B. subtilis* and *E. coli* suggests that 30S ribosomes in *B. subtilis* are more rigid than 30S ribosomes in *E. coli*.

4.3.2. Investigating the anti-SD sequence of 30S ribosome in *B. subtilis*

The last nine nucleotides at the 3’-end of the processed16S ribosomal RNA can base pair with the SD-like regions on mRNA during translation initiation. Although this “anti-SD” sequence has a universally conserved function, its exact sequence is organism-dependent. For example, the anti-SD sequence of *E. coli* and *B. subtilis* are known to be 5’-ACCUCCUUA-3’ and 5’-ACCUCCUUU-3’, respectively. However, the large similarity between these two sequences raised a question about whether or not the reported one-nucleotide difference has a distinguishable effect. Here, we designed two different 5’ UTR sequences that are identical in
their standby site region (14 nt unstructured sequence) and spacing region (6 nt poly-As), but different in their SD region: a canonical SD for *E. coli* (5’-UAAGGAGGU-3’) versus a colloquially known SD for *B. subtilis* (5’-AAAGGAGGU-3’) (106, 129). The measured steady-state fluorescence levels showed 5.2-fold higher translation rate for the mRNA with *E. coli*’s SD sequence, compared to the mRNA with *B. subtilis*’ SD (Figure 4-2). This large difference suggests that *E. coli*’s SD sequence is more favorable to bind to the *B. subtilis* 30S ribosome. Using our biophysical model calculations, we compared the predicted translation initiation rates of these two mRNAs under two scenarios: the anti-SD sequence is either *E. coli*’s anti-SD or *B. subtilis*’ anti-SD (Figure 4-2). The anti-SD sequence controls the rate of translation initiation through its hybridization energy with the mRNA’s SD sequence ($\Delta G_{mRNA-rRNA}$). A perfect sequence-complementarity between SD and anti-SD releases the highest free energy, and makes the mRNA-ribosome binding stronger. Consequently, our model calculations showed that only when the *E. coli*’s anti-SD was used, the predictions were in agreement with the measurements (Figure 4-2). These results provide evidence that the actual anti-SD sequence of *B. subtilis* is the same as *E. coli*’s anti-SD. Although further experimental measurements is needed to support this finding, incorporation of *E. coli*’s SD in the synthetic 5’ UTR sequences seems to be a robust design rule to achieve a high level of protein expression in *B. subtilis*. 
Figure 4-2. 30S ribosomes in both *E. coli* and *B. subtilis* have similar anti-SD sequence. The measured steady-state fluorescence levels (grey bars) are shown for 5’ UTR sequences with two different SD regions. The data points are the average of three measurements in one day (n=3). The predicted translation initiation rates for both 5’ UTR sequences are calculated according to formula: 
\[
    r = 70 \times \exp \left( -0.45 \Delta G_{\text{total}} \right),
\]
where \( \Delta G_{\text{total}} \) was calculated using either *E. coli*’s anti-SD (light blue) or *B. subtilis*’ anti-SD (dark blue).

4.3.3. Characterizing mRNAs with structured 5’ UTR in *B. subtilis*

Most natural riboswitches have been discovered in the genomic 5’ UTRs in *B. subtilis* (112–124). In general, these natural riboswitches have complex secondary and tertiary structures and control the gene expression through both translation initiation and transcription termination. Therefore, understanding how *B. subtilis* deals with the structured 5’ UTRs during translation initiation is of great interest. In Chapter 2, we showed that structured 5’ UTRs regulate the
mRNA-ribosome binding in *E. coli* through several mechanisms, including platform distortion, partial RNA unfolding, and platform’s sliding on the mRNA surface. Here, we sought to investigate whether or not these mechanisms are still functional in *B. subtilis*. Therefore, we selected 38 mRNA sequences from our previously characterized structured 5’ UTRs in *E. coli* (Chapter 2), and transferred their entire 5’ UTR section to an expression vector (pDG1661) suitable for characterization in *B. subtilis* (see Materials and methods). The selected 5’ UTRs were divided into three categories: 5’ UTRs that regulate distortion energy penalty, 5’ UTRs that undergo partial RNA unfolding, and 5’ UTRs with multiple hairpins to demonstrate the sliding mechanism. Although these transferred 5’ UTRs were structurally identical to their original mRNAs tested in *E. coli*, there existed a few sequence-dissimilarities. First, a 6 bp poly-A sequence was used as an optimal spacing in the new 5’ UTRs. Second, the colloquially known *B. subtilis’* SD was used (5’-AAAGGAGGU-3’). Third, due to the cloning purposes, the 5’-end nucleotide of most mRNA sequences, and the N-terminal of mRFP1 coding section were different. Importantly, by characterizing a no-hairpin control 5’-UTR with similar SD, spacing, and mRFP1 coding section as the structured 5’ UTRs, we eliminated these confounding factors that can affect the expression of mRFP1 in *B. subtilis*.

We measured the steady-state fluorescence level of these structured 5’ UTRs in *B. subtilis* and compared them with their corresponding measurements in *E. coli* (Figure 4-3). Overall, the mRFP1 expression levels in both bacterial hosts showed a relatively similar trend ($R^2 = 0.36, p = 0.0001$), where most of the outliers belong to the category of mRNAs with diverse distortion energy penalty (red circles in Figure 4-3). This similarity in protein expression indicates that both bacterial species use similar mechanisms to control translation initiation. However, since the maximum change in fluorescence level across all the structured 5’ UTRs dropped from 160-fold in *E. coli* to only 20-fold in *B. subtilis*, the translation machinery at *B. subtilis* seems to be much more efficient in dealing with diverse structured 5’ UTRs.
Figure 4-3. **Comparison of mRFP1 expression levels in two bacterial hosts.** The measured steady-state fluorescence levels of 38 structured 5’ UTRs together with the no-hairpin control mRNA are compared between two bacterial hosts. The data points in *B. subtilis* are the average of three measurements in one day (n =3). The data points in *E. coli* are taken from Chapter 2. (red circles) 5’ UTRs with diverse distortion energy penalty, (blue circles) 5’ UTRs that undergo partial RNA unfolding, (yellow circles) 5’ UTRs with multiple hairpins showing sliding mechanism, and (green circle) no-hairpin control mRNA. R² = 0.36, p = 0.0001. All measurements are listed in Supplementary_Data_4.

4.3.4. **Ribosome platform in *B. subtilis* is more flexible to accommodate mRNAs with inaccessible standby sites**

From all 38 characterized mRNAs, 23 mRNAs were used to test the distortion energy penalty in *B. subtilis*. These mRNAs have highly stable hairpins (with GC-rich stems) that remain folded when binding to the ribosome platform. In addition, they have diverse geometry with wide range of distal lengths (D = 2 to 12 nt), proximal lengths (P =1 to 12 nt), and hairpin heights (H = 9 to 18 nt). We plotted their fluorescence levels with respect to their proximal length, distal
length, or hairpin heights (Figure 4-4ABC). Overall, we found that lengthening the proximal length (P) and distal lengths (D) increases the fluorescence levels, regardless of the hairpin height in the standby site (Figure 4-4AB), which is in accordance with our observations in E. coli. However, for most of the structured standby sites, we observed a much higher fluorescence level than the no-hairpin control mRNA, which is rather surprising and is in contrast with our observations in E. coli. The lower expression level of no-hairpin control mRNA may be caused by reduced mRNA stability due to its long single-stranded standby site with 12 nt poly-A sequence. More importantly, if the expression level of this no-hairpin control mRNA is in fact lower than the structured mRNAs, it can introduce a new, but unknown, regulatory mechanism for translation initiation in B. subtilis (see Discussion and conclusion). Regardless of the expression level of no-hairpin control mRNA, our data indicate that the presence of structures in standby site regulates translation initiation rate by up to 16-fold.

Another striking observation in B. subtilis measurements was that the effect of hairpin height on translation initiation rate did not follow the observed trend in E. coli (Figure 4-4C). In particular, increasing the hairpin height by lengthening the loop length reduced the fluorescence levels in E. coli (Figure 2-2E), whereas in B. subtilis, the hairpin loops longer than 6 nt showed no inhibition. In E. coli, when calculating the available surface area, the loop length was considered as a part of the hairpin height, which does not apply to the B. subtilis measurements. Therefore, as an alternative interaction, we hypothesized that, in B. subtilis, when the hairpin bends over the surrounding nucleotide regions, if the hairpin loop is long (loop length > 6 nt), a portion of hairpin loop behaves like a single stranded region that can interact with the positively charged surface of ribosome platform. We name this region of the hairpin “single-stranded loop”, where its length is calculated according to formula: (ssLoop = loop length – 6 nt). The lack of this interaction in E. coli suggests that, in contrast to B. subtilis, the 30S ribosome platform surface in E.coli is not wide enough to interact with single-stranded loop regions when the hairpin bends.
Figure 4-4. Available single-stranded surface area controls the distortion energy penalty in *B. subtilis*. (A, B, C) Translation rates are measured as proximal lengths, distal lengths, and hairpin heights are increased. The dashed lines are the expression level of no-hairpin control mRNA. (C) All distal lengths are 6 nt, and hairpin heights are 9 nt (green stars), 12 nt (black asterisks), 15 nt (red squares), or 18 nt (blue circles). (D) All proximal lengths are 1 nt, and hairpin heights are 18 nt. (E) All distal and proximal lengths are 6 and 5 nt, respectively. (blue circle) A hairpin with 6 bp stem and 10 nt loop, yielding a hairpin height of 11 nt. (red squares) Hairpins with an identical 12 bp stem, but with different loop lengths (from 4 to 12 nt), resulting in hairpin height of 14 to 18 nt. (D) Cartoon shows how the available single-stranded surface area is calculated in *B. subtilis*. The grey colored regions are available to interact with the ribosome platform surface. (E) 22 characterized 5' UTRs from parts ABC are combined to quantitatively relate the
available surface area to the standby free energy penalty. Dashed line is a best-fit quadratic equation ($R^2 = 0.83$, $p = 3.3 \times 10^{-9}$, average error = 0.59 kcal/mol). 5' UTRs are colored if they have varying distal length (red circles), varying proximal length (blue circles), and varying hairpin height (yellow circles). Data points are the average of three measurements in one day ($n = 3$). All measurements and model calculations are listed in Supplementary_Data_4.

It has been shown that, in *E. coli*, the available single-stranded surface area ($A_S$), calculated according to the formula: $A_S = 15 + D + P - H$, controls the distortion energy penalty. Here we show that the single-stranded surface area controls the distortion energy penalty in *B. subtilis* as well, but with a small modification in calculating $A_S$, as shown in Figure 4-4D. The hairpin height is calculated as the summation of base-pairs in the stem and half of the nucleotides on the loop. If the loop length is greater than 6 nt, only 6 nt of the loop is used to calculate the hairpin height and the rest of the loop becomes the single-stranded loop (ssLoop). Also since *B. subtilis* grows at 30°C, the RNA hairpins have ~20% less rigidity, preventing them from complete bending over the neighboring regions. This 20% reduction in rigidity is equivalent of 36 degree ($\theta = \pi/5$) reduction in hairpin’s rotational movement (see Chapter 3).

We calculated the apparent standby energy penalty of these structured 5’ UTRs according to Equation 4.1, and plotted them against their modified surface area $A_S$ (Figure 4-4E). Overall, a quadratic relationship, similar to *E. coli*, was observed between the calculated surface area and the apparent distortion energy penalty. A three parameter quadratic model was fitted to the data, according to Equation 4.5.

$$\Delta G_{\text{distortion}} = C_1(A_S)^2 + C_2(A_S) + C_3$$ (4.5)

, where $C_1$, $C_2$ and $C_3$ are the fitted parameter values and equal to 0.018 kcal/mol/nt$^2$, $-0.724$ kcal/mol/nt and 4.064 kcal/mol, respectively ($R^2 = 0.83$, $p = 3.3 \times 10^{-9}$, average error = 0.59
kcal/mol). It is noteworthy that because the expression level of no-hairpin control mRNA was lower than most of the structured mRNAs, the apparent distortion energies yielded negative values. Negative value means that the binding is favorable for the ribosome platform. Importantly, although the sign of apparent distortion energy depends on the reference mRNA, the maximum change in distortion energy between most accessible ($A_S = 20$ nt) and least accessible ($A_S = 0$ nt) standby sites remains constant (7.24 kcal/mol). Interestingly, this maximum change in distortion energy is 10.1 kcal/mol lower than *E. coli*’s maximum distortion energy penalty (17.36 kcal/mol), indicating that 30S ribosome platform in *B. subtilis* can more efficiently accommodate inaccessible standby sites than in *E. coli*. This difference can arise from the specific structural conformation of 30S ribosome in each bacterial host. A lower energy penalty for distorting the *B. subtilis* platform surface suggests that either its surface is wider and can make more interactions with mRNA (as proposed for single-stranded loop region), or it has more flexibility to distort than ribosome platform in *E. coli*.

### 4.3.5. Modest distortion energy penalty in *B. subtilis* can eliminate the need for partial RNA unfolding

We have shown in Chapter 2 that *E.coli*’s ribosome platform partially unfolds RNA hairpins to increase the standby site’s available surface area and decreases the distortion energy penalty. The main reason behind the partial RNA unfolding is the lack of energy resource during translation initiation, and thus the need for thermodynamic minimization. Here, we sought to investigate if *B. subtilis* also follows this thermodynamic minimization. However, since the distortion energy penalty in *B. subtilis* is much weaker than *E. coli*, our calculations suggest that
B. subtilis’ ribosome platform may not need to unfold RNA hairpins, and the ribosome can still be under thermodynamic minimization without unfolding any RNA hairpin.

From the 38 characterized mRNAs, 11 mRNAs have a single hairpin in their standby sites, while their stem was destabilized using AU-rich sequences. They have a wide range of distal sites (1 to 10 nt), proximal sites (1 to 4 nt), and hairpin heights (9 to 18 nt), resulting in a diverse range of available surface areas (A_S between 4 and 16 nt). Using the model calculations, we predicted the standby energy penalties for each mRNA under three different scenarios: no RNA unfolding, partial RNA unfolding (thermodynamic minimization), and complete RNA unfolding. Similar to the results in E. coli, the complete RNA unfolding scenario poorly predicted the standby energy penalty (R^2 = 0.27, p = 0.10, average ΔΔG_{standby} = 20.65 kcal/mol) (Figure 4-5), showing that ribosome in B. subtilis also tends to save its energy during translation initiation and do not expend it on RNA unfolding if not necessary. More importantly and as expected before, both “no RNA unfolding” and “partial RNA unfolding” scenarios resulted in similar predictions, indicating that, due to the weak distortion energy penalties, ribosome does not need to even partially unfold RNA hairpins (R^2 = 0.79, p = 2.4x10^-4, average ΔΔG_{standby} = 1.46 kcal/mol for both scenarios) (Figure 4-5). However, the intriguing observation about these mRNAs was that the apparent standby energy penalty for all 11 mRNAs was ~1.5 kcal/mol less than the model predictions. We hypothesize that B. subtilis’ ribosomes use some extra, but unknown, energy resources or catalytic activities (equivalent of 1.5 kcal/mol) to partially unfold these weakly stable RNA hairpins, which requires further investigations. It is also noteworthy that the distortion energy calculations for these 11 mRNAs were highly accurate (R^2 = 0.79, p = 2.4x10^-4) (Figure 4-5), ruling out the possibility of mis-predicting the distortion energy during thermodynamic minimization calculations.
Figure 4.5. Investigating the partial RNA unfolding in *B. subtilis*. Scatter plots comparing the apparent and model predicted $\Delta G_{\text{standby}}$ from 11 synthetic 5’ UTRs. The model predicts that ribosome in *B. subtilis* does not need to even partially unfold RNA hairpins because its distortion energy penalty is weak enough to support efficient ribosome binding ($R^2 = 0.79$, $p = 2.4 \times 10^{-4}$, and average $\Delta \Delta G_{\text{standby}} = 1.46$ kcal/mol for “no RNA unfolding” and “partial RNA unfolding”). Similar to *E. coli*, the “complete RNA unfolding” scenario is quite unfavorable for *B. subtilis*’ ribosome ($R^2 = 0.27$, $p = 0.10$, and average $\Delta \Delta G_{\text{standby}} = 20.65$ kcal/mol). Data points are the average of three measurements in one days ($n = 3$). The diagonal dashed lines are the $x = y$ line.

4.3.6. Ribosome platform in *B. subtilis* slides on mRNA surface with less energy penalty

When the standby site module with most accessible surface area is located far upstream from initiation site, the ribosome platform binds to the standby site and slides on the mRNA surface toward the initiation site, during which it has to push aside the impeding hairpins at the expense of a sliding energy penalty. In Chapter 2, we showed that *E. coli*’s ribosome expends 0.2 kcal/mol/nt energy to push aside the hairpins during the sliding process. Here, we selected two of those mRNA sequences with multiple hairpins in their standby site (M₂ and M₄) (Figure 4-6). Their measured steady-state fluorescence levels demonstrated that these highly-structured 5’
UTRs can support efficient translation initiation in \textit{B. subtilis}. In both mRNAs, ribosome binds to the upstream standby site (shown in roman number I), but has to bypass different number of internal hairpins (shown with asterisks symbols). Our calculations show that without sliding energy, the model under-predicts the standby energy (-3.2 kcal/mol model prediction versus ~-2.2 kcal/mol apparent standby energy). However, when calculating the sliding energy using formula:

\[
\Delta G_{\text{sliding}} = 0.05 \left( \sum H_i \right),
\]

where \( \Sigma H_i \) is the summation of the heights of all internal hairpins (shown with asterisk symbol), the model accurately predicted the apparent standby energy penalties. As explained in Chapter 3, the hairpin heights used in calculating the sliding energy are the reduced hairpin heights at lower temperature (30 °C), according to \textbf{Equation 4.3}. Importantly, the new \textit{B. subtilis}' sliding coefficient (0.05 kcal/mol/nt) is 4-fold smaller than the sliding coefficient in \textit{E. coli} (0.20 kcal/mol/nt), indicating that ribosome platform in \textit{B. subtilis} is more flexible in sliding on mRNA, and can easily bypass RNA hairpins that impede its movements. This finding provides further evidence that the ribosome platform in \textit{B. subtilis} is more flexible in binding to structured 5' UTRs than ribosome platform in \textit{E. coli}.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4-6}
\caption{Ribosome platform in \textit{B. subtilis} efficiently slides on mRNA surface. (A) Synthetic 5' UTRs with one upstream (labeled with roman number I) and multiple internal standby site modules (labeled with asterisk symbols) are shown. (B) The apparent \( \Delta G_{\text{standby}} \) (yellow bars) are compared with the predicted \( \Delta G_{\text{standby}} \) with sliding energy (light blue bars) and without sliding energy (dark blue bars).}
\end{figure}
4.3.7. Biophysical model predicts the mRNA’s standby energy and translation initiation rate in *B. subtilis*

We incorporated the new fitted parameters for distortion energy, as well as the new sliding coefficient into our biophysical model to predict the standby energy penalty for any given mRNA sequence in *B. subtilis*. Overall, the biophysical model predicted the standby energy penalty of all 38 structured 5' UTRs characterized in this work by less than 1.0 kcal/mol error ($R^2 = 0.56, p = 5.8 \times 10^{-8},$ average $\Delta \Delta G_{\text{standby}} = 0.97$ kcal/mol) (Figure 4-7A). Finally, using our biophysical model calculations, we accurately predicted the translation initiation rate of all 47 mRNAs studied in this work ($R^2 = 0.93, p = 1.3 \times 10^{-27},$ average $\Delta \Delta G_{\text{total}} = 0.82$ kcal/mol) (Figure 4-7B).

**Figure 4-7. Biophysical model predictions in *B. subtilis*.** (A) Apparent versus predicted standby energies for all 38 structured 5’ UTRs and the no-hairpin control 5’ UTR. (red circles) 5’ UTRs with diverse distortion energy penalty, (blue circles) 5’ UTRs to test the partial RNA unfolding, (yellow circles) 5’ UTRs with multiple hairpins to test sliding mechanism, and (green circle) no-hairpin control mRNA ($R^2 =$...
0.56, p = 5.8x10⁻⁸, average ΔG_{standby} = 0.97 kcal/mol). (B) Measured fluorescence levels versus predicted translation initiation rates for all 47 mRNAs in this work. Translation initiation rates are calculated according to formula: \( r = K \exp\left(-0.45 \Delta G_{\text{total}}\right) \), where K is a proportionality constant, and is set to 70 and 20 for mRNAs with varying standby sites (light blue circles) and spacing regions (red circles), respectively (\( R^2 = 0.93, p = 1.3x10^{-27} \), average ΔΔG_{total} = 0.82 kcal/mol).

4.4. Discussion and conclusion

In summary, by characterizing 47 mRNA sequences with carefully designed standby sites or spacing regions, we have shown that the translation initiation in *Bacillus subtilis* is fundamentally similar to the translation initiation in *E. coli*. However, a few key structural differences between the 30S ribosomes in these bacterial species resulted in significant dissimilarities in their mRNA-ribosome interactions. First, we showed that similar to *E. coli*, 30S ribosome either compresses or stretches to accommodate mRNAs with non-optimal spacing, at the expense of a spacing energy penalty (*Figure 4-1*). However, our data analysis revealed that the 30S ribosome in *B. subtilis* is much more rigid compared to *E.coli’s* 30S ribosome, causing it to become very sensitive to short or long non-optimal spacing lengths. Second, we found that mRNAs with the canonical *E. coli’s* SD sequence have higher translation rate in *B. subtilis* host than mRNAs with colloquially known *B. subtilis’* SD sequence, suggesting that the 30S ribosome’s anti-SD sequences in both *E. coli* and *B. subtilis* are identical (*Figure 4-2*). Although our biophysical model calculations confirmed these conclusions, further investigation is required to find the actual ribosome’s anti-SD by sequencing the 3’-end of 16S ribosomal RNA from *B. subtilis* genomic DNA.

Next, we showed that ribosome platform in *B. subtilis* can support efficient translation rate even from highly structured 5’ UTRs (*Figure 4-3*). Interestingly, we found that mRNAs with
structured 5' UTRs have much higher expression level than the no-hairpin control mRNA. There can be two explanations for this observation. First, due to the long single-stranded region (14 nt) upstream of SD region on the reference mRNA, the mRNA stability can drop, lowering the observed fluorescence level. We have previously shown in Chapter 2, that single-stranded regions longer than 12 nt are subject to mRNA degradation (Figure A-2). In fact, one of the characterized 5' UTRs in this work (Figure 4-4A, 5’ UTR with P = 12 nt), showed almost 2-fold reduction in fluorescence level, while it was supposed to have the maximum fluorescence level, indicative of a potential mRNA degradation. Therefore, by measuring the mRNA level of no-hairpin control mRNA and couple of structured 5’ UTRs we can quantify how much of the reduction in fluorescence level was due to mRNA degradation. The second explanation is that RNA structures in the standby site can promote an enhanced translation initiation by binding to some unknown RNA-binding proteins that can bring the mRNA to the close vicinity of 30S anti-SD sequence, thus increasing the protein synthesis rate. Importantly, this enhancing effect of RNA structures on translation rate has been previously shown in mammalian cells, in which the hairpins located close to the 5’-cap of eukaryotic mRNAs increased the protein expression by up to 10-fold (131). However, this hypothesis requires further investigation.

Regardless of the expression level of no-hairpin control mRNA, the diverse standby site structures across 23 characterized mRNAs resulted in 20-fold increase in the expression level of mRFP1 fluorescent protein reporter. This 20-fold increase is equivalent of 7.24 kcal/mol reduction in the 5’UTR’s apparent distortion energy penalties (Figure 4-4). Although the change in distortion energy is a quadratic function of mRNA’s surface area, the modest distortion energy penalty in B. subtilis compared to E. coli suggests that higher flexibility and wider surface of ribosome platform in B. subtilis allows it to more efficiently, and with less energetic penalty accommodate structured standby sites on its surface. Our data showed that the platform surface in
\textit{B. subtilis} may even interact with the single-stranded nucleotides on the loop, if the loop length is longer than 6 nt.

The higher flexibility of ribosome platform surface in \textit{B. subtilis} is also advantageous when dealing with the RNA unfolding and ribosome sliding interactions. Because the ribosome follows the thermodynamic minimization principle, partial unfolding of RNA hairpins becomes unfavorable for \textit{B. subtilis} because the distortion energy is weak enough that accommodating RNA hairpins without partially unfolding them is still energetically favorable (Figure 4-5). Moreover, when the ribosome encounters a hairpin during its sliding motion, the higher flexibility of ribosome platform allows it to more efficiently bypass the hairpin with much lower energy penalty to push the hairpin aside (Figure 4-6). Overall, in this work we investigated the mechanisms of translation initiation rate in \textit{B. subtilis} in great details, and showed how both the similarities and the difference between 30S ribosomes in \textit{E. coli} and \textit{B. subtilis} can drastically alter mRNA’s expression levels in these bacterial species.
Chapter 5

Precise Quantification of Translation Inhibition by RNA Structures that Overlap with the Ribosome Footprint at N-terminal Coding Sections

This work was done in collaboration with Iman Farasat, Ashlee Smith, Walker Huso, and Natasha Lundgren. Measurements and model predictions are presented in Supplementary_Data_5 and can be found online at: http://salislab.net/files/Espah_Borujeni_ThesisSupplementaryData.zip.

5.1. Introduction

Translation initiation is a rate determining step in gene expression, whereby 30S ribosome binds to mRNA’s standby site, hybridizes to Shine-Dalgarno sequence (SD), and loads mRNA into its narrow “mRNA channel” (11, 22, 31) before starting the peptide elongation. mRNA sequences often fold into several secondary structures that attenuate the 30S assembly process depending on their positions on mRNA (Figure 5-1A). Some secondary structures fold within the mRNA’s standby site and regulate mRNA’s binding to 30S ribosome platform according to their shape and geometry (10, 12, 44). Other secondary structures sequester SD sequence and inhibit mRNA-ribosome hybridization via their energy and folding kinetics (8, 21). A third class of secondary structures fold exclusively within the N-terminal coding section, which may overlap with the ribosome footprint. The double helical diameter of overlapping structures is larger than the diameter of mRNA channel (~ 20 Å) (22, 27), forcing 30S ribosome to unfold them upon their entry. This is likely to be done by the combination of thermal fluctuations, cooperative action of S3, S4, and S5 ribosomal proteins, and coordinated rotational movement of
head and body subunits (20, 22, 27), and is thermodynamically supported by the energy released from GTP hydrolysis (132–134). However, due to the absence of GTP hydrolysis during initiation, the 30S’s helicase activity lacks an energy resource, causing the overlapping structures to impede the mRNA loading process and inhibit translation initiation. The extent of this inhibitory effect on in vivo translation rate is yet to be precisely characterized.

Figure 5-1. N-terminal secondary structures inhibit translation initiation by overlapping with 30S ribosome footprint. (A) A biophysical model quantifies the total free energy change between free mRNA and 30S initiation complex, and relates it to the rate of translation initiation. RNA secondary structures control the 30S binding according to their position on mRNA. (B) Sequence of expression system with an AC-rich insulator containing an unstructured N-terminal coding section with only fast codons is shown. (C)
Exclusive RNA hairpins were introduced to the N-terminal coding section at varying distance from start codon. Only RNA hairpins that overlap with 30S footprint are unfolded.

The mRNA’s coding section controls protein synthesis rate according to a coupled relationship between several processes including translation initiation, translation elongation, translational pausing, translational coupling, co-translational protein folding, and ribosomal frameshifting (5, 6, 13, 15, 19, 21, 22, 28–30, 135–141). It has been a long-standing question that how protein coding sequence controls overall translation rate. A recent combinatorial study concluded that, despite a common notion, N-terminal RNA structures have a more dominant effect on gene expression than codon rarity (14), which is consistent with the finding that evolution tends to reduce inhibitory RNA structures at the gene start to support efficient translation initiation (1, 142). Although elimination of N-terminal structures seems to be a straightforward approach for optimizing protein coding sequence, it is still unclear until what extent ribosome must unfold the downstream RNA structures and whether the size or energy of RNA structures would influence the unfolding process. It has been suggested that structures spanning from -4 to +37 positions are inhibitory (143), whereas other in vitro studies have shown that only those structures that overlap with ribosome footprint inhibit translation (2). However, the precise length of ribosome footprint is also unknown. Several in vitro studies reported that ribosome footprint can cover a range between 12 and 19 nucleotides of N-terminal coding section (2, 16, 20, 22, 27, 31).

In this chapter, we sought to precisely measure the ribosome’s in vivo footprint, and to determine exactly which mRNA hairpin need to be unfolded to initiate translation. We designed 27 expression systems, in which we introduced an exclusive RNA hairpin to the N-terminal coding section of mRFP1 fluorescent protein reporter at varying distances from start codon (positions +4 to +40). Importantly, we also decoupled the effect of secondary structure formation
from codon usage efficiency by selecting fast codons for the N-terminal sequences. We observed up to 6000-fold change in protein expression level upon moving RNA hairpin away from start codon, indicating that not all RNA hairpins are inhibitory for translation initiation, and only those that overlap with ribosome footprint regulate translation initiation. Using biophysical model calculations, we identified the ribosome footprint to cover the first 14 to 15 nucleotides of coding section, which is in agreement with other reported in vitro studies. Incorporation of precise ribosome footprint to the biophysical model of translation initiation significantly improved the model predictions for a large set of 496 previously characterized mRNA sequences. Overall, our computational design and experimental approach allow us to study the exclusive effects of N-terminal coding sequence on translation initiation, and provides the fundamentals for building an integrated model of translation rate from mRNA’s coding sequence.

5.2. Materials and methods

5.2.1. Plasmids and strains

The pFTV1 plasmid (ColE1, Cm\textsuperscript{R}) (21), containing mRFP1 fluorescent reporter, was used to construct different mRFP1 coding section variants. A σ\textsuperscript{70} constitutive promoter (BioBrick #J23100) controlled the transcription of all the variants. To construct each mRFP1 variants, first a unique Ndel site (5’-CAUAUG-3’) was inserted in between XbaI and SacI sites in the original pFTV1 plasmid. The AUG in the NdeI site served as the start codon for mRFP1.
5.2.2. Cloning strategy

Using standard cloning techniques, a synthetic 5’ UTR sequence containing a strong and stabilizing 6 base-pair hairpin (5’-CCCGCCAUAUACGGCGGG-3’) was inserted in between XbaI and NdeI sites. The 5’ UTR provided a highly accessible standby site that supported efficient binding to the 30S ribosomal platform ($\Delta G_{\text{standby}} = 0.06 \text{ kcal/mol}$ (12)). A near-full Shine-Dalgarno (SD) sequence (5’-CAAGGAGAC-3’, $\Delta G_{\text{SD-antiSD}} = -8.68 \text{ kcal/mol}$) and a 3 nucleotide spacing between SD and start codon (5’-CAU-3’, $\Delta G_{\text{spacing}} = 1.52 \text{ kcal/mol}$) was chosen to avoid undesired secondary structures in the 5’ UTRs, yielding a high 30S ribosome binding rate. Next, 27 systematically designed N-terminal variants for mRFP1 coding section were ligated into the vector between NdeI and SacI sites, followed by transformation of electrochemically competent *Escherichia coli* DH10B cells. All the inserts were created by either annealing of oligonucleotides with correct overhangs (Integrated DNA Technologies), or by PCR assembly of oligonucleotides followed by restriction digest. Correct clones were confirmed by DNA sequencing. The sequences of all the designed mRFP1 variants are presented in the Supplementary_Data_5.

5.2.3. Growth and characterization

*E. coli* cells containing the correct clones were characterized using a 12 hr culture, as described in Ref (12). Briefly, for each construct, two clones were used to inoculate overnight cultures in 700 ul LB/Cm media within a 96-well deep-well plate and grown at 37 °C and 200 rpm orbital shaking. Next, 10 ul of cultures were diluted into 200 ul of fresh LB/Cm media and grown in a 96-well microplate at 37 °C and with proper shaking. Three cycles of growth/dilution were repeated to maintain the cells in exponential growth phase and to achieve a steady state
mRFP1 protein expression level. During each cycle, the OD$_{600}$ and fluorescence levels were recorded every 10 min using spectrophotometer TECAN M1000, until OD$_{600}$ reaching 0.15 indicating that cells are at exponential growth phase and the protein expression levels reached steady state. This experiment was repeated in two separate days for all the constructs (n = 4).

Single-cell protein expression levels were measured by collecting 10 ul of cultures at the end of the second and third growth cycles (at steady state conditions), and transferring them to 200 ul of 1X PBS solution (12). A Fortessa flow cytometer (BD Biosciences) with 100,000 events threshold was used to measure the fluorescence levels per single cell. All the fluorescence distributions were unimodal. The corrected fluorescence levels were obtained by subtracting the background E. coli DH10B auto-fluorescence. All the reported fluorescence levels were the average of 2 measurements in two separate days (n=4), and are listed in Supplementary Data_5.

5.2.4. Biophysical model of translation initiation

The biophysical rules governing mRNA-ribosome interactions during translation initiation have been explained in detail in Chapter 2. Here we present a brief summary of the biophysical model.

A statistical thermodynamic-based model predicts the mRNA’s translation initiation rate ($r$) using Equation 5.1, in which the model calculates the total Gibbs free energy change when a free mRNA binds to a 30S ribosome ($\Delta G_{\text{total}}$), as shown in Equation 5.2 (Figure 5-1A). $\beta$ is the apparent Boltzmann constant (21).

$$r \propto \exp(-\beta\Delta G_{\text{total}})$$

$$\Delta G_{\text{total}} = \Delta G_{\text{mRNA-rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} + \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}}$$
This model has been validated using 496 synthetic mRNAs in different bacterial species (12, 21, 28, 144). The current biophysical model uses a two-state binding assumption (initial and final states), in which each state reaches to its thermodynamic equilibrium. $\Delta G_{\text{mRNA}}$ is the folding free energy of mRNA at initial state. $\Delta G_{\text{mRNA-rRNA}}$ is the free energy contribution of RNA folding and RNA/RNA hybridization in the final state, according to Equation 5.3.

$$
\Delta G_{\text{mRNA-rRNA}} = \Delta G_{\text{pre-SD}} + \Delta G_{\text{SD-antiSD}} + \Delta G_{\text{post-footprint}}
$$

(5.3)

$\Delta G_{\text{SD-antiSD}}$ is the hybridization energy of the last nine nucleotides at 3'-end of 16S ribosomal RNA to mRNA’s SD region. When mRNA is hybridized to ribosome, the mRNA region within the ribosomal footprint must be unfolded (Figure 5-1A), while other non-overlapping mRNA regions are allowed to refold. $\Delta G_{\text{pre-SD}}$ is the refolding free energy of non-clashing mRNA structures upstream from SD region. $\Delta G_{\text{post-footprint}}$ is the refolding free energy of mRNA structures downstream from ribosomal footprint. This energy strongly depends on the length of 30S ribosome footprint, which was determined in this study to cover the first 15 nucleotides of coding section. The folding and hybridization free energies are calculated using the semi-empirical RNA free energy models (52, 53), provided by the Vienna RNA suite (version 1.8.5) (54).

Three additional interactions control the assembly of 30S initiation complex (30SIC) and consequently the rate of translation initiation. $\Delta G_{\text{start}}$ is the hybridization energy of tRNA$^{\text{fMet}}$ anticodon loop to the start codon at the P-site (most negative for AUG and GUG). However, a non-optimal nucleotide distance between start codon and SD sequence penalizes the 30S ribosome binding with a free energy penalty $\Delta G_{\text{spacing}}$ (positive value), which was characterized previously (21). The optimal aligned spacing for 30S ribosome is five nucleotides (4). Finally, binding of 30S ribosome platform to a less accessible standby site, upstream from SD, will cause a free
energy penalty $\Delta G_{\text{standby}}$ (positive value) for the binding of 30S ribosome to mRNA, according to Equation 5.4.

$$
\Delta G_{\text{standby}} = \Delta G_{\text{distortion}} + \Delta G_{\text{unfolding}} + \Delta G_{\text{sliding}}
$$

Equation 5.4 summarizes the three energetic interactions that can take place during 30S binding to standby site (12). First, the geometry of standby site determines its accessibility ($A_S$) for platform binding. The 30S platform undergoes a conformational change to accommodate the standby site with low accessibility, but with a distortion free energy penalty $\Delta G_{\text{distortion}}$ to sustain the distorted state. Second, the 30S platform can partially unfold RNA structures only to minimize its binding energy penalty to the structured 5' UTR. Unfolding base pairs increases the accessibility ($A_S$), and lowers $\Delta G_{\text{distortion}}$, but at the expense of an RNA unfolding free energy penalty $\Delta G_{\text{unfolding}}$. Third, the 30S platform can bind at upstream standby sites and slide on the mRNA surface toward the start codon. However, hairpins impeding the 30S migration penalize the 30S platform according to their height. A longer hairpin requires more sliding energy penalty $\Delta G_{\text{sliding}}$ to be pushed aside by the 30S platform.

5.3. Results

5.3.1. Design of expression systems to investigate ribosome footprint

A main challenge in studying secondary structures within the N-terminal coding section is that their effect on translation initiation is strongly coupled to the codons usage efficiency during elongation. To investigate the exclusive effects of RNA secondary structures, we carefully designed a specific expression system, in which a 39-nucleotide AC-rich insulator sequence was fused to the N-terminal coding section of mRFP1 fluorescent protein (Figure 5-1B). The
insertion of insulator sequence had two main advantages. First, it created a fully single-stranded ribosome binding site by preventing the undesired long-range base-pairing across the start codon. Second, its nucleotide sequence had been chosen to only encode for fast codons to avoid a rate limiting elongation step. Next, we introduced RNA secondary structures within the insulator sequence at increasing distances from start codon (Figure 5-1C). By varying the RNA hairpin position we sought to determine when an RNA hairpin is absolutely necessary to be unfolded. These RNA hairpins were systematically designed to be thermodynamically stable, fold fast (average folding times as computed using Kinfold program (145) are listed in Supplementary_Data_5), encode only for fast codons (codon usage frequency ratios were greater than 0.16), and have no more than eight consecutive base pairing to minimize mRNA degradation (55). In addition, the designed RNA hairpins contained either a short (6 bp) or a long (11 bp) hairpin stem, while their loop length varied between 6, 7, or 8 nucleotides to maintain in-frame translation for mRFP1 protein.

5.3.2. RNA hairpins that overlap with ribosomal footprint substantially inhibit translation initiation

We first designed 16 mRFP1 N-terminal variants, in which a single RNA hairpin was introduced at positions between +4 and +40. The RNA hairpins contained a 6 bp stem, with folding energies between -9.5 and -11.0 kcal/mol. The direct effect of these RNA hairpins on translation initiation was characterized by measuring the steady-state fluorescence level of mRFP1 variants using long-run cultures maintained at exponential growth phase (Materials and methods section). The fluorescence level of the no-hairpin control mRNA was also measured under the identical growth conditions.
The absence of secondary structure resulted in a high fluorescence level for the no-hairpin control mRNA (62027.2 ± 6971.4 au) (dashed line, Figure 5-2A). However, introducing an RNA hairpin next to the start codon (position +4) significantly dropped the fluorescence level by 2800-fold (22.8 ± 12 au). We observed the same inhibitory effect even when the RNA hairpin was moved 7 nucleotides further downstream to the position +11, where the measured fluorescence level was only 49.4 ± 7.3 au (Figure 5-2A). These minimum fluorescence levels indicate that 30S ribosome unfolds any RNA hairpin that overlap with its footprint, regardless of hairpin’s distance from the start codon. We next observed a sharp increase in fluorescence level by over 40- and 65-fold when RNA hairpin was moved from position +11 to positions +12 and +13, respectively. However, these measured fluorescence levels were still much lower than the no-hairpin control mRNA (Figure 5-2A), suggesting that RNA hairpins have been partially unfolded by 30S ribosome as they reached a less inhibitory distance from start codon at the edge of ribosome footprint.

Figure 5-2. Position of N-terminal secondary structures regulates the extent of translation inhibition.

Presence of (A) a short, 6 base pair RNA hairpin and (B) a long, 11 base pair RNA hairpin at the N-terminal coding section of mRFP1 reporter protein caused a large sigmoidal increase in fluorescence level.
The measured fluorescence level of no-hairpin control mRNA is shown by dashed line. Data are the average of 2 measurements in two separate days (n=4), and are listed in Supplementary_Data_5.

When the RNA hairpin was moved one more nucleotide downstream to the position +14, the fluorescence level reached its maximum (91506.6 ± 7161.0 au), and remained at a plateau even when the RNA hairpin was moved as far as 30 nucleotides away from the start codon (Figure 5-2A). These high fluorescence levels indicate that RNA hairpins are located outside the ribosome footprint, and therefore are not inhibitory for initiation anymore. In addition, RNA hairpins were not inhibitory for elongation as well, confirming the presence of a helicase activity for elongating 70S ribosomes. Interestingly, folding of RNA hairpins outside the ribosome footprint seems to be beneficial for translation, as the fluorescence levels of mRNAs with RNA hairpins were higher than the no-hairpin control mRNA. Overall, we observed up to 6000-fold sigmoidal increase in fluorescence levels across a wide range of RNA hairpin positions. This sigmoidal trend is consistent with previously reported in vitro assays (2, 16).

5.3.3. Ribosome footprint unfolds RNA hairpins regardless of their size

We next investigated whether the size of RNA hairpins controls their inhibitory effect on the mRNA loading process. We designed 9 mRFP1 N-terminal variants, where exclusive long RNA hairpins (11 bp stem, and folding free energies between -17.2 and -18.6 kcal/mol) were located at positions +8 to +22. The measured fluorescence levels showed a similar sigmoidal increase in gene expression as long RNA hairpins were shifted away from start codon (Figure 5-2B). When located inside the ribosome footprint, the long RNA hairpins had relatively similar inhibitory effect on translation initiation as the short RNA hairpins (over 1700-fold reduction with respect to no-hairpin control mRNA). However, mRNAs with long hairpins reached the
maximum fluorescence level at position +16 (67126.0 ± 5304.6 au), which is 2 nucleotides longer than when short hairpins were used. This difference suggests that the size of RNA hairpin may influence the ribosome footprint coverage but not the amount of work needed to unfold hairpin and load mRNA.

5.3.4. Determination of ribosome footprint length using biophysical model calculations

The sigmoidal increase in fluorescence level was observed at positions +12 and +14 when using short and long RNA hairpins, respectively, indicating that ribosome footprint coverage depends on the overlapping RNA hairpin. To determine the exact ribosome footprint coverage we employ a biophysical model of translation initiation (Materials and methods section) that calculates the total Gibbs free energy change between a free mRNA and a 30S-bound mRNA, $\Delta G_{\text{total}}$, and relates it to translation initiation rate, $r$, according to Equation 5.1 (Figure 5-1A). Several interactions contribute to $\Delta G_{\text{total}}$, including folding free energy of unbound mRNA, mRNA-ribosome hybridization free energy, and energy penalties for non-optimal spacing lengths and inaccessible standby sites (Materials and methods section). Importantly, when a 30S ribosome assembles on mRNA, all secondary structures that overlap with ribosome footprint must be unfolded, while other non-clashing structures are allowed to refold, including the post footprint region on the coding sequence. The folding free energy of post footprint region, $\Delta G_{\text{post-footprint}}$, strongly depends on the ribosome footprint length and can play a significant role in $\Delta G_{\text{total}}$ calculations. Therefore, by varying the ribosome footprint length in the biophysical model calculations and comparing model predictions with the measured experimental data, we sought to identify the optimal ribosome footprint length. We should also emphasize that, to maintain the entirety of RNA hairpins in model calculations, we varied the mRNA 3’-cutoff between 76 and
112 nucleotides, depending on the size and position of RNA hairpins introduced to the N-terminal coding section (Supplementary_Data_5).

Figure 5-3. Determination of ribosome footprint length. Predictions of a biophysical model using wide range of footprint lengths (0 to 35 nt) are compared to the measured fluorescence levels when (A) a short 6 bp hairpin and (B) a long 11 bp hairpin are located at diverse positions in mRFP1’s N-terminal coding section. Red color is the measured fluorescence level. Numbers are the footprint length used in model calculations. (C) Footprint length of 15nt minimizes the average error in predicted ∆G_{total} across 26 characterized mRNAs in this study. (D) Incorporation of ribosome footprint into biophysical model calculations increases its accuracy for 496 previously characterized mRNAs (R^2 = 0.66, p = 1.1x10^{-116} and R^2 = 0.81, p = 1.8x10^{-177} for footprint length of 0 and 15 nt, respectively).
For each mRFP1 variant, we calculated the translation initiation rates by varying ribosome footprint length from 0 to 35 nt. All model predictions at different footprint lengths are listed in Supplementary_Data_5. We found that ribosome footprint length of 14 and 15 nt most accurately predict the observed sigmoidal increase in translation initiation rate when short and long hairpins are used, respectively (Figure 5-3AB). This one nucleotide difference between predicted footprint lengths can be due to the breathing of the closing base pairs at the bottom of different hairpins. Footprint length of 14 and 15 nt also yielded minimum average error in predicted ΔG_{total} across all 26 characterized mRFP1 variants (Figure 5-3C), where error for each mRNA was calculated according to ΔΔG_{total} = abs(ΔG_{total,apparent} – ΔG_{total,predicted}). The apparent ΔG_{total} was calculated using formula: ΔG_{total,apparent} = - ln (fluorescence/K)/β. K is a proportionality constant.

Next, we used a large set of 496 previously characterized mRNA sequences (12, 21, 28, 144) to further examine the accuracy of our identified ribosome footprint length. In contrast to the carefully designed mRNA sequences in this study, this large set of mRNA sequences have dissimilar 5’ UTR and protein coding sequences, and their coding sections contained RNA structures with diverse shapes, energies, and positions, providing a stringent test case for the identified ribosome footprint length. We predicted the translation initiation rates of this large set of mRNAs both with and without ribosome footprint assumption (footprint length = 0 and 15 nt) (Supplementary_Data_5), and compared model predictions with their measured expression levels. We found that incorporation of the correct ribosome footprint length (15 nt) increased the accuracy of biophysical model predictions such that the average error in ΔG_{total} calculations dropped from 6.46 to 1.89 kcal/mol, and consequently increased the accuracy of model predicted translation initiation rates (R^2 = 0.66, p = 1.1x10^{-116} versus R^2 = 0.81, p= 1.8x10^{-177} for footprint
length of 0 and 15 nt, respectively) (Figure 5-3D). For comparison, the model predictions using footprint lengths of 14 and 15 nt were identical.

5.3.5. Overlapping RNA hairpins inhibit translation initiation according to their folding free energy

Previous studies have suggested that reducing RNA stability at the gene start increases translation initiation rate. To test whether RNA stability can also affect the ribosome footprint’s unfolding capability, we designed and characterized a new mRFP1 variant containing a weakened RNA hairpin at position +13 (Figure 5-4A). According to the model calculations, 30S footprint unfolds only three base pairs of RNA hairpins located at position +13. The two G-C base pairs at the base of a stable RNA hairpin were replaced with the A-U base pairs, while maintaining the fast codon requirement. This new mRFP1 variant showed 9.5-fold increase in fluorescence level compared to its GC-rich counterpart, indicating that a weaker secondary structure is thermodynamically more favorable for 30S ribosome to unfold (Figure 5-4B). Consistent with the measurements, the biophysical model also predicted a 5.3-fold increase in translation initiation rate with respect to its original GC-rich mRFP1 variant (Figure 5-4B), confirming that folding free energy controls the extent of hairpin unfolding by the ribosome footprint.

We further investigated how RNA folding energy controlled the translation of our characterized mRNA sequences, for which we calculated the amount of free energy that 30S ribosome expended to unfold inhibitory hairpins ($\Delta G_{\text{unfold,apparent}}$) using the following formula:

$$
\Delta G_{\text{unfold,apparent}} = -\frac{1}{\beta} \ln \left( \frac{X}{X_{\text{ref}}} \right)
$$

(5.5)
where $X$ is the fluorescence level of a sample mRFP1 with structured N-terminal coding section and $X_{\text{ref}}$ is the fluorescence level of no-hairpin control mRNA. We also computed the model-predicted unfolding free energies using the same formula in Equation 5.5, but instead of fluorescence levels we used the predicted translation initiation rate $r$ (Equation 5.1). We found that the computed unfolding free energies are proportional to the apparent unfolding free energies ($R^2 = 0.68, p = 4.5 \times 10^{-5}$) (Figure 5-4C), validating our previous results that hairpin folding free energy controls the extent of inhibition. A very stable RNA hairpin needs more work to be unfolded than a weakly stable RNA hairpin, and therefore inhibits translation initiation with higher extent.

Previous works also aimed to relate the stability of N-terminal coding section to the efficiency of ribosome binding process. Kudla et al. suggested that folding free energy of an mRNA region spanning from -4 to +37 correlates with translation rate (143). Similarly, Kosuri et al. and Goodman et al. also related the fold change in protein expression level to the stability of mRNA regions spanning from the transcription start site (TSS) to either +30 or +96 nucleotides, respectively (14, 50). To test whether these relationships can explain the diverse protein expression levels observed for our designed mRNA sequences, we calculated the folding free energy of each mRNA sequence using their suggested ranges. We found that our characterized mRNA sequences did not follow these sequence-independent relationship ($R^2 = 0.24, 0.41, \text{ and } 0.06$ for Kudla et al., Kosuri et al. and Goodman et al., respectively) (Figure 5-4DEF).

An mRNA 3’-cutoff at +37 position is too short to include the entire sequence of an RNA hairpin inside the coding section, especially when it is long or is shifting away from start codon, causing an inaccurate calculation for the true RNA folding free energy. On the other hand, a constant mRNA 3’-cutoff at +96 position is long enough to include most of RNA hairpins in N-terminal coding section, but it does not distinguish the inhibitory RNA hairpins that overlap with ribosome footprint from non-inhibitory RNA hairpins, resulting in an inaccurate prediction for
ribosome binding efficiency. These results indicate that a constant mRNA cutoff may exclude relevant structural information from folding energy calculation, and the entire RNA hairpin should be used to predict translation initiation rate.

**Figure 5-4. RNA hairpin’s folding free energy controls translation inhibition.** (A and B) A weakly stable RNA hairpin was located at position +13, showing a less inhibition in translation rate than its stable RNA hairpin counterpart, which was also confirmed by the biophysical model predictions. (C) The apparent unfolding energy penalties strongly correlate with the predicted unfolding free energies of overlapping RNA hairpins ($R^2 = 0.68, p = 4.5 \times 10^{-5}$). Yellow and blue circles are the short and long RNA hairpins, respectively. 17 mRNA variants with overlapping RNA hairpins are shown. (D and E) The folding free energy of mRNA region spanning either from -4 to +37, or from transcription start site (TSS) to +30, respectively, correlate poorly with the measured fluorescence levels ($R^2 = 0.24, p = 0.01$; and $R^2 = 0.41, p = 3.1 \times 10^{-4}$). (F) The difference in folding free energy of mRNA region spanning from TSS to +96,
with respect to the no-hairpin control mRNA, does not correlate with the relative fold change in fluorescence level ($R^2 = 0.06$, $p = 0.23$). (D,E,F) Color codes are the same as C.

### 5.4. Discussion and conclusion

In contrast to the 5’ UTR sequence that only controls translation initiation (4, 8, 12, 21, 26, 44, 72, 146), mRNA’s coding section interacts with ribosome through several coupled processes including initiation and elongation. Studying the direct effect of each process requires decoupling from other confounding interactions. Using our carefully designed expression systems, we investigated how N-terminal coding sequence exclusively controls translation initiation, while minimizing the effects of other interactions such as elongation, pausing, and mRNA degradation.

Presence of a stable short RNA hairpin (-11 kcal/mol) inside the 30S ribosome footprint, even at 7 nucleotides downstream from start codon, inhibited translation by over 2800-fold with respect to a fully single-stranded mRNA (Figure 5-2A). We observed a sigmoidal increase in translation when the hairpin shifted away from the start codon until translation reached a maximum, indicating that ribosome footprint covers a limited portion of N-terminal coding section. The fluorescence levels remained at maximum plateau when RNA hairpin was shifted to the position +34, where the level of plateau was higher than the no-hairpin control mRNA. It was only at position +40 that the fluorescence level dropped below the no-hairpin control mRNA ($24743.9 \pm 2737.8$ au) (Figure 5-2A). We hypothesize that the presence of a post-footprint RNA hairpin within the AC-rich insulator region discontinued a long stretch of single-stranded region that is otherwise present in the no-hairpin control mRNA, resulting in a higher mRNA stability and elevated fluorescence level. A single-stranded region greater than 20 nt may be long enough
to promote mRNA degradation for hairpin at position +40. We also do not rule out the possibility of changing the protein folding efficiency, due to the additional peptides attached to the mRFP1 peptide sequence as a result of the RNA hairpin insertion.

A long stable RNA hairpin (-17.2 kcal/mol) showed a similar sigmoidal increase in fluorescence level than the short RNA hairpin (Figure 5-2B). However, the maximum fluorescence level of mRNAs with long hairpins was lower than both the no-hairpin control mRNA and the mRNAs with short hairpins. In particular, the maximum fluorescence levels showed a decline as the long RNA hairpins were moved away from ribosome footprint. These observations suggests that, in contrast to the short RNA hairpins, long RNA hairpins may have an inhibitory effect on elongation rate, which could be due to the lack of sufficient energy to translocate ribosomes through such highly stable hairpins (-17.2 kcal/mol). In particular, our results could be explained using a physics-based model proposed by Qu et al. (20), which quantifies the ribosome translocation rate when facing RNA hairpins within coding section. However, the exact applied force to unfold either short or long RNA hairpins needs to be measured for the model calculations, which is beyond the scope of this study.

By comparing our experimental data with the predictions of a biophysical model of translation initiation (12, 21), we determined that 30S ribosome covers either 14 or 15 nucleotides of coding section during in vivo initiation (Figure 5-3ABC), which is consistent with previous in vitro studies (2, 16, 20). Incorporation of the predicted footprint length to the biophysical model calculations resulted in a higher model accuracy for a large set of 496 previously characterized mRNAs (Figure 5-3D). We also showed that the extent of translation inhibition is proportional to the hairpin folding free energy such that a more stable RNA hairpin inside the ribosome footprint inhibits translation with greater extent (Figure 5-4C). However, our calculations show that the apparent energy exerted by 30S ribosome to unfold N-terminal RNA hairpins is 1.6-fold higher than the unfolding energies predicted using conventional RNA energy models (1.6 is the slope in
Figure 5-4C). This finding was rather surprising because 30S’s input work for unfolding hairpins that sequester SD and start codon correlates well with their predicted folding free energies (21). The reasons why 30S ribosome functions differently in the presence of overlapping RNA hairpins at N-terminal coding section is currently unknown and required further investigations.

Overall, our computational design approach provides a useful method for studying the effect of mRNA’s coding section, and provides initial steps for building an initiation-elongation integrated model for translation rate. In addition, our results can facilitate rational design of translation-regulating riboswitches or small RNAs that target N-terminal coding section for tunable control of protein synthesis (147, 148).
Chapter 6

Automated Physics-Based Design of Synthetic Riboswitches from Diverse RNA Aptamers

This chapter has been previously published in Nucleic Acids Research as a NAR Breakthrough Article (149). This work was done in collaboration with Dennis M. Mishler and Jingzhi Wang from Department of Chemistry at Emory University, and Walker Huso from Penn State University. All the measurements and model predictions for this work are presented in Supplementary_Data_6 and can be found online at:


6.1. Introduction

Riboswitches are RNA-based sensors that use an aptamer domain to bind a chemical ligand, change shape, and alter gene expression level; typically, by modulating translation initiation or transcriptional termination of mRNAs (150–152). While natural riboswitches have evolved as exquisite sensors, it remains difficult to engineer non-natural riboswitches that utilize different aptamers to detect and respond to ligands of interest, even though their application as biosensors or medical diagnostics would be transformative (153, 154). To date, hundreds of aptamers have been discovered, or generated by SELEX (systematic evolution of ligands by exponential enrichment), to bind specifically to cellular metabolites (cofactors, nucleotides, amino acids), chemicals of interest (antibiotics, explosives, pesticides, toxins), and proteins indicative of infection or disease (HIV envelope protein, human RUNX1 tumor suppressor) (155–
However, most aptamers have not been converted into functional riboswitches as existing approaches have relied on qualitative design by experts, combinatorial library generation, and high-throughput screening, which have limited the breadth of riboswitch-based applications (37, 161–168). In particular, high-throughput screening for functional riboswitches cannot be carried out when using ligands that are cytotoxic, insoluble, highly reactive, or impermeable to cell membranes.

Computational RNA design offers a scalable approach to solve these challenges. For example, small RNAs have been rationally engineered to control gene expression through quantification of their RNA-RNA interactions and activation barriers (57, 59, 169, 170). Theophylline-binding riboswitches have been engineered by modeling the kinetics of their binding interactions (37). The translation initiation rates of mRNA sequences can be predicted using a thermodynamic model of ribosome-RNA interactions, enabling the automated design of synthetic ribosome binding sites to control protein expression (21, 42). Such RNA devices, including ribozymes and aptazymes, can be modeled and readily combined to tune protein expression levels, engineer signal-responsive genetic circuits, and dynamically control metabolic pathways (43). However, a model has yet to be developed that can predict a riboswitch's function from its sequence, using in principle any RNA aptamer. Creating such a model from thermodynamic first principles would critically test our understanding of the physics of ligand-induced RNA conformational changes, while enabling the design and optimization of a toolbox of riboswitch-based RNA devices that can sense different ligands.

We therefore created a sequence-to-function physics-based model and automated optimization algorithm that converts RNA aptamers into synthetic riboswitches. As will be explained in this chapter, we validated our computational approach by characterizing 77 synthetic riboswitches with highly diverse RNA aptamers, sequences, and structures, performing both in vivo and in vitro cell-free measurements to quantify their function. To illustrate the versatility of
our approach, we selected RNA aptamers that bind specifically to both ideal and non-ideal ligands, including theophylline, tetramethylrosamine (TMR), fluoride, dopamine, thyroxine, and 2,4-dinitrotoluene (DNT). Our automated approach generated functional riboswitches with high activation ratios, up to 383-fold. The validated model provides a comprehensive, quantitative, and falsifiable physical mechanism for riboswitch function, including how changing the aptamer's structure and affinity, surrounding mRNA sequence, switching free energy, ligand and mRNA concentrations, and macromolecular crowding collectively control riboswitch activation.

6.2. Materials and methods

6.2.1. Plasmids and strains

All riboswitch and promoter sequences are listed in Supplementary_Data_6. Theophylline, fluoride, and DNT riboswitches were constructed and inserted into an mRFP1 fluorescent protein expression vector, derived from plasmid pFTV1 (ColE1 origin, Cm<sup>R</sup>) (21). Three theophylline riboswitches (Theo-40, Theo-41, and Theo-45) also controlled the translation of a fusion mRFP1 protein. To create the fusion protein, 4 or 5 non-rare codons were introduced between the start codon and SacI restriction site within mRFP1 coding section. All the riboswitches were constructed using standard molecular cloning. Briefly, DNA fragments were computationally designed, synthesized, and assembled using either annealing of oligonucleotides, PCR assembly of oligonucleotides, or PCR amplification of gBLOCK DNA fragments (Integrated DNA Technologies). DNA fragments were then digested by XbaI and SacI restriction enzymes, followed by ligation with digested plasmid, transformation, plating on selective media, and verification of purified plasmid by sequencing. Similarly, promoter replacements were performed by annealing designed pairs of oligonucleotides, followed by digestion with...
BamHI/XbaI restriction enzymes, ligation, transformation, selective plating, and verification by sequencing.

The promoters AEB-3, J23100, and LmrA were selected or designed to significantly vary riboswitch transcription rates (Figure B-17). The promoter AEB-3 is a result of mutating the -10 and -35 hexamers of promoter J23100, resulting in 10-fold lower transcription rate. The promoter LmrA is a near-consensus promoter with a 5-fold higher transcription rate. Unless noted otherwise in the text, all theophylline and fluoride riboswitches use the J23100 promoter, while all DNT riboswitches used the AEB-3 promoter.

Theophylline and TMR riboswitches were then constructed in plasmids expressing the luciferase reporter protein using standard molecular cloning. The plasmid is derived from the pBESTluc vector (Promega) initially using a pUC19 origin, where the riboswitch-reporter mRNA is transcribed by a tac promoter. As described in the text, plasmid origins were replaced with either pBAC, p15A, or pFTV1 by PCR amplifying the expression cassette (promoter to transcriptional terminator) and digesting it with BamHI and SpeI, followed by ligation to the corresponding digested vectors, transformation, selective plating, and verification by sequencing.

Dopamine and thyroxine riboswitches were constructed in pFTV1-derived plasmids containing the luciferase expression cassette, where pFTV1 was previously modified to insert AatII and HindIII restriction sites after the tac promoter and after the start codon, respectively. Riboswitch-encoding DNA fragments were PCR-amplified from designed gBLOCKs, followed by digestion with AatII and HindIII, ligation to digested plasmids, transformation, selective plating, and verification by sequencing.
6.2.2. Ligand, media, and buffer conditions

Theophylline, tetramethylrosamine (TMR), sodium fluoride, dopamine, thyroxine, 2,4-Dinitrotoluene (DNT), DMSO, ascorbic acid, and sodium chloride were purchased from Sigma-Aldrich. Luria broth Miller (LB) media was purchased from VWR. All ligand conditions are provided in the Supplementary_Data_6.

DNT has a solubility of 1.48 mM (270 µg/mL) in water. To increase DNT solubility, 1% (v/v) DMSO was added to the media when characterizing DNT riboswitches and no-aptamer controls using the 0 mM DNT and 1 mM DNT conditions. Dopamine is enzymatically oxidized by E. coli DH10B cells, and spontaneously oxidized by molecular oxygen. To minimize dopamine oxidation, 5 mM ascorbic acid was added to cell-free transcription-translation assays to characterize dopamine riboswitches and no-aptamer controls using the 0 and 1 mM dopamine conditions. Thyroxine has a solubility of 135 nM (0.105 µg/mL) in neutral water. To increase thyroxine solubility, 10 mM NaOH was added to create a 1 mM thyroxine working stock solution, followed by adding 2 µl of stock solution to 11 µl cell-free transcription-translation assays (154 µM thyroxine). 100 mM NaCl was then added to cell-free transcription-translation assays to maintain physiological RNA folding. Characterization of thyroxine riboswitches and no-aptamer controls took place in the same salt and pH conditions, using either 0 and 154 µM thyroxine.

6.2.3. Growth, fluorescence, and luminescence measurements

All in vivo fluorescence measurements were carried out using the strain E. coli DH10B during long-time cultures that maintain cells in the exponential growth phase. Cells harboring cloned plasmids expressing the mRFP1 reporter were grown overnight in 700 µl LB media and
50 µg/ml Cm antibiotic in a 96 deep-well plate at 37 °C and 200 rpm orbital shaking. Cultures were then diluted to an OD_{600} of 0.01 into 200 µl LB media, 50 ug/ml Cm, and pre-defined concentrations of ligand and buffer solution (see Supplementary_Data_6 for all ligand concentrations). OD_{600} and fluorescence values were recorded by a M1000 spectrophotometer (TECAN) every 10 min until the OD_{600} reached about 0.15. Cells were then diluted to 0.01 into fresh media with the same composition, and grown until reaching mid-exponential phase. A third dilution was carried out in the same way. Time course fluorescence measurements per cell (FLPC) were analyzed by evaluating the equation:

\[
\text{FLPC} = \frac{F - F_M}{OD_{600} - OD_{600,M}} - \frac{F_{NC} - F_M}{OD_{600,NC} - OD_{600,M}}
\]

(6.1)

where F is the fluorescence of cells, F_M is the fluorescence of media, F_{NC} is the auto-fluorescence of non-transformed cells, OD_{600} is the optical density of cells, and OD_{600,M} is the optimal density of media. Figure B-8 shows examples of FLPC time series data.

Single-cell fluorescence levels were recorded by a Fortessa flow cytometer (BD Biosciences) and used to calculate activation ratios. During growth, and before each dilution, 10 µL samples were taken and added to 200 µL PBS with 2 mM kanamycin antibiotic for measurement. All distributions were unimodal. Auto-fluorescences of non-transformed cells were subtracted. Averages and standard deviations calculated from at least two independent measurements. The number of measurements for each riboswitch are listed in Supplementary_Data_6.

The in vivo luminescence measurements for theophylline and TMR riboswitches were conducted by transforming E. coli DH10B using plasmids expressing the luciferase reporter and measuring luminescence in a Biotek Synergy HT plate reader, as described in Ref (163). Briefly, cells were diluted 1:500 in fresh LB/AMP from an overnight culture and grown to mid-exponential phase. Cultures were then prepared for measurement using dry ice and 1x CCLR.
(Promega) following the manufacturer’s procedure. The luminescence data are normalized relative to OD$_{600}$ values at the time of harvesting.

### 6.2.4. Cell-free transcription-translation assays and luminescence measurements

Cell-free transcription-translation assays were conducted using the S30 *E. coli* extract (Promega, L1020). TMR riboswitches were characterized by adding 20 ng/µl plasmid DNA to cell-free transcription-translation assays, and incubating at 37 °C for 30 minutes. Quenching was carried out by placing the reaction on ice and adding dilution buffer. To measure luciferase activity, equal volumes of cell-free product and luciferase reagent (Promega) were combined in black, opaque 96-well plates. Luminescence was recorded using a Biotek Synergy HT plate reader with a sensitivity of 150, and divided by amount of added plasmid. The same procedure was applied when adding increasing plasmid concentrations for the three TMR riboswitches. Dopamine and thyroxine riboswitches were characterized similarly, except using white, opaque 96-well plates, a 60 minute incubation period, and a SpectraMax M5 plate reader with an integration time of 1500 ms.

### 6.2.5. Transcription-translation decoupling experiment for TMR riboswitches

Decoupling experiments were conducted as previously described (163). Briefly, 40 ng/µl of DNA template was used, as for a normal cell-free reaction. Amino acids were omitted during the initial 5 minute incubation. Rifampicin was then added at 250 µg/ml final concentration in the reaction to stop transcription. After a one minute incubation, amino acids were added to start translation. 40 µM TMR was added at time 0 or 6 min. The mixtures were incubated for up to 10 minutes after translation has started before quenching and measuring luminescence.
6.3. Results

6.3.1. A biophysical model of translation-regulating riboswitches

The biophysical model considers a riboswitch as a long mRNA whose RNA-RNA, RNA-ribosome, and RNA-ligand interactions control its translation initiation rate, and thereby the output protein's expression level (Figure 6-1A). As a single mRNA is transcribed, its folding trajectory is dictated by contact with ligand (Figure 6-1B). In the absence of ligand, the mRNA freely folds towards its minimum free energy structure (state 1) and binds to the ribosome with free energy $\Delta G_{\text{total,OFF}}$. We calculate the ribosome's binding free energy to the mRNA in state 1 ($\Delta G_{\text{total,OFF}}$) using the RBS Calculator v2.0 model (12, 21) (Figure 6-2A), which includes the energy needed to unfold inhibitory mRNA structures ($\Delta G_{\text{mRNA}}$); the energy released when a canonical or non-canonical Shine-Dalgarno (SD) sequence binds to the last nine nucleotides of the 16S rRNA ($\Delta G_{\text{mRNA-rRNA}}$); an energetic penalty when the ribosome is stretched or compressed, due to a non-optimal distance between the SD and start codon ($\Delta G_{\text{spacing}}$); the energy released when the tRNA$^{\text{fMet}}$ binds to the start codon ($\Delta G_{\text{start}}$); and an energetic penalty controlled by the single-stranded surface area of upstream standby sites ($\Delta G_{\text{standby}}$) (12). RNA-RNA interactions are calculated using established RNA free energy models that account for both secondary structures (52) and pseudoknots (171, 172) (see also supplementary notes in Appendix B.2.4).

In the absence of ligand, the mRNA's translation initiation rate is predicted according to $r_{\text{OFF}} = \exp(-\beta \Delta G_{\text{total,OFF}})$ on a 100,000-fold proportional scale. When the ligand is available, a newly transcribed mRNA can rapidly bind to it via co-transcriptional folding (path a), reaching an intermediate structure that contains the ligand-bound aptamer (state 2), and continuing to fold (path b) towards an alternate minimum free energy structure (state 3) that now can bind more...
tightly to the ribosome with free energy $\Delta G_{\text{total,ON}}$ (Figure 6-1B). In the mRNA’s state 3, the RNA aptamer's structure is now locked into a ligand-bound state, while the remaining portions of mRNA are allowed to freely fold towards their minimum free energy structure (Figure 6-2B).

We then use the same free energy model to calculate the ribosome's binding free energy to mRNA’s state 3, $\Delta G_{\text{total,ON}}$, and its translation initiation rate $r_{\text{ON}} = \exp(-\beta \Delta G_{\text{total,ON}})$, where the differences between $\Delta G_{\text{total,ON}}$ and $\Delta G_{\text{total,OFF}}$ arise from ligand-induced changes in $\Delta G_{\text{mRNA-rRNA}}$, $\Delta G_{\text{mRNA}}$ and $\Delta G_{\text{standby}}$. As the model's first generalized sequence-to-function prediction, the maximum increase in a riboswitch's translation initiation rate (maximum activation ratio, AR$_{\text{max}}$) is determined by the difference in the ribosome's binding free energy, according to $\text{AR}_{\text{max}} = r_{\text{ON}} / r_{\text{OFF}} = \exp(-\beta [\Delta G_{\text{total,ON}} - \Delta G_{\text{total,OFF}}])$. The maximum possible activation is only obtained when co-transcriptional folding is present, when excess amounts of ligand are added, and when the mRNA-ligand complex (state 3) is stable.

Figure 6-1. A biophysical model of riboswitch regulation. (A) A biophysical model was developed to predict the function of translation-regulating riboswitches. The model uses the riboswitch's mRNA sequence, the aptamer's ligand-bound structure, the concentrations of the mRNA and ligand, and their molar volumes to calculate the riboswitch's translation initiation rates in the ON and OFF states ($r_{\text{ON}}, r_{\text{OFF}}$).
as well as its maximum, concentration-limited, and actual activation ratios (AR\textsubscript{max}, AR\textsubscript{conc}, AR\textsubscript{actual}). (B) A reaction coordinate diagram showing the states, energies, and transition paths during mRNA co-transcriptional folding, ligand-binding, and translation initiation. Without ligand, the ribosome binds to a folded mRNA (state 1) with free energy $\Delta G_{\text{total,OFF}}$. When excess ligand is present, the co-transcriptional folding of mRNA and ligand through paths a and b ends with a mRNA-ligand complex (state 3) that binds the ribosome with free energy $\Delta G_{\text{total,ON}}$. The stability of the mRNA-ligand complex is controlled by the switching free energy ($\Delta \Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}$).

6.3.2. Initial validation of the biophysical model

The theophylline RNA aptamer has been frequently utilized to engineer riboswitches, as theophylline binds tightly to its aptamer and can be added to excess without toxicity (37, 43, 152). We therefore initially tested the model's predictions for AR\textsubscript{max} by characterizing 15 previously designed theophylline-binding riboswitch variants that regulate luciferase expression (163) (Figure 6-2C), measuring their steady-state in vivo luminescences after adding either 0 or 2 mM theophylline. The calculated ribosome binding free energies, $\Delta G_{\text{total,OFF}}$ and $\Delta G_{\text{total,ON}}$, correctly predicted the changes in translation rate and luminescence ($R^2 = 0.68$, $p = 2 \times 10^{-8}$), according to the expected Boltzmann log-linear relationship (Figure 6-2D). The apparent slope ($\beta = 0.42 \pm 0.03$ mol/kcal) is similar to our previous translation rate measurements (12, 21). The riboswitches' activation ratios varied from 1.9 to 326.6-fold and were well-predicted by the model's calculated AR\textsubscript{max} ($R^2 = 0.68$, $p = 1.5 \times 10^{-4}$) (Figure 6-2E). These results suggested that the model's thermodynamic calculations can adequately predict riboswitch activation when a mRNA stably binds to a ligand, changes its structure, and thereby alters the strengths of the ribosome-mRNA interactions controlling translation rate. We then expanded the biophysical model to predict riboswitch activation when the ligand concentration is limiting, when the
mRNA-ligand complex is not stable, and when macromolecular crowding causes the amount of free volume to be limiting.

Figure 6-2. Biophysical model calculations. (A) Without ligand, when a ribosome binds to a folded mRNA, it undergoes several structural changes with corresponding free energy changes in the mRNA-ribosome interactions. Here, the ribosome binds poorly to the mRNA’s initial state because it has an inaccessible standby site and sequestered (brown) Shine-Dalgarno sequence. In the final mRNA state, the ribosome’s 16S rRNA and tRNA\textsuperscript{fMet} hybridize to the Shine-Dalgarno sequence and start codon, respectively, with significant refolding of the mRNA. The model uses the mRNA sequence to calculate the total free energy change. (B) When the ligand is present, the same free energy calculation is performed,
while constraining the (blue) aptamer sequence to its ligand-bound structure. (C) 15 synthetic theophylline riboswitches from Ref. (163) were characterized to test the model's AR\textsuperscript{max} prediction. (D) Predicted $\Delta G_{\text{total}}$ values in the (white circles) OFF and (orange circles) ON states are well-correlated to the measured luminescences according to the expected log-linear relationship ($R^2 = 0.68$, $p = 2 \times 10^{-8}$). (E) The calculated free energy differences $\Delta \Delta G_{\text{total}}$ are well-correlated to the measured activation ratios according to the expected log-linear relationship ($R^2 = 0.68$, $p = 1.5 \times 10^{-4}$). Each data point and bar represents the mean and s.d. of 3 measurements.

6.3.3. Riboswitch stability and concentration-dependence

We determine how changing ligand and mRNA concentrations affect riboswitch activation by calculating the ensemble partitioning of riboswitch mRNA into states 1 and 3 (shown in Figure 6-1B), within a water-filled, constant-volume system where the total amount of available volume is low to mimic the crowded environment found inside cells. The concentrations of mRNA in states 1 and 3 ($C_{\text{mRNA},1}$ and $C_{\text{mRNA},3}$) are solved by minimizing the system's total free energy, constrained by the energies of formation for states 1 and 3, the total mRNA ($C_{\text{mRNA, total}}$), and the total ligand concentration ($C_{\text{ligand, total}}$) as well as their molar volumes ($V_{\text{mRNA}}$ and $V_{\text{ligand}}$) (Figure B-1, and supplementary notes in Appendix B.2.2). At a specified total mRNA and ligand concentration, the riboswitch's activation ratio will be $AR_{\text{conc}} = (C_{\text{mRNA},1} / C_{\text{mRNA, total}}) + (C_{\text{mRNA},3} / C_{\text{mRNA, total}}) AR_{\text{max}}$. 
By converting concentrations to volume fractions and equivalently substituting \((C_{\text{ligand,total}} - C_{\text{ligand,free}})\) for \(C_{mRNA,3}\), we rearranged this equation to explicitly show how the volume fractions and molar volumes of the mRNA and ligand control riboswitch activation (see supplementary notes in Appendix B.2.1):

\[
AR_{\text{conc}} = \frac{V_{\text{mRNA,1}}}{V_{\text{mRNA,total}}} + \frac{V_{\text{mRNA}}}{V_{\text{ligand}}} \left[ \frac{V_{\text{ligand,total}} - V_{\text{ligand,free}}}{V_{\text{mRNA,total}}} \right] AR_{\text{max}}
\]

We next quantify the switching free energy to calculate a riboswitch's activation ratio when its mRNA-ligand complex is not stable. We first consider an alternative structure (state 4) that is not bound by ligand, but is otherwise structurally identical to state 3 (Figure 6-1B). The mRNA in state 4 is not stable because it has a higher free energy than the mRNA in state 1. The difference in free energy between states 1 and 4 is always positive \((\Delta\Delta G_{mRNA} > 0)\). Therefore, for the mRNA's ligand-bound state 3 to become thermodynamically favorable, the ligand must bind sufficiently tight to stabilize the mRNA-ligand complex; the switching free energy must be negative \((\Delta\Delta G_{mRNA} + \Delta G_{\text{ligand}} < 0)\). However, if the switching free energy is positive, then the ligand could disassociate from the mRNA-ligand complex prior to its degradation (path c), leading to state 1, and it will be unfavorable for the ligand to bind again without the presence of co-transcriptional folding (paths d and e). We refer to riboswitches with positive (negative) switching free energies as metastable (stable). Assuming equilibrium between mRNA states 1 and 3, the fraction of mRNA that remains stably bound to the ligand can be readily derived (see supplementary notes in Appendix B.2.1), leading us to the following sequence-to-function expression for the riboswitch's actual activation ratio, \(AR_{\text{actual}}\), whenever the switching free energy \((\Delta\Delta G_{mRNA} + \Delta G_{\text{ligand}})\) is positive:

\[
AR_{\text{actual}} = 1 + \frac{\exp(-\beta(\Delta\Delta G_{mRNA} + \Delta G_{\text{ligand}}))}{1 + \exp(-\beta(\Delta\Delta G_{mRNA} + \Delta G_{\text{ligand}}))} (AR_{\text{conc}} - 1)
\]
Altogether, the inputs into the biophysical model are the riboswitch's mRNA sequence, the structure of the RNA aptamer's ligand-bound state, the protein coding sequence, and the concentrations of the mRNA and ligand. There are no additional free or fit parameters. The model first calculates the translation initiation rates of a riboswitch in its ON and OFF states, $r_{ON}$ and $r_{OFF}$, according to the ribosome's binding free energies in the ligand-bound ($\Delta G_{\text{total,ON}}$) and ligand-free state ($\Delta G_{\text{total,OFF}}$). The model then uses Equation 6.2 to decrease $r_{ON}$ according to the specified mRNA and ligand concentrations. The riboswitches' switching free energy is then calculated and, if it's positive, Equation 6.3 is applied to further reduce $r_{ON}$ to account for metastability of the mRNA-ligand complex. The predicted activation ratio is the ratio between $r_{ON}$ and $r_{OFF}$. The model assumes that the mRNA, ribosome, and ligand reach chemical equilibrium faster than the mRNA's half-life, which requires two key conditions: first, the ligand must rapidly bind to the mRNA, for example, via co-transcriptional folding; and second, the mRNA must rapidly, but reversibly, change shape from its OFF to ON state after the ligand has bound. The model's accuracy also depends on the accuracy of the inputs; for example, the RNA aptamer's ligand-bound structure must be correctly specified. In addition, the model does not account for potential changes in mRNA stability, for example, when the translation rate of the mRNA is so low that it becomes a target for RNase activity.

6.3.4. Automated design of synthetic riboswitches

Using the complete model, we developed an automated optimization algorithm to computationally convert any RNA aptamer into a synthetic riboswitch. The inputs are the RNA aptamer's sequence, its ligand-bound structure, and the output protein's coding sequence. The algorithm then identifies an optimal pre-aptamer sequence and post-aptamer sequence to create a
non-natural riboswitch sequence (**Figure 6-3A**). Starting with a population of randomized pre- and post-aptamer sequences, the genetic algorithm performs rounds of in silico random mutation, prediction, selection, and recombination to efficiently search for pre- and post-aptamer sequences that maximize a selected objective function, such as maximizing $AR_{\text{actual}}$, minimizing the switching free energy, or selecting specific translation rates in the OFF and ON states (**Figure B-2**). The computational search space contains about $10^{36}$ sequences, which greatly surpasses the library generation and measurement capacities of high-throughput approaches (**Figure B-3**, and supplementary notes in **Appendix B.2.5**).

We then employed automated computational optimization to design 52 synthetic riboswitches using 3 RNA aptamers with the objective of testing the model’s ability to correctly predict how different pre- and post-aptamer sequences controlled riboswitch function. Specifically, we selected the theophylline TCT8-4 (173), tetramethylrosamine (TMR) (174), and fluoride 78-Psy (155) aptamers (**Figure 6-3B**), because they have varying ligand binding energies, ligand sizes, and ligand toxicities ($\Delta G_{\text{ligand}} = -9, -10.2,$ and $-6 \text{ kcal/mol}$; $V_{\text{ligand}}/V_{\text{water}} = 10, 20,$ and $1$; $C_{\text{toxicity}} = 5 \text{ mM, 20 } \mu \text{M, and 150 mM for theophylline, TMR, and fluoride, respectively}$). Our choice of RNA aptamers provided appropriate test cases for sensing ligands at low concentrations (TMR) and sensing ligands that bind weakly to their RNA aptamer (fluoride) together with a commonly engineered RNA aptamer that serves as a baseline for comparisons (theophylline). The fluoride RNA aptamer also has a highly pseudoknotted tertiary structure (175), creating a more complex structural constraint. Initially, our optimization algorithm's objective function was tailored to design both stable and metastable riboswitches by varying the targeted switching free energies from $-8.4$ to $11.52 \text{ kcal/mol}$. As a result, 37 of the designed riboswitches have positive switching free energies (**Figure B-4**). The designed riboswitches have dissimilar pre-aptamer and post-aptamer sequences, which do not resemble any natural sequences.
or structures (sequence entropy 63%, Figure B-5). All riboswitch structures were also predicted to change significantly when the ligand is bound (examples shown in Figures B-6 and B-7). 3 of the previously designed theophylline riboswitches also have positive switching free energies.

Figure 6-3. Automated design of synthetic riboswitches. (A) An optimization algorithm converts RNA aptamers into synthetic riboswitch sequences through rounds of mutation, prediction, selection, and recombination. (B) The ligand-bound structures of the theophylline, TMR, fluoride, dopamine
(dopa1.3/c.3), thyroxine (ApT4-J-min), and 2,4-dinitrotoluene aptamers are shown (see also Figure B-16). Pseudoknotted base pairs are indicated by dotted lines. Pre- and post-aptamer sequences varied after optimization. (C) The error in the model's calculated $AR_{conc}$ is compared to the calculated switching free energy, showing that the result of Equation 6.2 loses accuracy as the switching free energy grows. Riboswitches are colored according to their aptamer: (dark blue) 12 designed theophylline riboswitches; (red) 12 theophylline riboswitches with mutated aptamer; (brown) 6 theophylline riboswitches with different mRFP1 coding sequences; (green) 15 previously engineered theophylline riboswitches (163); (yellow) 12 designed fluoride riboswitches; (light blue) 10 TMR riboswitches. The boxed outliers are riboswitches Theo-42 and Theo-44, suggesting a systematic malfunction in their function. (D) Model predictions using Equation 6.3 ($AR_{actual}$) are in good agreement with the measured activation ratios for 59 of 67 riboswitches (Spearman $R = 0.69$, $p = 1.2 \times 10^{-10}$, $N = 67$; Pearson $R^2 = 0.61$, $p = 2.6 \times 10^{-13}$, $N = 59$). (Shaded blue) The model predicted the activation ratios of 37 riboswitches to within 2-fold. (E) The predicted translation initiation rates in the riboswitches' (white circles) OFF and (orange circle) ON states are compared to their measured fluorescence or luminescence levels (Spearman $R = 0.68$, $p = 3.7 \times 10^{-19}$, $N = 134$; Pearson $R^2 = 0.44$, $p = 4.3 \times 10^{-18}$, $N = 134$). Each data point and bar represents the mean and standard deviation of 2 to 4 measurements.

6.3.5. Validating model-predicted requirements for riboswitch activation

To test the model's predictions, we measured reporter expression levels in the riboswitches' ON and OFF states by using spectrophotometry and flow cytometry to record mRFP1 fluorescence and luciferase luminescence levels, with and without added ligand, inside *E. coli* DH10B cells maintained in exponential growth for 12 hours (Materials and methods section and Figure B-8 for time series data). Either 2 mM theophylline, 20 µM TMR, or 150 mM fluoride were added throughout the long-time cultures to measure the riboswitches' ON state expression levels. We performed the same measurements on a set of no-aptamer controls to
measure the ligands' non-specific effects on reporter expression levels under identical *E. coli* growth conditions. For example, 2 mM theophylline and 150 mM fluoride both increased the no-aptamer control's reporter expression by 1.4-fold and 1.3-fold, while 20 µM TMR decreased it by 2.6-fold. We removed these non-specific effects by dividing the riboswitch's measured ON/OFF expression ratio by the no-aptamer control's measured ON/OFF expression ratio, measured under identical growth and ligand conditions. The resulting measurement is the riboswitch's specific activation ratio (AR). Altogether, we validated the model’s predictions by performing these measurements on 67 riboswitches, including the 15 previously designed theophylline riboswitches. All sequences, model calculations, riboswitch measurements, and no-aptamer control measurements are provided in **Supplementary Data 6**.

We first evaluated the original biophysical model that only accounts for the ligand-induced changes in mRNA-ribosome interactions, using our formula for AR\textsubscript{max}. We found that a riboswitch must have a high model-calculated AR\textsubscript{max} for it to successfully activate translation in response to ligand binding. For example, five of our characterized riboswitches have AR\textsubscript{max} less than 3.0; when they bind theophylline, all five activate translation by 3-fold or less. However, by itself, the model-calculated AR\textsubscript{max} is not a quantitative predictor of a riboswitch’s activation ratio as quantified by a log-log comparison across the 67 riboswitches (Pearson R\textsuperscript{2} = 0.002, p = 0.71, N = 67) (**Figure B-9**). Designing a riboswitch with a high model-calculated AR\textsubscript{max} is necessary, but not sufficient, for achieving high activation.

Instead, we found that incorporating concentration-dependence into the biophysical model enabled it to accurately predict riboswitch activation, but only for stable riboswitches (**Figure 6-3C**). Using the formula for AR\textsubscript{conc} (**Equation 6.2**, and supplementary notes in **Appendix B.2.6**), 25 out of the 27 total riboswitches with negative switching free energies had well-predicted activation ratios with two outliers (**Figure B-10A**) (Spearman R = 0.72, p =
2.5x10^5, N = 27; Pearson R^2 = 0.62, p = 3.5x10^-6, N = 25). The activation ratios of all 10 TMR riboswitches were well-predicted, due to their high-affinity RNA aptamers and modestly high structural free energy changes (ΔΔG_mRNA less than 8.4 kcal/mol) (Figure B-10A). In contrast, without accounting for the instability of the mRNA-ligand complex, the model-calculated AR_{conc} could not predict the activation ratios of the remaining 40 metastable riboswitches, which included 28 theophylline riboswitches that undergo energetically costly switching (ΔΔG_mRNA > 9 kcal/mol) and all 12 low-affinity fluoride-binding riboswitches (Spearman R = 0.27, p = 0.09, N = 40; Pearson R^2 = 0.04, p = 0.22, N = 40) (Figure B-10B).

Finally, when we account for the instability of the mRNA-ligand complex, the biophysical model could predict the activation ratios of both stable and metastable riboswitches (Figure 6-3D). Using the formula for AR_{actual} (Equation 6.3), the calculated activation ratios for 59 out of the 67 riboswitches agreed with their measured activation ratios (ARs) across a 105-fold range (Spearman R = 0.69, p = 1.2x10^-10, N = 67; Pearson R^2 = 0.61, p = 2.6x10^-13, N = 59). The eight outliers were theophylline riboswitches, and five of them had higher-than-predicted activation ratios. The model errors in AR_{actual}, r_{ON}, and r_{OFF} were log-normally distributed and 55% of all riboswitches had their activation ratios correctly predicted to within 2-fold (Figure B-11). Moreover, the model's accuracy did not depend on the RNA aptamer employed to engineer the riboswitch; the apparent error in the model's free energy calculations, quantified by the metric |log(AR/AR_{actual})|, was 0.64, 0.41, and 0.59 for the theophylline, TMR, and fluoride riboswitches, respectively.

Interestingly, the riboswitches' activation ratios were more accurately calculated than their translation rates in their ON and OFF states (Spearman R = 0.68, p = 3.7x10^-19, N = 134; Pearson R^2 = 0.44, p = 4.3x10^-18, N = 134) (Figure 6-3E and Figure B-12) as the error in predicting r_{ON} and r_{OFF} for the same riboswitch had a high degree of correlation (Spearman R =
0.78, \( p < 10^{-30}, N = 67 \); Pearson \( R^2 = 0.87, p = 3.6 \times 10^{-27}, N = 59 \). This shared uncertainty factor was eliminated when calculating the \( r_{ON} / r_{OFF} \) ratio (Figure B-13). This source of error could arise from any interaction, not included within our model, that affects riboswitch expression similarly in both ON and OFF states. By analyzing the physical interactions distinguishing the riboswitch inliers from outliers, we also found that riboswitches with lower-than-predicted activation ratios bound much more tightly to the ribosome's 16S rRNA (highly negative \( \Delta G_{SD,\text{antiSD}} \)) and had a high energetic cost to switching (high \( \Delta \Delta G_{\text{mRNA}} \)) (Figure B-14), which could be responsible for their higher-than-predicted \( r_{OFF} \) and lower-than-predicted \( r_{ON} \), respectively.

Overall, by developing a sequence-to-function biophysical model for riboswitches and coupling it to an automated optimization algorithm, we successfully converted the theophylline, TMR, and fluoride aptamers into highly active riboswitches, while performing a small number of measurements (Figure 6-4A, Figure B-15). The model explains why some RNA aptamers are more conducive to riboswitch engineering, compared to others; for example, the highest possible \( AR_{\text{actual}} \) for a fluoride riboswitch is 22-fold lower than for a theophylline riboswitch, due to their differences in ligand affinity and ligand-bound structure. Further, because TMR is toxic and limits its maximum concentration, the highest possible \( AR_{\text{actual}} \) for a TMR riboswitch is 62-folder lower than for a theophylline riboswitch.

### 6.3.6. Design of dopamine, thyroxine, and DNT-binding riboswitches

We next tested whether our computational design approach could convert uncharacterized RNA aptamers that bind non-ideal ligands into functional riboswitches. As a demonstration, we selected the dopamine (dopa2/C.1 and dopa1.3/C.3) (176), thyroxine (ApT4-J-min) (177), and 2,4-dinitrotoluene (DNT) (167, 178) aptamers and ligand targets (Figure 6-3B
Dopamine and thyroxine are rapidly oxidized by bacterial membrane-bound monoamide oxidases (179), and thyroxine is highly insoluble in water, motivating the characterization of such riboswitches in cell-free transcription-translation assays. Dopamine and thyroxine riboswitches have not been engineered yet, and a DNT riboswitch with 10-fold activation was found by elegantly screening a $10^{18}$ sequence library (167).

Figure 6-4. *in vivo* and cell-free characterization of designed riboswitches. (A) For each aptamer, the (circles) measured expression levels in the ON and OFF states and (bars) the activation ratios for the two best riboswitch variants are shown. To characterize their ON states, added ligand concentrations were 2 mM theophylline, 30 μM TMR, 150 mM fluoride, 1 mM DNT, 1 mM dopamine supplemented with 5 mM ascorbic acid, and 150 μM thyroxine supplemented with 100 mM NaCl and 1.5 mM NaOH (see Figure B-15). Each data point and bar represents the mean and standard deviation of 2 to 4 measurements. (B) Cell-
free transcription-translation assays are used to characterize riboswitch function during (yellow) the absence of ligand, (blue) post-transcriptional ligand binding, and (red) co-transcriptional ligand binding. (C) Using cell-free transcription-translation assays, the expression levels of three TMR riboswitches and a no-aptamer control were measured to determine the role of co-transcriptional folding. As described in Ref. (163), expression levels are normalized to the highest luminescence level during the 10 min assay for each individual construct. Each data point and bar represents the mean and standard deviation of 3 to 4 measurements.

Accordingly, we applied our biophysical model and automated optimization algorithm to design 5 dopamine, 2 thyroxine, and 3 DNT riboswitches. We measured the activation ratios of the dopamine and thyroxine riboswitches using cell-free transcription-translation assays and characterized the activation of the DNT riboswitches inside E. coli DH10B cells (Materials and methods section). We also measured the non-specific effects of these ligands by recording the no-aptamer control’s reporter expression levels under identical conditions; 1 mM DNT increased the no-aptamer control’s expression by 4.8-fold, while 1 mM dopamine and 150 µM thyroxine decreased it by 1.67- and 2.5-fold, respectively. As before, the ligands’ non-specific effects were removed from the riboswitches’ measured activation ratios (supplementary notes in Appendix B.2.9). As a result, the best DNT, thyroxine, and dopamine riboswitches achieved activation ratios of 11.1-fold, 2.4-fold, and 2.0-fold, respectively (Figure 6-4A, and Figure B-15), which are the highest reported activation ratios for engineered riboswitches using these RNA aptamers.

In contrast to high-throughput screening, these prototypical examples demonstrate how physics-based design can engineer functional riboswitches even when the ligand is reactive, insoluble, or impermeable to cell membranes.
6.3.7. Co-transcriptional folding is required for riboswitch activation

With additional use of cell-free transcription-translation assays, we next applied our model to evaluate how co-transcriptional folding and macromolecular crowding affects riboswitch function inside cells. First, we characterized the activation of three TMR riboswitches in cell-free assays either without ligand, without co-transcriptional ligand binding, or with co-transcriptional ligand binding (Figure 6-4B and Materials and methods section). When ligand was added after transcription had been stopped, the riboswitches were inactive and behave similarly as the no-aptamer control (Figure 6-4C). The riboswitches only activated reporter expression when ligand was added during transcription. With co-transcriptional ligand-binding, all 10 TMR riboswitches activated similarly in both cell-free and in vivo environments (Figure B-15EF). These results show that co-transcriptional folding is required for reaching a fast mRNA-ligand equilibrium, a key model assumption, which occurs either inside cells or within cell-free transcription-translation assays. Our results are consistent with our earlier work showing that co-transcriptional folding is essential even when ligand is present in excess (163).

6.3.8. Macromolecular crowding controls riboswitch activation

An intriguing and counter-intuitive model prediction is that higher riboswitch mRNA levels will actually decrease riboswitch activation ratios inside cells, but not within a relatively dilute cell-free assay. Inside cells, high concentrations of native RNA and protein occupy a significant portion of the intracellular volume, called macromolecular crowding (180, 181). As a result of volume exclusion, there is significantly less free volume available to a bulky mRNA. According to the model, when more mRNA is expressed inside the cell, the amount of free volume will further decrease, and could eventually lead to the exclusion of the ligand from the
same volumetric space as the mRNA with a correspondingly lower amount of mRNA-ligand complex. We calculate the free mRNA, free ligand, and ligand-bound mRNA volume fractions by minimizing the thermodynamic free energy of the constant volume system, constrained by the total mRNA and ligand concentrations and their molar volumes (supplementary notes in Appendix B.2.2). In Figure 6-5A, we show how these volume fractions vary as more riboswitch mRNA is expressed. Using these volume fractions and Equation 6.2, the model qualitatively predicts that expressing more riboswitch mRNA will lead to higher translation rates in both the ON and OFF states (higher $r_{ON}$ and $r_{OFF}$), however the riboswitch's activation ratio will actually decrease because $r_{ON}$ increases less than $r_{OFF}$. In contrast, in a dilute system, there is enough free volume that adding more riboswitch mRNA should not exclude ligand from the system. Riboswitch activation will increase if ligand-binding follows second order kinetics (supplementary notes in Appendix B.2.7).

We tested these predictions by adding increasing amounts of riboswitch DNA template to cell-free transcription-translation assay, and by increasing riboswitch mRNA levels inside E. coli cells using modified promoters and plasmid copy numbers (Figure B-17). We found that increasing the mRNA levels of the TMR-10 riboswitch by about 1000-fold in cell-free assays increased its reporter expression levels, while also enhancing its activation ratio from 3.9- to 14.8-fold (Figure 6-5B). However, when we increased the in vivo mRNA levels for the same TMR riboswitch, by replacing the single-copy BAC origin with either low-copy or high-copy origins of replication (p15A, ColE1, or pUC19), the measured activation ratio substantially dropped from 12.3- to 5.7-fold (Figure 6-5C and Figure B-17A). The same differences in cell-free versus in vivo activation were observed for three different TMR riboswitches. We then verified that changing mRNA levels, and not plasmid burden, were responsible for modulating riboswitch activation by introducing three different promoters to vary the transcription rates of 3 fluoride and 2 theophylline riboswitches. A 50-fold increase in promoter transcription rate
(Figure B-17B) dropped the F-11 riboswitch’s activation ratio from 4.0- to 1.6-fold, and the Theo-44 riboswitch’s activation ratio from 12.6- to 6.3-fold (Figure 6-5E). The same trend was observed for all five riboswitches without finding a reduction in the hosts’ growth rates, compared to wild-type or no-aptamer controls (Figure B-18A). We also found that the reporter expression levels from the no-aptamer controls always exceeded the riboswitches' expression levels, indicating that the host's maximum translational capacity had not been reached even as mRNA levels were increased (Figure B-18B). These results are qualitatively consistent with our model-calculated activation ratios (Figure 6-5D) and the expectation that volume exclusion will have a more pronounced effect on larger ligands (theophylline, TMR), compared to tiny fluoride. Consequently, as a general design principle, lowering a riboswitch's mRNA level will increase its activation ratio inside cells.
Figure 6-5. Volume exclusion controls riboswitch function inside cells. (A) The relationship between riboswitch mRNA level and the volume fractions (v) of free mRNA, free ligand, mRNA-ligand complex, and free volume in either dilute or crowded systems were calculated according to Equation 6.2 for the TMR-10 riboswitch. (B) Cell-free transcription-translation assays were performed for three TMR riboswitches using increasing DNA template concentrations (0.1, 1, 10, and 100 ng/µl). Colors, symbols, and ligand concentrations are the same as Figure 6-4. (C) The TMR riboswitches’ in vivo expression levels were measured at zero and 20 µM TMR after increasing the plasmid copy numbers, using the pBAC, p15A, pColE1, and pUC19 replication origins. (D) Model-predicted AR_{actual} for the TMR-10 riboswitch when varying the total concentrations of mRNA and ligand. The effect of changing the plasmid’s copy number on mRNA level is shown. (E) The in vivo expression levels and activation ratios for theophylline and fluoride riboswitches were measured after increasing riboswitch mRNA levels, using three promoters with steadily
increasing transcription rates. Colors, symbols, and ligand concentrations are the same as Figure 6-4. Each data point and bar represents the mean and standard deviation of 2 to 4 measurements.

6.3.9. Model-predicted limits of riboswitch-based sensors

Finally, we illustrate the potential limits of riboswitch-based sensors by designing an idealized “perfect” riboswitch that uses a hypothetical aptamer and highly optimized pre, post, and coding sequences for maximum possible detection ($AR_{\text{max}} = 16000$, $\Delta G_{\text{mRNA}} = 15.2$ kcal/mol) (Figure 6-6A, and supplementary notes in Appendix B.2.8). When a very high-affinity aptamer is available ($K_D \sim 10$ pM) and its ligand is plentiful, such a riboswitch could activate with a high dynamic range ($AR_{\text{actual}} > 1000$-fold). However, when faced with the most stringent sensing application, for example, sensing nanomolar ligand levels, even the most ideal riboswitch will only activate expression by 20-fold using a high-affinity aptamer, and 3-fold using a low-affinity aptamer (Figure 6-6B). These limits provide an approximate best-case for engineering novel riboswitches, regardless of aptamer sequence and structure, and help determine when downstream signal amplification will be necessary.

6.4. Discussion and conclusion

We have developed a parsimonious statistical thermodynamic model that predicts the sequence-structure-function relationship for translation-regulating riboswitches, accounting for the interactions between the ribosome, mRNA, and ligand. We experimentally validated the model by designing and characterizing 62 synthetic riboswitches, using 6 different RNA aptamers, that bind to their respective ligands (theophylline, tetramethylrosamine, fluoride, dopamine, thyroxine, or 2,4-dinitrotoluene) and activated translation rate by up to 383-fold. We
used both cell-free \textit{in vitro} transcription translation assays as well as \textit{in vivo} bacterial expression assays to measure riboswitch function. Importantly, we also performed extensive measurements of no-aptamer controls to ensure that the non-specific effects of the ligands on gene expression levels are correctly subtracted. For 55\% of these riboswitches, the model was able to correctly calculate the ligand-induced changes in the ribosome-mRNA binding free energy to within 1.5 kcal/mol.

Figure 6-6. A “perfect” riboswitch to test the limits of sensing. (A) Optimization was used to design a “perfect” riboswitch using a hypothetical aptamer, which is shown in its ligand-free and ligand-bound states. (B) Model calculations show the effects of the aptamer’s binding free energy (affinity) and maximum ligand concentration on the “perfect” riboswitch’s actual activation ratio to illustrate the best possible expression changes under potential sensing scenarios.

Using our thermodynamic model, we experimentally confirmed several mechanisms that control riboswitch function. First, some RNA aptamers are more conducive to riboswitch engineering than others, though for previously unknown reasons. We demonstrated that an aptamer’s ligand binding free energy places an upper limit on the possible RNA conformational
changes that may occur, according to its switching free energy, which then places a maximum limit on the riboswitch’s activation ratio. The same thermodynamic limit will govern the ligand-induced conformational changes of all RNA regulators. Second, the effect of volume exclusion on biomolecular interactions inside crowded cells has remained poorly understood, though it is particularly important when expressing bulky mRNAs. A key model prediction is that higher mRNA concentrations will lower riboswitch activation ratios by reducing the amount of free volume. We systematically increased intracellular mRNA concentrations to show that, indeed, riboswitch activation ratios drop by considerable amounts. This effect was not observed in several in vivo no-aptamer controls and was found to depend on the ligand’s molar volume, as expected according to the theory. Importantly, we repeated these measurements in the 20-fold more dilute cell-free assay and expectedly found the opposite behavior. The principles that we experimentally validated to develop this model are not specific to riboswitches, and encompass RNA molecules whose ligand-induced conformational changes affect their interactions with proteins, other RNAs, or supramolecular complexes (e.g. the ribosome).

While all models are imperfect, here we systematically quantified the thermodynamic model’s predictive accuracy across a large and diverse data-set, enabling us to distinguish the differences between stable and meta-stable riboswitches and further improve our understanding of the physics of RNA shape-change. Importantly, throughout this study, our model assumes thermodynamic equilibrium between the ribosome, mRNA, and ligand, and does not attempt to predict changes in mRNA stability. We envision that eliminating these assumptions will further improve the model’s ability to predict riboswitch function from sequence.

Finally, by automating our physical chemistry calculations, both experts and non-experts alike can perform identical calculations to convert a wide array of RNA aptamers into engineered riboswitches, broadly enabling the rapid prototyping of new biomolecular sensors for use in medical diagnostics, environmental remediation, metabolic engineering, and future fields (43).
web-interface to the automated design method, called the Riboswitch Calculator (see supplementary notes in Appendix B.2.3), is accessible at http://salislab.net/software.
Chapter 7

Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism

All the measurements and model predictions in this work are presented in Supplementary Data 7 and can be found online at:


7.1. Introduction

Inside cells, the self-assembly kinetics of DNA, RNA, and protein controls the rates of its cellular processes, such as transcription and translation. It has been a long-standing question to elucidate the genetic code’s sequence-structure-function relationship, particularly to determine how changes in non-equilibrium self-assembly contribute to biological function. Much of our understanding of biological self-assembly arises from cell-free, in vitro experiments where the system eventually reaches thermodynamic equilibrium. However, most cellular processes undergo continuous cycles, which operate under non-equilibrium conditions (182–184).

In this chapter, we show how a mRNA’s nucleotide sequence affects the rate of its non-equilibrium assembly with the ribosome, directly controlling the rate of translation initiation, which is the first step in protein synthesis. We introduce a new mechanism, called Ribosome Drafting, that explains how changes in RNA folding kinetics can control protein synthesis rates by pushing the system into non-equilibrium states. Our results demonstrate a rational design
approach for studying self-assembly inside cells and provide the theoretical foundation for building predictive non-equilibrium sequence-to-function models of translation.

7.2. Materials and methods

7.2.1. Plasmids and strains

The expression plasmids are derivatives of pFTV1 vector (21) with ColE1 origin of replication, Chloramphenicol (Cm) resistance gene, and mRFP1 fluorescent protein reporter. All mRNAs are transcribed using constitutive J23100 promoter. The designed inserts that replace the mRNA’s 5’-UTR or 5’-CDS regions were prepared using either annealing of oligonucleotides with proper overhangs (Integrated DNA Technologies), or PCR assembly of overlapped oligonucleotides, followed by digestion with XbaI/Ndel or Ndel/SacI enzymes, respectively. In particular, NdeI restriction site contained the start codon AUG. The inserts were then ligated to the backbones digested with corresponding restriction enzymes, followed by transformation of electrochemically competent Escherichia coli DH10B cells using electroporation, and plating on selective media. Clones with correct sequences were cryo-stored for further characterizations.

7.2.2. Growth and characterization

Cells containing the cloned plasmids were characterized using long-run “serial dilution” experiment (12) to measure the protein expression levels at steady state conditions. Briefly, two clones per each construct were grown overnight in 700 ul of LB miller media (VWR) with 50 ug/ml Cm antibiotic (Sigma-Aldrich) in a 96-well deep well plate at 37 °C and 200 rpm orbital shaking. Next, 10 ul of cultures were diluted into 200 ul of fresh LB/Cm media and grown in a
96-well microplate at 37 °C and with proper shaking. The OD<sub>600</sub> and fluorescence levels were recorded every 10 min using spectrophotometer TECAN M1000. At mid-exponential growth phase (OD<sub>600</sub> = 0.15) the protein expression levels reached steady state. To keep the cultures at the exponential growth and maintain them at steady state conditions, cultures were diluted by transferring 10 ul of culture into 200 ul of fresh LB/Cm media in the second 96-well microplate and grown until the corrected OD<sub>600</sub> reached 0.15. Finally, the same dilution cycle was repeated one more time by transferring cells into third 96-well microplate with fresh media, and growing them to mid-exponential growth (OD<sub>600</sub> = 0.15). This experiment was repeated in two separate days for all the constructs (n = 4).

To measure the single-cell protein expression levels, 10 ul of cultures were collected at the end of the second and third growth cycles (steady state conditions), and transferred to 200 ul of 1X PBS solution (12). The fluorescence levels per single cell were measured using BD Fortessa flow cytometer with 100,000 events threshold. All the fluorescence distributions were unimodal (see Figure 7-4D for example). The average fluorescence levels were corrected for the background E. coli DH10B fluorescence. All the reported fluorescence values are the average of 2 measurements in two separate days (n=4), and are listed in Supplementary_Data_7.

7.2.3. Design of the expression system

To demonstrate the significance of RNA folding kinetics on translation initiation rate, we first designed a control mRNA harboring an mRFP1 fluorescent protein reporter. Similar to Chapter 5, the control mRNA was designed to have no secondary structure around its translation initiation site by using an AC-rich standby site sequence upstream of the SD region (5’-TCTAGAAACCGCCATATACGCGGGACACACA-3’), and fusing a 39 nt AC-rich insulator sequence to the beginning of mRFP1 coding section (all non-rare codons) (5’-
AAACAGAACAAAGAACAGATCACCACACAGACATCAAC-3\textsuperscript{\textprime}) (Figure 7-4B). The designed standby site contained a stable hairpin (underlined above) that together with the insulator sequence prevented the formation of any undesired long range base-pairings across the start codon. In addition, the standby site had a sufficient single-stranded surface area ($A_S = 20$ nt), which provided a proper binding site for 30S ribosome platform (12). Finally, the mRNA’s SD sequence (5’-CAAGGAGAC-3’) was designed to strongly bind the 30S ribosome’s antiSD (5’-ACCUCUUUA-3’). Overall, our thermodynamic model calculations using RBS Calculator v2.0 (12) predicted that the no-hairpin control mRNA binds tightly to 30S ribosome ($\Delta G_{\text{complex}} = -8.3$ kcal/mol), resulting in a very high predicted translation initiation rate (104,720 au). The experimental data corresponding to this no-hairpin control mRNA is shown in Figures 7-4D and 7-11A.

We further designed another no-hairpin control mRNA with a lower predicted translation initiation rate (1744 au). This no-hairpin control mRNA was similar to the previous one, except that its SD sequence was mutated (5’-CAACGAGGC-3’) and the spacing length was reduced to 2 nt, resulting in a more positive (weaker) mRNA-ribosome complex free energy ($\Delta G_{\text{complex}} = 0.8$ kcal/mol). The experimental data corresponding to this no-hairpin control mRNA is depicted in Figures 7-11C.

Finally, to show the effect of 30S binding rate on the ribosome drafting mechanism in a systematic way, we designed a library of six no-hairpin control mRNAs that were identical in their sequences except for their SD sequence (Figure 7-6A). They all contained the previous 39 nt insulator sequence fused to mRFP1 coding section, and their standby site was redesigned to support the no-hairpin requirement around the start codon (5’-TCTAGACCCGCAGCCATATACGGCAGCGGGAAAA- 3’) (single stranded surface area $A_S =$
13 nt). All six SD sequences were carefully designed using the RBS Calculator v2.0 (12) to vary the ribosome binding rate across 1000-fold range, as depicted in **Figures 7-6B**.

After constructing different no-hairpin control mRNAs, we designed their corresponding structured mRNAs by inserting exclusive RNA hairpins into their translation initiation region such that the RNA hairpin sequester the SD and start codon (**Figures 7-4B** and **7-6A**). These designed RNA hairpins have diverse folding energies and folding time scales. Overall, these expression systems allowed us to examine the direct effects of hairpin folding kinetics on the mRNA’s translation initiation rate by comparing the fluorescence level of structured mRNAs (both fast-folding and slow-folding hairpins) with the fluorescence level of no-hairpin control mRNA.

### 7.2.4. Predicting the RNA hairpin folding time using Kinfold calculations

We employ the Kinfold program from ViennaRNA package (145) to compute the RNA hairpin folding kinetics. Kinfold is a nucleotide-based stochastic simulation that predicts the folding trajectory of a given RNA sequence by computing the number of intermediate folding and unfolding events that occur during transition from unfolded to fully folded RNA. Due to the heuristic nature of Kinfold simulation, the computed folding times are not unique. Therefore, by repeating the Kinfold simulation for 1000 times we obtain a distribution of folding times for a given RNA hairpin sequence (**Figure 7-4C**). Interestingly, we found that each RNA hairpin has a distinguished folding trajectory, and two RNA hairpins with similar folding energies can fold at very different rates (**Figure 7-4C**). This introduces a new layer of sequence-dependent information that needs to be considered when designing RNA devices for complex genetic circuits.
7.2.5. Biophysical model of translation initiation

We have previously developed a biophysical model for translation initiation process (21), that predicts the mRNA’s translation initiation rate (TIR) directly from its sequences, according to Equation 7.1.

\[
\text{TIR} = K \cdot \exp(-\beta \Delta G_{\text{total}})
\]  

(7.1)

This equation is derived from the statistical thermodynamic principles, and relates the initiation rate to the total Gibbs free energy change from binding of ribosome to mRNA, \(\Delta G_{\text{total}}\), which is calculated from Equation 7.2.

\[
\Delta G_{\text{total}} = \Delta G_{\text{mRNA-rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} + \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}}
\]  

(7.2)

In Equation 7.1, \(\beta\) is the Boltzmann constant and \(K\) is a proportionality constant (21), and the five energetic terms in Equation 7.2 are the energetic interactions between mRNA and ribosome during initiation, which have been extensively explained in Chapters 2 and 5 (see Materials and methods section in Chapter 5). Here, in this chapter, our focus is to explain the translation initiation process from a kinetic point of view. According to our proposed “Ribosome Drafting” mechanism, after the 30S initiation complex is formed and the initiation step is completed, the elongating 70S complex departs for elongation, which leaves the initiation site unfolded. Therefore, the initiation site can either start to refold such that the whole mRNA reaches its MFE state, or quickly bind to a new 30S ribosome to form the next initiation complex. Each of these two processes has specific sequence-dependent rate, allowing the RNA folding kinetics to regulate the overall translation initiation rate. Therefore, for the purpose of Ribosome Drafting mechanism, we rewrite Equation 7.2 according to Equation 7.3.

\[
\Delta G_{\text{total}} = \Delta G_{\text{complex}} - \Delta G_{\text{redfold}}
\]  

(7.3)
\( \Delta G_{\text{complex}} \) is the energy released from the binding of 30S ribosome to an *unfolded* mRNA (no unfolding step is needed for 30S ribosome), and anchoring mRNA to the ribosomal P-site. A successful mRNA-ribosome complex formation depends on four energetic interactions: the SD-antiSD hybridization, the hybridization of start codon and tRNA\(^{\text{Met}}\) anti-codon loop, the nucleotide distance between SD and start codon, and the standby site accessibility. Therefore, the Gibbs free energy of mRNA-ribosome complex can be calculated according to Equation 7.4.

\[
\Delta G_{\text{complex}} = \Delta G_{\text{SD-antiSD}} + \Delta G_{\text{spacing}} + \Delta G_{\text{standby}} + \Delta G_{\text{start}}
\]  

(7.4)

On the other hand, \( \Delta G_{\text{refold}} \) is the energy released when an unfolded mRNA (mRNA with unfolded initiation site) refolds to its fully folded state, which is calculated according to Equation 7.5. In this process, the local structures folded within the pre-SD and post-footprint regions should unfold to allow the entire mRNA sequence to refold to its MFE state, hence the negative signs for \( \Delta G_{\text{pre-SD}} \) and \( \Delta G_{\text{post-footprint}} \) in Equation 7.5.

\[
\Delta G_{\text{refold}} = \Delta G_{\text{mRNA}} - \Delta G_{\text{pre-SD}} - \Delta G_{\text{post-footprint}}
\]  

(7.5)

The \( \Delta G_{\text{refold}} \) is the maximum energy penalty that 30S ribosome has to expend to unfold mRNA. It happens only when a sufficient time is given to mRNA to refold to its MFE state (thermodynamic equilibrium assumption). However, when competing with the 30S ribosome binding, the RNA folding kinetics determines the extent of non-equilibrium mRNA folding. Interestingly, as we show in the Results section, when the RNA folding kinetics is slow, plus RNA is weakly stable and 30S binding is fast, the ribosome drafting mechanism does not allow the mRNA to refold at all. This complicated relationship between RNA folding kinetics and 30S ribosome binding provides valuable insight on the events that can happen during translation initiation of mRNAs with long and complex structures.

Finally, we should note that in this study, we designed 22 different inhibitory RNA hairpins that sequestered SD and start codon with diverse folding energy and folding kinetics.
Our design goal was to avoid any undesired hairpin formation in the post-footprint region. However, due to the complexity of DNA sequence design, 7 out of 36 designed mRNAs have weak hairpin in their post-footprint region, from which only 3 has hairpins with folding energy greater than 1 kcal/mol. Nonetheless, we considered these extra energy terms into our data analysis according to Equation 7.6.

\[ \Delta G_{\text{fold,actual}} = \Delta G_{\text{fold}} - \Delta G_{\text{post-footprint}} \]  

(7.6)

7.2.6. A Markov model for Ribosome Drafting mechanism

We modeled the Ribosome Drafting mechanism using a simplified stochastic Markov model with four states (Figure 7-1B). Below, the state’s numberings are the same as those in Figure 7-1B. The probability of each state at steady-state (Pi) is calculated by analytically solving the following equations.

\[ (P_3 + P_4) a_{\text{elong}} - P_1 a_1 - P_1 a_{\text{fold}} = 0 \]

\[ P_1 a_{\text{fold}} - P_2 a_2 = 0 \]

\[ P_1 a_1 - P_3 a_{\text{elong}} = 0 \]

\[ P_2 a_2 - P_4 a_{\text{elong}} = 0 \]

The \( a_{\text{elong}} \) is the rate at which the elongating 70S complex clears the initiation site. In this study, by selecting the fast codons for both the designed hairpins and the insulator sequence at the N-terminal of mRFP1 coding section, we made sure that the elongation rate (\( a_{\text{elong}} \)) is not rate-limiting. Consequently, in the Markov model calculations, we assumed the \( a_{\text{elong}} \) to be 1000-times faster than \( a_1 \).
\( a_{\text{fold}} \) is the RNA hairpin folding rate, \( a_1 \) is the 30S binding rate when the initiation site is unfolded (Equation 7.7), and \( a_2 \) is the 30S binding rate when the mRNA is fully folded (Equation 7.8). The 30S binding rates are calculated using a statistical thermodynamic model of mRNA-ribosome interactions (21), as follows:

\[
a_1 = A \exp(-\beta \Delta G_{\text{complex}}) \\
(7.7) \\
a_2 = A \exp(-\beta (\Delta G_{\text{complex}} - \Delta G_{\text{refold}})) \\
(7.8)
\]

\( \beta \) is the Boltzmann constant (0.45 mol/kcal (21)), and \( A \) is a proportional constant which is set to 100 in our Markov model calculations. \( \Delta G_{\text{complex}} \) is the energy released from forming mRNA-ribosome complex (Equation 7.4), and \( \Delta G_{\text{refold}} \) is the mRNA’s refolding energy from unfolded state to fully folded state (Equation 7.5).

Finally, in the Markov model, translation initiation rate is calculated according to Equation 7.9. For the sake of comparison, the translation initiation rate under thermodynamic assumption is calculated using the same formula but with \( a_{\text{fold}} \) much greater than \( a_1 \) (\( a_{\text{fold}} \rightarrow \infty \)).

\[
\text{TIR} = (P_3 + P_4)a_{\text{elong}} \\
(7.9)
\]

7.3. Results

7.3.1. A Ribosome Drafting mechanism for controlling translation initiation

Translation initiation is a rate-limiting step in gene expression, whereby the 30S ribosomal subunit binds to a mRNA to form a 30S initiation complex (30SIC) (22, 31), followed by 50S recruitment, 70S assembly, and peptide elongation. The probability that a 30S ribosome binds to a mRNA, and its corresponding translation initiation rate \( r \), are determined by the total Gibbs free energy change whenever a 30S ribosome binds to the 5’ untranslated region of a
mRNA, according to $r \propto \exp(-\beta \Delta G_{\text{total}})$ (Figure 7-1A, left) (12, 21, 31). $\Delta G_{\text{total}}$ is calculated using a sequence-dependent free energy model that quantifies the strengths of several mRNA and mRNA-ribosome interactions that affect a mRNA’s translation initiation rate (1, 4, 8, 11, 12, 14, 21, 24, 28, 30–32, 185) (see also Materials and methods section). In particular, the 30S ribosome must unfold any mRNA structures that overlap with its footprint; the free energy to unfold these structures penalizes the ribosome’s binding free energy. Once translation initiates and protein synthesis begins, another 30S ribosome binds to the mRNA to continue the cycle. Several cycles of translation typically take place per mRNA.

![Figure 7-1A](image)

**Figure 7-1.** A Ribosome Drafting mechanism for controlling translation initiation. (A) The first ribosome binds to folded mRNAs with near-equilibrium states, but successive ribosomes bind to pre-folded
mRNAs maintained in non-equilibrium states. (B) A Markov model shows how RNA hairpin refolding kinetics partitions the system into two states with distinct translation rates. (C) We show how the ribosome’s binding rate \( a_1 \) and hairpin refolding rates \( a_{\text{fold}} \) control the translation rate according to either (dark blue) a non-equilibrium Markov model utilizing Ribosome Drafting or (light blue) a previously developed thermodynamic equilibrium model. Here, \( \Delta G_{\text{refold}} \) is -15 kcal/mol, causing \( a_1 \) to be 1000-fold higher than \( a_2 \). \( a_{\text{elong}} \) is 1000-fold faster than \( a_1 \).

This model has been critically tested across 496 characterized mRNA sequences in diverse bacterial species (12, 21, 28, 144) with translation rates that varied by 100,000-fold. However, the model assumes that the pool of ribosomes and mRNAs have reached thermodynamic equilibrium. While the model’s average error is only 1.89 kcal/mol, there is a large subclass of characterized mRNA sequences that appear to fold slowly and therefore invalidate the equilibrium assumption (Figure 7-2). In particular, a long, slow-folding hairpin caused a 10-fold increase in translation rate beyond the thermodynamic prediction (Figure 7-3). While the potential influence of RNA folding kinetics on translation was considered 20 years ago (32, 33), recent studies have largely ignored its effects, and a quantitative mechanism has yet to be tested.

We propose that Ribosome Drafting takes place anytime when the 30S ribosome binds to a mRNA faster than the folding kinetics of its inhibitory RNA structures. While the first ribosome that initiates translation must unfold inhibitory RNA structures, the second and thereafter ribosomes can bind to the mRNA at pre-equilibrium states if the RNA structures fold slowly and if the ribosomes bind quickly (Figure 7-1A). Consequently, as ribosomes bind and translation initiates in continuous cycles, the mRNA never reaches an equilibrium folded state. Importantly, ribosomes that bind the mRNA in pre-equilibrium states will have a higher rate of translation initiation, and therefore the presence of Ribosome Drafting will accelerate protein synthesis rates.
Figure 7-2. A large sub-class of long and slow-folding mRNAs shows higher than predicted translation rates. (A) From 496 previously designed mRNA sequences (12, 21, 28, 144) we analyzed 485 mRNAs that expressed three reporter proteins (mRFP1, sfGFP, and luciferase) and were characterized within 6 different bacterial species. The total Gibbs free energy changes ($\Delta G_{\text{total}}$) of all mRNA sequences were calculated using the free-energy model of RBS Calculator v2.0 (12), and were related to the translation initiation rate according the formula: $TIR = K e^{(-\beta \Delta G_{\text{total}})}$ (see Materials and methods section). $\beta$ is 0.45 mol/kcal, as empirically measured in Ref (21). $K$ is the proportionality constant, which can vary under different transcription rates, reporter proteins, bacterial species, etc. Therefore, for each study, we obtained a specific $K$ value by fitting the predicted TIR values of only those mRNAs that have short (less than 35 nt) inhibitory RNA hairpins sequestering SD and start codon, or have slow-binding 30S ribosome ($\Delta G_{\text{complex}} > -7$ kcal/mol). Under these conditions, the effect of non-equilibrium Ribosome Drafting is assumed to be minimum, and therefore mRNAs are controlled mainly by thermodynamics of mRNA-ribosome binding. Finally, the fitted $K$ values were then applied to predict the TIR values for the rest of mRNA sequences. Overall, the model accurately predicted the translation initiation rate of this diverse mRNA data set ($R^2 = 0.75$, $p = 10^{-30}$, $N = 485$), with 2.4 kcal/mol average error for $\Delta G_{\text{total}}$. However, out of 485 mRNA sequences, 133 mRNAs were identified (red circles) to have long and slow-folding inhibitory hairpins, with a fast-binding ribosome, suggesting that they are influenced by Ribosome Drafting. In fact,
67% of these mRNAs have higher-than-predicted expression levels. (B) The error distribution indicates that non-equilibrium Ribosomal Drafting mechanism has a large impact on the model performance (28 mRNAs have more than 10-fold higher-than-predicted expression). The average error in $\Delta G_{\text{total}}$ for thermodynamically controlled mRNAs (blue region) was 2.0 kcal/mol, whereas the non-equilibrium driven mRNAs (red region) caused a larger error in predicted $\Delta G_{\text{total}}$ (3.4 kcal/mol). All mRNA sequences, with their references, and the model calculations are presented in Supplementary_Data_7.

Figure 7-3. A slow-folding mRNA is largely under-predicted by the thermodynamic model. (A) The in vivo measured fluorescence level of three mRNAs from Salis et al. 2009 (21) (I: 4-lacI(27nt)-mRFP1, II: 2-lacI(27nt)-mRFP1, III: 5-mRFP1) are shown against their predicted $\Delta G_{\text{total}}$ using RBS Calculator v2.0 (12). The blue dotted line is the model prediction based on the formula: $TIR = 3820 \, e^{(-0.45 \, \Delta G_{\text{total}})}$. The mRNA-II shows 10-fold higher expression level than the model prediction. (B) The RNA hairpins that sequester the SD and start codon of each mRNA (I, II, and III) are shown. Orange and blue colors are SD and start codon, respectively. Folding time distribution is from 1000 folding trajectories calculated using the Kinfold program (145). Consistently, the RNA hairpin-II folds much slower than other two RNA hairpins with well-predicted expression levels.
To incorporate Ribosome Drafting into a non-equilibrium model of translation, we developed a four-state Markov model (Figure 7-1B, Materials and methods section) where the 5’ untranslated region of the mRNA is either free or bound by a 30S ribosome, and the 5’ untranslated region is either unfolded or contains a fully folded translation-inhibiting structure. Given a mRNA sequence, we calculated the energy needed to refold the translation-inhibiting RNA structure (\(\Delta G_{\text{refold}} < 0\)) using a nearest-neighbor free energy model and folding algorithm (52–54). We also simulated 1000 folding trajectories to determine the structure’s folding time distribution (145), its exponentially distributed folding time \(\tau\), and its folding rate \(a_{\text{fold}}\). We then calculated the binding free energy of the 30S ribosome to the fully unfolded mRNA (\(\Delta G_{\text{complex}}\)), which strongly depends on the hybridization energy between the 16S ribosomal RNA and the Shine-Dalgarno sequence within the 5’ untranslated region. We used these calculations to determine the Markov model’s transition rates.

When the 5’ untranslated region is fully unfolded, the ribosome’s binding rate is \(a_1 \propto \exp(-\beta \Delta G_{\text{complex}})\). However, when it has a fully folded translation-inhibiting hairpin, the ribosome’s binding rate is slower, according to \(a_1 = a_2 \exp(-\beta \Delta G_{\text{refold}})\). According to the steady-state solution of this Markov model, whenever the 30S ribosome binds faster than the RNA folding time, the non-equilibrium Ribosome Drafting mechanism can substantially increase the mRNA’s translation rate, compared to a thermodynamic equilibrium model (e.g. up to 1000-fold, Figure 7-1C).
7.3.2. Validating the Ribosome Drafting mechanism

We validated the Ribosome Drafting mechanism by measuring mRNA translation rates using fluorescent protein reporters under steady-state conditions, while incorporating computationally designed, translation-inhibiting RNA hairpins to systematically vary folding times ($a_{\text{fold}}$), ribosome binding free energies ($\Delta G_{\text{complex}}$), and RNA refolding free energies ($\Delta G_{\text{refold}}$). We designed hairpin sequences by combining RNA folding simulations with an optimization algorithm to identify RNA hairpins that have equivalent refolding free energies, but distinctly different folding times (Figure 7-4A). Further, to avoid confounding interactions, hairpins were only selected if they exhibited a single folding transition to satisfy the model’s 2-state assumption (Figure 7-5); if their coding sequence portions utilized fast codons to maximize translation elongation rates; and if they contained no more than eight consecutive base pairings to minimize RNase-mediated degradation.

Initially, we designed a pair of fast-folding and slow-folding RNA hairpins, introduced them into our mRFP1 fluorescent protein reporter expression system using identical, constitutive promoters (Figure 7-4B), and measured their single-cell fluorescence levels inside *E. coli* DH10B cells after a 12 hour culture maintained in the exponential growth phase (Materials and methods section). The simulated folding time distributions of these RNA hairpins had averages 5-fold apart (Figure 7-4C), though their calculated refolding free energies were within 0.3 kcal/mol. Both mRNAs were also designed to bind tightly to the ribosome ($\Delta G_{\text{complex}} = -8.3$ kcal/mol, see Materials and methods section), yielding a high ribosome binding rate $a_1$. The measured fluorescence levels showed that the mRNA with a slow-folding RNA hairpin had a 1103-fold higher protein expression level than the mRNA with a fast-folding hairpin (Figure 7-4D). This very large difference in protein expression cannot be explained by the thermodynamic equilibrium model because both mRNAs have similar $\Delta G_{\text{complex}}$ and $\Delta G_{\text{refold}}$, and similar total
ribosome binding free energies ($\Delta G_{\text{total}}$). We performed the same characterization on a mRNA that lacked such an inhibitory RNA hairpin (a no-hairpin control), and found that its reporter expression level was only 1.6-fold higher than the mRNA with a slow-folding RNA hairpin, suggesting that the ribosome did not need to exert much work to unfold the slow-folding RNA hairpin and initiate translation. These results provided initial evidence that the Ribosome Drafting mechanism can greatly affect a mRNA’s translation rate.

Figure 7-4. Demonstrating the Ribosome Drafting mechanism. (A) Computational design of RNA hairpins with desired folding time distributions. $\tau$ is the mean folding time from 1000 RNA folding trajectories calculated using Kinfold (145). (B) Expression systems used to measure the effect of RNA hairpins on translation rates, including designed insulators to prevent undesired structures. (C) Two designed RNA hairpins with similar folding energies ($\Delta G_{\text{refold,fast}} = -7.6$, $\Delta G_{\text{refold,slow}} = -7.3$ kcal/mol), but
disparate folding times ($\tau_{\text{fast}} = 105 \text{ au}$, $\tau_{\text{slow}} = 503 \text{ au}$). (D) The expression systems' steady-state, single-cell fluorescence levels were measured in *E. coli* DH10B cells. BG, background auto-fluorescence.

**Figure 7-5. Designed RNA hairpins follow a single folding transition.** The folding time distributions of two designed RNA hairpins (characterized in Figure 7-4) are fitted to an exponential distribution function with rate constant $= 1/\tau_{\text{mean}}$. $\tau_{\text{mean}}$ is the average folding time of 1000 folding trajectories calculated using Kinfold program. Fitted functions suggest that folding of these RNA hairpins follows a 2-state folding process with only an unfolded RNA and a fully folded RNA at equilibrium.

### 7.3.3. The effect of 30S binding rate on the Ribosome Drafting

We then investigated how the ribosome’s binding rate $a_1$ affected its ability to bind mRNA in a pre-equilibrium state. We systematically varied $a_1$ by introducing rationally designed Shine-Dalgarno sequences that alter the ribosome’s binding free energy ($\Delta G_{\text{complex}}$) from 4.7 to -
10.3 kcal/mol, and therefore vary $a_1$ by about 1000-fold (Figure 7-6A, see Materials and methods section). For each Shine-Dalgarno variant, fast-folding and slow-folding RNA hairpins were computationally designed, introduced into mRNAs together with no-hairpin control mRNAs, and their expression levels characterized as before. All calculated RNA hairpin structures are shown in Figure 7-7. All sequences, measurements, and calculations are included in the Supplementary_Data_7.

According to the data, the equilibrium model was able to accurately predict the translation initiation rates of the mRNAs without inhibitory hairpins ($R^2 = 0.94$, $p = 0.0012$, Figure 7-6B) and with fast-folding inhibitory RNA hairpins ($R^2 = 0.74$, $p = 0.027$, Figure 7-8), but the equilibrium model could not explain why the translation rates of the mRNAs with slow-folding hairpins increased sigmoidally (Figure 7-6C).

Instead, we found that the non-equilibrium Markov model, utilizing the Ribosome Drafting mechanism, could explain how both the ribosome’s binding rate and the RNA hairpin’s folding time controlled the mRNA’s translation rate (Figure 7-6D). As the ribosome’s binding rate was increased, there was less time for the slow-folding hairpins to refold, yielding a higher fraction of unfolded mRNA and a higher translation initiation rate. The model predicts the observed sigmoidal relationship where slow-binding ribosomes must unfold the fully folded hairpin, but fast-binding ribosomes employ Ribosome Drafting to bind pre-folded mRNAs. Our measurements also qualitatively agree with the model’s predictions at very slow or very fast ribosome binding rates. A hairpin’s folding kinetics will not affect a mRNA’s translation rate if the ribosome binds very slowly. When the ribosome binds very fast, the translation rate of a slow-folding mRNAs will be similar to a mRNA without any hairpin (the no-hairpin control).
Figure 7-6. The extent of Ribosome Drafting is controlled by the 30S binding rate. (A) Design of mRNAs with varying SD sequences and 30S binding rates. Thermodynamically equivalent fast-folding and slow-folding hairpins were designed to inhibit translation. (B) The no-hairpin expression system's measured fluorescence levels were well-predicted by the thermodynamic model. (C) Measured fluorescence levels and (D) Markov model solutions with respect to 30S binding rates. $\Delta G_{\text{refold}} = -8.0$ kcal/mol. $a_{\text{fold, fast}} = 2000$ au. $a_{\text{fold, slow}} = 20$ au. $a_{\text{elong}}$ is 1000-fold faster than $a_1$. Data points and error bars are the average and s.d. of four independent measurements on two separate days.
Figure 7-7. Sequence and structure of designed RNA hairpins to test the effect of 30S binding rate.

The SD numbers are the same as those in Figure 7-6A. The SD and start codons are shown in orange and blue colors, respectively.
Figure 7-8. Thermodynamic model predictions for mRNAs with fast-folding RNA hairpins. Thermodynamic model predictions are proportional with the measured fluorescence levels of mRNAs with fast-folding hairpins ($R^2 = 0.74, p = 0.027$). The fluorescence levels did not increase sigmoidally, as seen for mRNAs with slow-folding hairpins (Figure 7-6C), indicating that thermodynamic model is accurate for fast-folding mRNAs.

7.3.4. The effect of RNA hairpin’s folding times and folding free energies on the Ribosome Drafting

According to the Markov model, the extent and impact of the Ribosome Drafting mechanism are controlled by two factors. First, the time-scale separation between ribosome binding and RNA folding determines the extent of its non-equilibrium behavior. The probability of the ribosome being bound to the mRNA’s pre-folded state ($P_3$) or equilibrium state ($P_4$) is related according to $P_3/P_4 = a_1 / a_{\text{fold}}$, indicating simply that when the kinetics of RNA folding are 10-fold slower than the ribosome’s binding rate ($a_1$), then 90% of the mRNAs will be maintained in a non-equilibrium, pre-folded state, yielding a higher translation initiation rate. Second, when
Ribosome Drafting takes place, its acceleration of the translation initiation rate depends on the inhibitory RNA hairpin’s folding free energy via $a_1 = a_2 \exp(-\beta \Delta G_{\text{refold}})$. These effects are independent because there is no strict relationship between an RNA hairpin’s GC content, its length, its folding free energy, or its folding time. Some weakly stable RNA structures fold quickly, while other stable RNA structures have many intermediate states and therefore fold slowly (186, 187).

We demonstrated how these factors control Ribosome Drafting by computationally designing fast-folding and slow-folding inhibitory hairpins with varying refolding free energies ($\Delta G_{\text{refold}}$ from -11.3 to -4.7 kcal/mol) and inserting them into mRNAs that bound ribosomes with either fast or slow rates ($\Delta G_{\text{complex}} =$ -8.3 or 0.8 kcal/mol, respectively) (Figures 7-9 and 7-10). The translation rate measurements illustrated that the Ribosome Drafting mechanism had a very large impact (a 102-fold translation rate change) when ribosomes bound rapidly to a mRNA with a slow-folding, highly stable hairpin ($\Delta G_{\text{refold}} = -$11.3 kcal/mol), compared to its equivalent fast-folding hairpin (Figure 7-11A). In contrast, when the same measurements were performed on the least stable hairpin ($\Delta G_{\text{refold}} = -$4.7 kcal/mol), the impact on translation rate was much lower (a 7-fold change), which agrees well with the model calculations (Figure 7-11B). In both cases, the large time-scale separation between ribosome binding and RNA folding prevented the system from reaching equilibrium, and the impact of the Ribosome Drafting mechanism was controlled by the hairpin’s refolding free energy. However, when ribosomes bound very slowly to the mRNAs, their translation rates were mainly affected by the hairpin’s refolding free energies, and not their folding times (Figure 7-11C). A large change in folding time had a small effect on measured translation rates (between 4- and 16-fold). Here, the system operated closer to equilibrium conditions, where both the non-equilibrium Markov model and the thermodynamic equilibrium model have similar predictions (Figure 7-11D and Figure 7-12).
Figure 7-9. Sequence and structure of designed RNA hairpins to test the effect of folding free energy (the case of fast-binding ribosome). $\Delta G_{\text{complex}}$ for these RNA hairpins is -8.3 kcal/mol. Experimental data corresponding to these hairpins are shown in Figure 7-11A. The SD and start codons are shown in orange and blue colors, respectively.

$\Delta G_{\text{refold,I}} = -4.7$ kcal/mol, $\tau_1 = 120$ au
$\Delta G_{\text{refold,II}} = -4.8$ kcal/mol, $\tau_2 = 350$ au

$\Delta G_{\text{refold,I}} = -7.6$ kcal/mol, $\tau_1 = 105$ au
$\Delta G_{\text{refold,II}} = -7.3$ kcal/mol, $\tau_2 = 503$ au

$\Delta G_{\text{refold,I}} = -8.7$ kcal/mol, $\tau_1 = 60$ au
$\Delta G_{\text{refold,II}} = -9.1$ kcal/mol, $\tau_2 = 740$ au

$\Delta G_{\text{refold,I}} = -11.3$ kcal/mol, $\tau_1 = 50$ au
$\Delta G_{\text{refold,II}} = -11.2$ kcal/mol, $\tau_2 = 1075$ au
Figure 7-10. Sequence and structure of designed RNA hairpins to test the effect of folding free energy (the case of slow-binding ribosome). $\Delta G_{\text{complex}}$ for these RNA hairpins is 0.8 kcal/mol. Experimental data corresponding to these hairpins are shown in Figure 7-11C. The SD and start codons are shown in orange and blue colors, respectively.
Figure 7-11. Ribosome Drafting’s impact on translation rate is controlled by the RNA hairpins’ folding times and folding free energies. (A) Measured fluorescence levels from gene expression systems containing inhibitory RNA hairpins with fast-binding ribosomes ($\Delta G_{\text{complex}} = -8.3$ kcal/mol) and varying refolding energies ($\Delta G_{\text{refold}}$). (B) The Markov model solution with $a_1 = 4190$ au and varying $\Delta G_{\text{refold}}$. (C) Measured fluorescence levels from gene expression systems containing slow-binding ribosomes ($\Delta G_{\text{complex}} = 0.8$ kcal/mol) and varying refolding energies. (D) The Markov model solution with $a_1 = 70$ au and varying $\Delta G_{\text{refold}}$. RNA folding times in Markov models are the same as in Figure 7-6. Fluorescence levels from no-hairpin controls are shown as horizontal dashed lines. Data points and error bars are the average and s.d. of four independent measurements on two separate days.
Figure 7.12. Non-equilibrium Markov model solutions are close to thermodynamic model predictions when 30S binding rate is slow. Markov model predictions are compared with the thermodynamic model predictions at a wide range of RNA folding stability ($\Delta G_{\text{refold}}$ varies from -15 to -3 kcal/mol). $a_{\text{fold,fast}} = 2000$ au, $a_{\text{fold,slow}} = 20$ au, $a_{\text{elong}}$ is 1000-fold faster than $a_1$.

7.4. Discussion and conclusion

In summary, the Ribosome Drafting mechanism explains how the ribosome’s initial binding rate, the mRNA’s refolding kinetics, and the mRNA’s folding free energies collectively control its translation initiation rate. Using designed fast-folding and slow-folding 2-state RNA hairpins with similar folding energetics, we show experimentally that Ribosome Drafting altered protein synthesis rates by over 1000-fold. By comparing our results to a simple Markov model, we derived the general conditions for when cycles of translation initiation inside cells operate under non-equilibrium conditions. Ribosome Drafting has the highest impact when the ribosome’s binding rate is fast, the mRNA’s folding kinetics are slow, and inhibitory hairpins are stable. By expanding our approach, it is now possible to develop more complex non-equilibrium...
Markov models that predict the translation rates of arbitrary mRNA sequences with multi-state folding pathways.

Our results have widespread implications because, while RNA plays such a central role in controlling several biological processes, the effects of RNA folding kinetics have largely been ignored. In addition to controlling translation initiation, RNAs are responsible for assembling viral particles, controlling transcriptional initiation and termination, controlling translational coupling, sensing metabolite and protein levels, and dynamically changing gene expression levels (11, 28, 188–191). Our results provide definitive evidence that RNA folding kinetics greatly affects the outcome of a biological process, while showing how to combine RNA folding simulations, rational sequence design, and Markov models to systematically study such non-equilibrium systems.
Chapter 8

An Integrated Non-Equilibrium Model of Translation Initiation

This work was done in collaboration with Ali Khodayari. All the measurements and model predictions are presented in Supplementary_Data_8 and can be found online at: http://salislab.net/files/Espah_Borujeni_ThesisSupplementaryData.zip.

8.1. Introduction

Predicting protein synthesis rate directly from DNA sequence has been a long-standing challenge. To date several physics-based models have been developed to explain how the sequence of DNA controls its interactions with other cellular machineries such as RNA polymerase, transcription factors, ribosome, RNases, and non-coding regulatory RNAs (12, 21, 55, 57, 59, 170, 192–200). These models commonly assume that their system can reach thermodynamic equilibrium, while ignoring the possibility that they can also operate under non-equilibrium conditions. In particular, translation initiation, which is a rate limiting step in protein synthetic, is assumed to be controlled by the thermodynamics of mRNA folding and mRNA-ribosome binding at equilibrium conditions (12, 21). However, as we showed in Chapter 7, due to the continuous nature of translation initiation, a Ribosome Drafting mechanism can shift the mRNA-ribosome binding toward non-equilibrium conditions, for which the kinetics of mRNA folding plays a crucial role. The extent of Ribosome Drafting is largely controlled by the time difference between the RNA folding and 30S ribosome binding. Although one can predict the
kinetics of RNA folding in silico, it is still unclear how the kinetics of ribosome binding can be predicted directly from DNA sequence.

According to Ribosome Drafting mechanism, fast-binding 30S ribosomes have access to the non-equilibrium, intermediate folding states of mRNA that are otherwise inaccessible for ribosome binding under thermodynamic equilibrium conditions. Several algorithms, including Kinfold (145), have been developed to predict the kinetics of RNA folding directly from its sequence. However, a main challenge in using the in silico predicted RNA folding kinetics is that their predictions may not represent the in vivo RNA folding timescales.

Similar to RNA folding process that includes many intermediate states, the binding of mRNA to ribosome takes place in multiple steps: 30S platform binds to mRNA’s standby site, 3’-end of 16S ribosomal RNA hybridizes to mRNA’s SD, mRNA is anchored at P-site by tRNA^{Met}, and mRNA is loaded into the 30S’s mRNA channel (12, 22, 23, 44). Although extensive body of evidences show how other molecular interactions such as tRNA binding, assembly and release of initiation factors, and ribosome translocation can kinetically control initiation process (17, 18, 25, 34, 201, 202), less is known about how mRNA-30S ribosome binding is kinetically controlled. The relative timescale by which mRNA-30S ribosome interactions take place with respect to RNA folding is currently unknown. Developing a model that combines the individual mRNA-30S ribosome binding steps with diverse RNA folding states can provide a broader picture of what happens during translation initiation in a molecular level.

Here we first designed and characterized 52 synthetic mRNA sequences that contain carefully designed RNA hairpins with diverse folding kinetics. We placed these RNA hairpins at the N-terminal coding sections of mRFP1 fluorescent protein reporter, and measured their steady-state fluorescence levels. We found that, in the presence of fast-binding 30S ribosomes, slow-folding mRNAs have higher protein expression levels than their fast-folding counterparts, demonstrating the generality of Ribosome Drafting mechanism during translation initiation.
However, the extent of Ribosome Drafting on N-terminal RNA hairpins was lower than SD-sequestering RNA hairpins, which cannot be explained using the current two-state model of mRNA-ribosome binding during translation initiation. In addition, by designing specific RNA hairpins with more than one folding transitions, we showed that the two-state model of RNA folding is not valid for hairpins with complicated folding pathways. These challenges motivated us to develop a non-equilibrium model of translation initiation that incorporates additional mRNA-ribosome interactions, and can be applied to mRNAs with complex folding transitions.

We propose a scalable computational framework that combines the Ribosome Drafting mechanism with RNA folding kinetics, mRNA transcription, mRNA degradation, and ribosome translocation, to predict the overall translation rate under both equilibrium and non-equilibrium conditions. This framework simplifies the formation of an RNA hairpin into a four-state folding pathway, for which the transition rates are computationally determined by minimizing the error between model predictions and Kinfold calculations. The 30S ribosome binding is also assumed to be a two-step process, whereby 30S ribosome first hybridizes to mRNA and anchors it to P-site using tRNA\textsuperscript{Met}, and then it loads the anchored-mRNA to the mRNA channel, followed by 50S subunit recruitment. In addition, the model simulates 70S ribosome translocation process within the first 22 codons of coding section to incorporate the effect of elongation on translation initiation. Importantly, the framework calculations are designed to be based on Kinfold timescale. The main advantage of this approach is that, once the timescale of Kinfold’\textquotesingle s predictions is experimentally validated inside cells, we will be able to relate the model predictions to actual protein synthesis rates inside cells.

Using our proposed computational framework, we developed a non-equilibrium Markov model of translation initiation that contains 443 distinct states of mRNA folding, 30S initiation, and 70S elongation, where the states are connected using 29 specific transition rates (all in Kinfold timescale). We parameterized the Markov model by identifying 12 unknown variables
that minimize the relative error between model predictions and measured fluorescence levels, for 119 synthetic mRNA sequences. Finally, our non-equilibrium Markov model can accurately predict the steady-state protein production rate by within 4.1-fold error. More importantly, the Markov model can predict the difference between fast-folding and slow-folding mRNAs, which otherwise cannot be predicted using the equilibrium-based model. Overall, our proposed integrated framework is the first step in building a global non-equilibrium model of translation rate. To achieve a greater accuracy, this framework can be expanded to include more intermediate RNA folding states, more distinct mRNA-ribosome binding steps, and a codon-dependent ribosome translocation process.

8.2. Materials and methods

8.2.1. Plasmids and strains

Eight expression plasmids (pFTV1 vector, ColE1, CmR) (Chapter 2) containing an mRFP1 fluorescent protein reporter, with unstructured N-terminal coding section, were used in this study as the initial vectors for cloning. A σ70 constitutive promoter (BioBrick #J23100) derived the expression of all mRNAs. The designed RNA hairpins with diverse folding kinetics were constructed using standard cloning techniques, and inserted at mRFP1 N-terminal coding section between NdeI and SacI restriction sites, followed by transformation of electrochemically competent Escherichia coli DH10B cells. All constructs were sequence-verified. All mRNA sequences are presented in the Supplementary_Data_8.
8.2.2. Growth and characterization

*E. coli* cells transformed with the correct clones were characterized using a 12 hr serial dilution experiment, as described in Chapter 7. The single-cell, steady-state fluorescence levels were measured using Fortessa flow cytometer (BD Biosciences) with 100,000 events threshold. The background *E. coli* DH10B auto-fluorescence was subtracted from all measured fluorescence levels. All the reported fluorescence levels are the average of 2 measurements in two separate days (n=4), and are listed in Supplementary_Data_8.

8.2.3. Equilibrium-based biophysical model of translation initiation

In Chapter 2, we described the free energy model of mRNA-ribosome binding to calculate the translation initiation rate under thermodynamic equilibrium conditions. Briefly, the rate of 30S ribosome binding to mRNA is controlled by its total free energy change, $\Delta G_{\text{total}}$, which is calculated according to Equation 8.1.

$$
\Delta G_{\text{total}} = \Delta G_{\text{standby}} + \Delta G_{\text{SD-antiSD}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} - (\Delta G_{\text{mRNA}} - \Delta G_{\text{pre-SD}} - \Delta G_{\text{post-footprint}})
$$

(8.1)

$\Delta G_{\text{standby}}$ is the energy penalty for binding of 30S platform to an inaccessible standby site upstream of Shine-Dalgarno sequence (SD) (12). $\Delta G_{\text{spacing}}$ is the energy penalty for non-optimal spacing between SD and start codon (4, 21). $\Delta G_{\text{start}}$ is the hybridization energy between tRNA $^{\text{fMet}}$ anti-codon loop and the mRNA’s start codon. $\Delta G_{\text{SD-antiSD}}$ is the hybridization energy between mRNA’s SD and 30S ribosome’s antiSD. When 30S ribosome binds to mRNA, all secondary structures overlapping with 30S ribosome binding must be unfolded, while non-clashing structures are allowed to refold. 30S ribosome footprint was measured to cover the first 15 nucleotides of coding section (Chapter 5). Therefore, $\Delta G_{\text{pre-SD}}$ and $\Delta G_{\text{post-footprint}}$ are the
folding free energy of mRNA at pre-SD and post-footprint regions when mRNA is bound to 30S ribosome. Finally, $\Delta G_{mRNA}$ is the folding energy of free mRNA before binding to 30S ribosome.

The above equation is derived based on a two-state ribosome binding model, in which mRNA exists either in freely folded state with minimum free energy (MFE), or in the ribosome-bound state. Therefore, based on an equilibrium assumption between the ensemble of 30S ribosomes and the ensemble of mRNAs inside cell, statistical thermodynamics principles predicts the translation initiation rate ($r$) according to Equation 8.2, where $\beta$ is the apparent Boltzmann constant (0.45 mol/kcal) (21) and $K$ is a proportionality constant.

$$ r = K \cdot \exp\left(-\beta \Delta G_{\text{total}}\right) \quad (8.2) $$

Importantly, this equilibrium model ignores several important interactions that can have significant impacts on overall translation rate, including the kinetics of mRNA folding, rates of mRNA degradation, or rates of ribosome elongation. Incorporation of such interactions can improve the model accuracy under both equilibrium and non-equilibrium conditions.

### 8.2.4. A Genetic Algorithm optimization for model identification

The model identification steps were performed using a real coded genetic algorithm (GA) implementation in MATLAB, as detailed below:

*Genotype and phenotype of individuals:* Each set of variables in the initial population was represented by a chromosome consisting of a set of genes, where each gene represents a real number between the variables’ lower and upper bounds. For our non-equilibrium Markov model of translation initiation, we defined 12 genes in each chromosome representing the 12 variables in the model. The lower and upper bounds were set to 0 and 0.05 (Kinfold units) for all variable except for $\beta$, which was allowed to change between or 0.3 and 2 (mol/kcal).
Population size: In all simulations the population size was set to 8 times of the number of problem variables (i.e., genes) in the chromosomes.

Initialization of population: Only one chromosome was initialized with the solution obtained in the previous iteration and the rest were randomly initialized, using the built-in GA function in MATLAB.

Crossover operator: We performed the optimization using different crossover probabilities. We found that a crossover probability between 0.6 and 0.8 for early generations, and between 0.8 and 0.9 for later generations provides the best performance.

Termination criterion: The GA procedure was terminated when the fitness of the elite chromosome was sufficiently high, or if no improvement was observed in the elite chromosome after 40 generations. We also set a maximum number of generations (2000) for each optimization problem after which the procedure was terminated.

The optimization simulations were performed on a single node with an Intel Xeon X5675, 3.07 GHz, 12 cores per node, 96 GB RAM on Penn State Lion-X clusters.

8.2.5. in silico determination of RNA folding kinetics using Kinfold program

The folding time distribution for every single RNA hairpin is calculated using 1000 Kinfold trajectories, as described in Chapter 7. In addition, the RNA hairpin’s folding free energy profile is also calculated as the average of 1000 folding free energy profiles across 100 time points that are equally binned in log scale between 0.1 and $10^5$ (kfold arbitrary units). Examples of folding time distribution and folding free energy profile are shown in Figure 8-6C.
8.2.6. Coarse-graining the Kinfold trajectories using a four-state RNA folding model

The complicated kinetics of RNA folding can be coarse-grained using a small number of RNA folding states. We simulate the dynamics of RNA folding using a four-state RNA folding model with six transitioning steps (Figure 8-6B), where the probability of each state can be calculated using a set of ordinary differential equations, as shown below.

\[
\frac{dU(t)}{dt} = -\left(K_1 + K_2\right)U(t)
\]
\[
\frac{dI_1(t)}{dt} = K_1 \cdot U(t) + K_3 \cdot I_2(t) - \left(K_4 + K_5\right)I_1(t)
\]
\[
\frac{dI_2(t)}{dt} = K_2 \cdot U(t) + K_4 \cdot I_1(t) - \left(K_5 + K_6\right)I_2(t)
\]
\[
\frac{dF(t)}{dt} = K_5 \cdot I_1(t) + K_6 \cdot I_2(t)
\]

U and F are the normalized probability of unfolded and fully folded RNA hairpin states, respectively, while I_1 and I_2 are the normalized probability of two intermediate states with unknown structures and energies. We employed the \textit{ode15s} function in Matlab to solve this set of ODEs, and determined the time-dependent probability of each state (Figure 8-6C). At time zero, all RNA hairpins exist in the unfolded state (U(t=0) = 1). Importantly, each of these states has a specific energy, where the summation of these energies determines the net energy of RNA folding ensemble per time, according to Equation 8.3.

\[
\Delta G_{\text{fold}}(t) = U(t) \cdot \Delta G_U + I_1(t) \cdot \Delta G_{I_1} + I_2(t) \cdot \Delta G_{I_2} + F(t) \cdot \Delta G_F
\] (8.3)

In above equation, \Delta G_U is 0 kcal/mol, \Delta G_F is the hairpin’s minimum free energy (MFE), and \Delta G_{I_1} and \Delta G_{I_2} are two unknown energies to be identified by optimization algorithm. In addition, by solving the probability of fully folded RNA state (F) at steady-state, we can determine the RNA hairpin folding time, \tau, also known as the first passage time (F(t = \tau) = 1).
Therefore, for any given RNA hairpin, we first calculate the folding time distribution and the folding free energy landscape, using 1000 Kinfold trajectories. Next, by incorporating the 
*ode15s* function into our optimization algorithm, we identify six K-values (K<sub>1</sub> to K<sub>6</sub>) and two energy contents (ΔG<sub>I-1</sub> and ΔG<sub>I-2</sub>) that can minimize the relative error between predicted ΔG<sub>fold</sub> and τ, and the their corresponding Kinfold calculations (Figure 8-6C).

### 8.2.7. An integrated non-equilibrium model of translation initiation

We consider the following molecular interactions in our non-equilibrium model of translation initiation: mRNA refolding, mRNA-30S ribosome hybridization, mRNA loading, 70S ribosome translocation, mRNA transcription, and mRNA degradation. Below we explain how we incorporate them into our model.

- *mRNA refolding*: As we show in the Results section, the folding kinetics of hairpins located exclusively inside the N-terminal coding section has a different effect on translation initiation than hairpins sequestering SD and start codon. Therefore, we divide all mRNA structures into two distinct categories: RNA hairpins that sequester SD and start codon, and RNA hairpins located inside the coding section (Figure 8-7B). A random mRNA can have RNA hairpins in both categories, or only from one of the categories, or with no hairpin altogether. We model the RNA hairpin formation in each category using our proposed four-state RNA folding, in which the transition rates and intermediate energy contents are determined as explained above. Importantly, since RNA hairpins in these two categories fold independently, the folding of entire mRNA sequence can be modeled using 16 distinct folding states (4 x 4 states) (Figure 8-7B). The horizontal RNA folding pathways in Figure 8-7B represent the folding of hairpins sequestering SD and start codon, while vertical pathways show the RNA hairpins that fold within the coding section. We assume that each RNA hairpin can start to refold only after its entire
sequence is freed from 70S ribosome (Figure 8-7A). 70S ribosome’s footprint is assumed to cover 30 nucleotides of mRNA, which is equivalent of 10 codons: 5 codons upstream of P-site and 5 codons downstream of P-site, including the P-site itself (see Chapter 5). Therefore, in our non-equilibrium model calculations, we identify the codon in which the hairpin’s 3’-end nucleotide is located, and we call it $C_{HP}$. Therefore, the refolding of RNA hairpin starts once the 70S ribosome’s P-site reaches a codon position greater than $C_{HP} + 6$.

- **mRNA-30S ribosome hybridization**: 30S ribosome can bind to mRNA though SD-antiSD hybridization followed by anchoring mRNA to the P-site using tRNA$^{\text{Met}}$ (Figure 8-7A). We assume that before hybridizing to SD, 30S ribosome binds to the mRNA’s standby site upstream of SD, and gets ready to hybridize once the elongating 70S ribosome clears the initiation site (70S ribosome’s P-site at the 11th codon). The net binding free energy between 30S ribosome and an unfolded SD region, $\Delta G_{\text{complex}}$, is calculated according to Equation 8.4.

$$
\Delta G_{\text{complex}} = \Delta G_{\text{standby}} + \Delta G_{\text{SD-antiSD}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}}
$$

(8.4)

All four free energies in above equation have been described previously, and can be calculated directly from mRNA sequence (see Chapter 2). We then relate $\Delta G_{\text{complex}}$ to the rate of 30S ribosome binding to mRNA, using statistical thermodynamic formulation shown in Equation 8.5, where $R_1$ and $\beta$ are the proportionality constant and the Boltzmann constant, respectively.

$$
V_i = R_1 \cdot \exp(-\beta \cdot \Delta G_{\text{complex}})
$$

(8.5)

All inhibitory RNA hairpins sequestering SD and start codon must be unfolded prior to the SD-antiSD hybridization. The rate of 30S ribosome binding in the presence of inhibitory RNA hairpins drops according to the hairpins’ unfolding free energy. Since we assume a four-state folding pathway for the formation of RNA hairpins, we can calculate the distribution of 30S ribosome binding to the ensemble of RNA folding states (Figure 8-7B), as shown below.
\[ V_2 = R_1 \cdot \exp(-\beta \cdot (\Delta G_{\text{complex}} - \Delta G_{I_1})) \] (8.6)

\[ V_3 = R_1 \cdot \exp(-\beta \cdot (\Delta G_{\text{complex}} - \Delta G_{I_2})) \] (8.7)

\[ V_4 = R_1 \cdot \exp(-\beta \cdot (\Delta G_{\text{complex}} - \Delta G_F)) \] (8.8)

\[ \Delta G_{I_1} \] and \( \Delta G_{I_2} \) are the folding free energy of intermediate folding states as described in previous section, and \( \Delta G_F \) is the hairpin’s minimum free energy (MFE).

**mRNA loading:** After mRNA has been anchored to P-site, it will be loaded into the mRNA channel within the 30S ribosome (Figure 8-7A). We assume that the rate of mRNA loading is sequence-independent and remains constant for all mRNA sequences. However, presence of inhibitory RNA hairpins reduces the loading rate. Similar to hybridization step, mRNA loading rate can be distributed between the ensemble of four 30-bound RNA folded states (Figure 8-7B), as shown below.

\[ V_5 = R_2 \] (8.9)

\[ V_6 = R_2 \cdot \exp(-\beta \cdot (-\Delta G_{I_1})) \] (8.10)

\[ V_7 = R_2 \cdot \exp(-\beta \cdot (-\Delta G_{I_2})) \] (8.11)

\[ V_8 = R_2 \cdot \exp(-\beta \cdot \alpha \cdot (\Delta G_{\text{HP\_post\_footprint}} - \Delta G_F)) \] (8.12)

In above equations, \( R_2 \) is the constant, sequence-independent RNA loading rate. In **Equation 8.12**, if the fully folded RNA hairpin partially overlaps with the 30S ribosome footprint, we only consider the energy required to unfold the overlapped portion of RNA hairpin. \( \Delta G_{\text{HP\_post\_footprint}} \) is the folding free energy of the remaining portion of RNA hairpin. In addition, as we observed in Chapter 5, the apparent unfolding free energy of N-terminal RNA hairpins that overlap with 30S ribosome footprint was 1.6-fold stronger than their predicted unfolding free energies. Therefore, we incorporated a new parameter \( \alpha \) in **Equation 8.12**, where \( \alpha = 1.6 \).
Furthermore, as we show in the Results section, when the ribosome binding is strong \((\Delta G_{\text{complex}} = -8.3 \text{ kcal/mol})\), presence of hairpins with low stability \((\Delta G_{\text{refold}} = -4.6 \text{ kcal/mol})\) in the N-terminal coding section is not inhibitory at all; it even increases the translation initiation rate beyond that of no-hairpin control mRNA (Figure 8-3B). We hypothesize that a strong ribosome binding, for example though stable SD-antiSD hybridization, creates a potential momentum for the 30S ribosome that can unfold weakly stable RNA hairpins located close to the start codon. Interestingly, we did not observe this striking effect of RNA hairpins when the ribosome binding itself was quite instable \((\Delta G_{\text{complex}} = 0.8 \text{ kcal/mol})\) (Figure 8-3A). Therefore, we included a constraint in our non-equilibrium model calculations that automatically stops the folding of RNA hairpins within the N-terminal coding section (all K-values = 0) if the ribosome binding energy is stronger than hairpin unfolding energy, according to:

\[
\Delta G_{\text{complex}} < \alpha \left( \Delta G_F - \Delta G_{\text{HP-post-footprint}} \right).
\]

- **70S ribosome translocation**: After mRNA is successfully loaded and 50S ribosome subunit has joined, the initiation step is complete and the ribosome (now called 70S) departs from the initiation site for peptide elongation. Several factors control the rate of 70S ribosome elongation, including tRNA arrival time, ribosome translocation speed, and formation of mRNA secondary structures inside the coding section (20, 29, 30, 138, 203, 204). Here for simplicity we assume that the rate of 70S ribosome translocation, \(V_g\), is constant for all codons, and secondary structures do not alter this rate, as observed previously in Chapter 5. Notably, this assumption can be readily modified in the future by incorporating a codon-dependent translocation rate. Regardless of translocation speed, the movement of 70S ribosome can be monitored at a codon-based resolution (Figure 8-8), until it reaches the stop codon at the end of coding section. However, for simplicity we only consider the first 21 translocation steps, where the last step places the 70S’s P-site at the 22\textsuperscript{nd} codon of coding section. This 22-codon distance is long enough
to allow all the designed RNA hairpins in our mRNA sequences to refold to their fully folded state. For example, our longest inhibitory RNA hairpin ends at the 16th codon of the coding section, which is just outside the 70S ribosome footprint when the P-site is at 22nd codon.

Importantly, once 70S ribosome reaches to this checkpoint, binding of new 30S ribosome to initiate the next cycle of translation becomes independent of 70S ribosome elongation (Figure 8-8). 70S ribosome continues its elongation steps with the same speed until reaching the stop codon, at a time equal to \((C_{\text{total}} - 21)/V_9\) , in Kinfold units. \(C_{\text{total}}\) is the total number of codons in the coding section, which is equal to 230 for our mRFP1 protein reporter. Therefore, at steady-state, we can assume that 70S ribosome travels the distance between 22nd codon and the stop codon, at a rate equal to \(V_{\text{elongate}} = V_9 / (C_{\text{total}} - 21)\).

After the first elongating 70S ribosome has reached the 11th codon, a new 30S ribosome can bind to mRNA, initiate translation, and form the second 70S ribosome that elongates on mRNA with the same speed as the first 70S ribosome (Figure 8-8). If the binding of 30S ribosome is fast enough, the first and second 70S ribosomes can translocate with zero codon distance (maximum ribosome density). However, if the 30S ribosome binding is much slower, for example due to formation of inhibitory structures, the two 70S ribosomes translocate at longer distances. Since we are only studying the first 22 codons, the distance between two 70S ribosomes can vary between 0 and 10 codons. In addition, at the lowest 30S ribosome binding rate, there will be only one elongating 70S ribosome within the first 22 codons (minimum ribosome density). In total, there are 13 different possibilities that a new translation can initiate with respect to the previous translation (Figure 8-8). For each of these initiation possibilities, we simulate the elongation step codon-by-codon, until the first 70S ribosomes reach the 22nd codon. For example, a single 70S ribosome has 21 different elongation states on mRNA, all connected with the translocation rate \(V_9\), whereas two 70S ribosomes with zero codon distance will have
only 11 elongation states (Figure 8-8). Overall, protein is produced from these 13 initiation possibilities (in total 89 distinct elongation states) (Figure 8-8).

- **mRNA transcription:** We assume that new mRNA molecules are generated with a constant transcription rate, \( T \) (Figure 8-7A). The newly transcribed mRNA molecules are assumed to be fully unfolded, with no hairpins sequestering SD and start codon, and no N-terminal structures within the first 16 codons of coding section.

- **mRNA degradation:** When the mRNA’s initiation site is not bound to 30S ribosome, it can be a target for RNA degradation. In addition, presence of secondary structures can promote RNA degradation (Figure 8-7A). Therefore, in our model, we assigned \( \delta_U \), \( \delta_i \), and \( \delta_F \) as constant, sequence-independent degradation rates for all 30S-unbound mRNAs with either no structures, intermediate folding structures, or fully folded structures, respectively. We also assume that mRNAs can be degraded while hybridized to 30S ribosome and waiting to be loaded to the mRNA channel. In particular, presence of a secondary structure that overlaps with the 30S ribosome footprint can inhibit the loading process, and therefore can accelerate the degradation rate. Therefore, we assigned \( \delta_{30S-U} \), \( \delta_{30S-i} \), and \( \delta_{30S-F} \) as additional constant, sequence-independent degradation rates for all 30-bound mRNAs with either no structures, intermediate folding structures, or fully folded structures within the 30S ribosome footprint region, respectively.

Finally, by combining all the interactions explained above, we developed a Markov model for non-equilibrium translation initiation that contained 443 distinct states of mRNA folding, 30S ribosome initiation, 70S ribosome elongation, where states were connected using 29 different transition rates. By solving the Markov model using ode15s function in Matlab, we calculate the time-dependent probability of all the states, including the state in which the elongating ribosomes complete the protein production (Figure 8-8). We assume that at time zero, all the states have zero probabilities. Therefore, for each mRNA sequence, we calculate the
steady-state probability of protein production, and compare it with the measured fluorescence levels at steady-state. An additional unknown variable, $V_{10}$, was considered as a proportionality constant that relates the calculated probabilities to the measured fluorescence levels. Finally, by employing our GA optimization algorithm, we identified a set of 12 variables that minimized the average error between the predicted and measured protein expression levels, across 119 mRNA sequences characterized in this chapter, and Chapters 5 and 7.

8.3. Results

8.3.1. Ribosome Drafting controls the rate of translation initiation when RNA hairpins are located within N-terminal coding section

Secondary structures that overlap with 30S ribosome footprint at mRNA’s N-terminal coding section must be unfolded before 30S ribosome can load mRNA into its narrow mRNA channel. We previously reported that location and thermodynamics of N-terminal RNA hairpins controls the extent of translation inhibition (Chapter 5). However, it is not clear how the folding kinetics of the N-terminal RNA hairpins can regulate translation rate. In particular, folding kinetics of RNA hairpins, sequestering SD and start codon, has been shown to control the rate of translation initiation via Ribosome Drafting mechanism (Chapter 7). Here, we designed and characterized 52 synthetic mRNA sequences with diverse ribosome binding rates, RNA folding energetics, and RNA folding kinetics to investigate whether or not Ribosome Drafting can control the function of N-terminal RNA hairpins during translation initiation.

By combining RNA folding simulations with an optimization algorithm (Figure 8-1A), we designed three groups of RNA hairpins with increasing stabilities ($\Delta G_{\text{refold}} = -4.6, -7.8,$ and -
11.9 kcal/mol), where each group contained 2 RNA hairpins with similar folding free energies but with different folding kinetics (Figure 8-2). In addition, to avoid confounding interactions, hairpins were only selected if they exhibited a single folding transition; if they introduced no additional start codon or stop codon; if their sequences only encode for fast codons to maximize translation elongation rates; and if they contained no more than ten consecutive base pairings to minimize RNAse-mediated degradation. Also since these hairpins were located exclusively inside the coding section, the design of their sequence was not constrained by the SD sequence (higher degrees of freedom).

We inserted these designed RNA hairpins into our expression system such that the hairpins were located immediately after the start codon of mRFP1 protein (position +4) (Figure 8-1B). It has been shown that hairpins at this position have the highest inhibitory effects on translation initiation (Figure 5-2). We then characterized these RNA hairpins using both slow-binding ribosomes ($\Delta G_{\text{complex}} = 0.8$ kcal/mol) and fast-binding ribosomes ($\Delta G_{\text{complex}} = -8.3$ kcal/mol) (see Chapter 7). The measured steady-state fluorescence levels show that, similar to SD-sequestering RNA hairpins, Ribosome Drafting mechanism is active at the presence of N-terminal RNA hairpins, only when the 30S ribosome binding is fast (Figure 8-3). In particular, when the 30S ribosome binding rate was fast ($\Delta G_{\text{complex}} = -8.3$ kcal/mol), thermodynamics of RNA folding controlled the impact of Ribosome Drafting on translation rate (1.5-, 10.5-, and 35.6-fold increase in translation rate for $\Delta G_{\text{refold}} = -4.6$, -7.8, and -11.9 kcal/mol, respectively) (Figure 8-3B). This finding is consistent with our Markov model of Ribosome Drafting (Chapter 7), in which RNA hairpins with higher stability have larger impact on translation rate. However, the observed impact was much lower than SD-sequestering RNA hairpins. For example, Ribosome Drafting caused 1110-fold increase in translation rate of an mRNA with SD-sequestering RNA hairpin ($\Delta G_{\text{refold}} = -7.8$ kcal/mol, $\Delta G_{\text{complex}} = -8.3$ kcal/mol) (Figure 7-11),
while an N-terminal RNA hairpin with similar folding energy and the same 30S ribosome binding rate altered translation rate by only 10.5-fold (Figure 8-3B).

Figure 8-1. Design and characterization of N-terminal RNA hairpins with diverse folding kinetics.

(A) By combining the Kinfold calculations with an optimization algorithm, we designed pairs of RNA hairpins that have equal folding energies but very different folding times. (B) We characterized their regulatory effects on translation initiation by placing them at the N-terminal coding section of mRFP1 fluorescent protein reporter, immediately after mRFP1 start codon (position +4). The A-C rich insulator sequence prevented the formation of undesired long-range base-pairings.
Figure 8-2. Sequence and structure of designed RNA hairpins to demonstrate Ribosome Drafting for N-terminal RNA hairpins. Folding time distributions are the result of 1000 Kinfold trajectories. Experimental data corresponding to these hairpins are shown in Figure 8-3.
Figure 8-3. Ribosome Drafting’s impact on translation rate is controlled by the folding free energy of N-terminal RNA hairpins. (A) Measured fluorescence levels from gene expression systems containing inhibitory RNA hairpins with slow-binding ribosomes ($\Delta G_{\text{complex}} = 0.8$ kcal/mol) and varying refolding energies ($\Delta G_{\text{refold}}$). (B) Measured fluorescence levels from gene expression systems containing fast-binding ribosomes ($\Delta G_{\text{complex}} = -8.3$ kcal/mol) and varying refolding energies. Fluorescence levels from no-hairpin controls are shown as horizontal dashed lines. Data points and error bars are the average and s.d. of four independent measurements on two separate days ($n = 4$).

Although it is expected that N-terminal RNA hairpins have a large inhibitory effect on translation rate, we found that mRNAs with the least stable N-terminal RNA hairpins ($\Delta G_{\text{refold}} = -4.6$ kcal/mol) have higher fluorescence levels than the no-hairpin control mRNA, suggesting that these weakly stable N-terminal RNA hairpin are not folded at all. Importantly, this was only observed when the ribosome binding was fast and stable ($\Delta G_{\text{complex}} = -8.3$ kcal/mol) (Figure 8-3B). We hypothesize that when 30S ribosome binding is fast (strong SD-antiSD hybridization),
30S ribosome gains an extra momentum that can unfold weakly stable N-terminal RNA hairpins without expending any unfolding energy penalty. However, a slow-binding ribosome does not create enough momentum to unfold hairpins without exerting extra unfolding energy, causing the same N-terminal RNA hairpins to be inhibitory for translation initiation rate (Figure 8-3A). Although further investigation is required to determine the exact role of weakly stable RNA hairpins inside the coding section, we have incorporated this hypothesis in our non-equilibrium model of translation initiation as an additional design rule to predict the effect of weakly stable RNA hairpins (see Materials and methods section).

8.3.2. The extent of Ribosome Drafting is reduced for N-terminal RNA hairpins

We further studied how the 30S binding rate controls the extent of Ribosome Drafting for the N-terminal RNA hairpins. For that purpose, we used a library of six no-hairpin control mRNAs, for which the 30S binding rate vary by 1000-fold across the library by mutating the mRNA’s SD sequence. We should mention that, this library was used previously to study the extent of Ribosome Drafting on SD-sequestering RNA hairpins (Figure 7-6). Next, for each member of the library, we inserted fast-folding and slow-folding RNA hairpins immediately after the start codon of mRFP1 coding section (position +4). The RNA hairpins have similar folding energies ($\Delta G_{\text{refold}} = -7.8$ kcal/mol) but different folding kinetics. We also repeat the same experiments with a more stable pair of RNA hairpins ($\Delta G_{\text{refold}} = -11.9$ kcal/mol) (Figure 8-4A).

Our measured steady-state fluorescence levels revealed that the impact of Ribosome Drafting for N-terminal RNA hairpins is much lower compared to the SD-sequestering RNA hairpins (Figure 8-4C; for comparison see Figure 7-6). While our Markov model of Ribosome Drafting predicts that a more stable RNA hairpin has a larger impact on translation initiation, our
measurements showed an opposite effect for N-terminal RNA hairpins. In particular, the maximum increase in translation rate between slow-folding and fast-folding RNA hairpins was 3.6-fold versus 12.2-fold for RNA hairpins with $\Delta G_{\text{refold}}$ of -11.9 and -7.8 kcal/mol, respectively (Figure 8-4BC). For comparison, the maximum increase in fluorescence levels of SD-sequestering RNA hairpins with $\Delta G_{\text{refold}} = -8.0$ kcal/mol was 1034.2-fold. These discrepancies suggest that the folding kinetics of N-terminal RNA hairpins control the rate of translation initiation very different from SD-sequestering RNA hairpins.

Two interactions may be responsible for the observed dramatic reduction in Ribosome Drafting activity. First, RNA hairpins that overlap with the 30S ribosome footprint and inhibit the loading process may promote RNAse activity, resulting in a reduced mRNA stability. This will lower the impact of RNA folding kinetics on non-equilibrium translation rate. Second, mRNA-ribosome binding process may take place in two consecutive steps, in which 30S ribosome first binds to mRNA through SD-antiSD hybridization and anchors the mRNA to the P-site, and second, 30S ribosome loads the bound mRNA into the mRNA channel. Since these two steps occur in series, both fast-folding and slow-folding RNA hairpins inside the coding section will have enough time to refold before the loading starts, which reduces the extent of Ribosome Drafting. Importantly, our current Markov model of Ribosome Drafting (Figure 7-1B) does not consider mRNA degradation rate, and only assumes a one-step mRNA-ribosome binding process. Incorporation of these two interactions into our non-equilibrium model of translation initiation can improve its accuracy for wide range of mRNA sequences.
Figure 8-4. Extent of Ribosome Drafting is reduced for N-terminal RNA hairpins. (A) Design of mRNAs with varying SD sequences and 30S binding rates. Thermodynamically equivalent, fast-folding and slow-folding RNA hairpins were inserted immediately after start codon (position +4) to inhibit translation. (B and C) Measured fluorescence levels of no-hairpin control mRNAs (black), slow-folding mRNAs (blue), and fast-folding mRNAs (red) are shown with respect to the predicted 30S binding rates, according to formula: $2500 \times \exp(-0.45 \Delta G_{\text{complex}})$. Data points and error bars are the average and s.d. of four independent measurements on two separate days ($n = 4$).
8.3.3. Effect of Ribosome Drafting on RNA hairpins with two folding transitions

All our characterized RNA hairpins were specifically designed to have a single folding transition (see Figure 7-5 for two examples). Our current Markov model of Ribosome Drafting (Figure 7-1B) is also based on a two-state RNA folding process with a constant folding rate. However, most natural RNA hairpins have complex folding pathways and will not exhibit a single folding transition. Therefore, to investigate how RNA hairpins with multiple folding transitions are affected by Ribosome Drafting, we designed three RNA hairpins that exhibit two distinct folding transitions with largely separate folding time distributions (i.e. both a fast and a slow) (Figure 8-5). These hairpins were designed to have a wide range of folding free energies (\(\Delta G_{\text{refold}} = -5.3, -8.3,\) and -11.8 kcal/mol). We inserted the most stable hairpins (\(\Delta G_{\text{refold}} = -8.3\) and -11.8 kcal/mol) immediately after the start codon of mRFP1 protein (position +4), while the least stable hairpin (\(\Delta G_{\text{refold}} = -5.3\) kcal/mol) was inserted at position +7, to avoid the undesired base-pairing, present at position +4. For the comparison, we also inserted our previously designed fast-folding hairpin with \(\Delta G_{\text{refold}} = -4.6\) kcal/mol at position +7.

The current Markov model of Ribosome Drafting incorrectly considers these RNA hairpins as slow-folding hairpins (their average folding time is large, \(\tau = 610, 2235,\) and 100 Kinfold units for \(\Delta G_{\text{refold}} = -5.3, -8.3,\) and -11.8 kcal/mol, respectively), and therefore predicts a higher translation initiation rate for these RNA hairpins with respect to their fast-folding hairpin counterparts. However, the measured steady-state fluorescence levels show that RNA hairpins with two folding transitions behave like a fast-folding RNA hairpins, as if 30S ribosome can only see their fast-folding transition pathway (Figure 8-5). Regardless of the 30S binding rates (\(\Delta G_{\text{complex}} = 0.8\) or -8.3 kcal/mol), translation initiation rates of mRNAs with two folding transitions, are either similar or lower than that of mRNAs with single folding transitions. These
results demonstrate another limitation of our current Markov model of Ribosome Drafting, which assumes all RNA hairpins have single folding transitions. We will address this problem in our non-equilibrium model of translation initiation by developing a generalized model of four-state RNA folding pathway, which can be applicable to all types of folding transitions (Figure 8-6).

Figure 8-5. Translation initiation rates for RNA hairpins with two folding transitions. By combining Kinfold calculation with an optimization algorithm, three RNA hairpins (shown with roman number II) were designed to have two separate folding transitions (gray folding time distributions). The fast-folding RNA hairpins (shown with roman number I) are taken from Figure 8-2 for comparison. Bar graphs show the measured fluorescence levels of RNA hairpins with single (white bars) and double (gray bars) folding
transitions. Fluorescence levels from no-hairpin controls are shown as horizontal dashed lines. Data points and error bars are the average and s.d. of four independent measurements on two separate days (n = 4).

8.3.4. A four-state model of RNA folding

Folding dynamics of an RNA hairpin can be measured experimentally using different techniques, such as isothermal titration calorimetry (ITC) (205), laser temperature-jump (206), and time-resolved SAXS (small-angle X-ray scattering) (207). However, obtaining experimental data for every single RNA hairpin involved in translation initiation is infeasible. As an alternative approach, we employ the Kinfold RNA folding algorithm (145) to calculate both the folding time distributions and the folding free energy landscapes for any given RNA hairpin (see Materials and methods) (Figure 8-6A). One can imagine many different pathways with diverse intermediate folding states that can explain the Kinfold’s results. However, we simplified the RNA folding process into a coarse-grained, quantifiable four-state system, consisting of a single-stranded RNA sequence, two intermediate RNA folding states with arbitrary structures, and a fully folded RNA hairpin (see Materials and methods) (Figure 8-6B). Similar four-state RNA folding models have been reported before to explain experimentally measured RNA folding pathways (208, 209). It is noteworthy that, when an RNA hairpin is subject to complete unfolding, the prediction of translation initiation rate only requires the folding free energy of RNA hairpins not their structure. Therefore, for every RNA hairpin we parameterize our proposed four-state system using 8 unknown variables: 6 transition rates that connect the four states, and 2 folding free energies that represent the energy content of intermediate folding states (Figure 8-6B). Using a minimization approach, we identify these eight variables by fitting the system to the calculated Kinfold’s folding time distribution and folding free energy landscape (see Materials and methods). Depending on the folding pathway of the RNA hairpins, transition rates vary between 0 and 1
(Kinfold timescales), while intermediate folding energy contents vary between 0 and 10 kcal/mol. All the eight fitted variables for the 52 designed RNA hairpins in this chapter and chapter 5 and 7 are presented in Supplementary_Data_8. Importantly, this simple four-state model of RNA folding is accurate enough to explain the behavior of both slow-folding and fast-folding RNA hairpins, as well as RNA hairpins with complicated folding trajectories (Figure 8-6C).

Figure 8-6. A four-state, coarse-grained model of RNA folding. (A) 1000 Kinfold trajectories are used to predict the folding kinetics of a given RNA sequence. (B) A four-state model of RNA folding coarse-
grains the results of Kinfold simulation. The intermediate folding states are arbitrary states with unknown structure and energy. (C) Performance of the four-state, coarse-grained model is demonstrated by comparing its results with Kinfold’s predictions for (top row) a fast-folding hairpin with $\Delta G_{\text{refold}} = -7.8$ kcal/mol, (middle row) a slow-folding hairpin with $\Delta G_{\text{refold}} = -7.5$ kcal/mol, and (bottom row) a hairpin with two folding transitions with $\Delta G_{\text{refold}} = -8.3$ kcal/mol. (left column) Folding time distortions from 1000 Kinfold trajectories (gray bars) are shown versus four-state model predictions (red lines). (middle column) Average folding energy profiles from 1000 Kinfold trajectories (black bold line) are shown versus four-state model predictions (red dashed line). The gray areas are the standard deviations from 1000 folding energy profiles. (right column) Folding fraction profiles are shown for each state in the model. Symbols are similar to part B.

8.3.5. An integrated non-equilibrium model of translation initiation

In contrast to a thermodynamics-based model of translation initiation, we aim to incorporate both the kinetics and thermodynamics of RNA folding to predict translation initiation under both equilibrium and non-equilibrium conditions. Translation initiation is a continuous process, in which several 30S ribosomes consecutively bind to an mRNA, initiate translation, and depart from initiation site to continue the elongation. During these continuous cycles, an mRNA undergoes multiple unfolding and refolding events, and therefore may exist at non-equilibrium states depending on the speed of 30S ribosome assembly. Previously, a simplified Markov model of Ribosome Drafting was proposed that explained this non-equilibrium driven translation initiation process (Figure 7-1B). However, as shown in previous sections, this model is not accurate enough to predict translation initiation rates when: 1- RNA hairpins have multiple folding transitions, 2- mRNA-ribosome binding has several intermediate states, and 3- mRNA degradation has a large effect on translation rate. In addition, folding kinetics of RNA hairpins is
strongly coupled with 70S ribosome elongation rate. For example, refolding of a long RNA hairpin inside the N-terminal coding section can be delayed until 70S ribosome clears the entire RNA hairpin sequence. Interestingly, this interaction creates a co-dependency between the rate of translation initiation and the ribosome elongation rate at N-terminal coding section. A slow elongation rate can further delay the formation of inhibitory RNA hairpins, resulting in more Ribosome Drafting, and therefore higher translation initiation rate. Therefore, a more detailed biophysical model is necessary to predict the rate of translation initiation from all these diverse molecular interactions.

As a first step toward developing a non-equilibrium model of translation, we propose an integrated framework that includes several cellular processes (Figure 8-7A), including mRNA folding kinetics, two-step mRNA-ribosome binding, ribosome elongation, mRNA transcription, and mRNA degradation, to predict translation rate directly from DNA sequence (see Materials and methods). In particular, we assume that the mRNA molecule is generated at a constant rate, which depends on the promoter DNA sequence. Similarly, we also assume that a free mRNA molecule, regardless of its folding states, is subject to mRNA degradation at constant, sequence-independent rates. Development of additional sequence-to-function relationships that can predict the rate of mRNA degradation can further improve our model accuracy. Free mRNA molecule can then bind to 30S ribosome to initiate translation, with a rate that depends on its sequence and presence of inhibitory RNA hairpins. After 30S initiation complex is successfully formed, we assume that 70S ribosome departs from initiation site to continue the elongation step with a constant, codon-independent rate. This elongation rate dictates when the new 30S ribosome can bind to mRNA and initiate the next of round of translation, and when the inhibitory RNA hairpins can refold to inhibit the initiation step (Figure 8-7A). The time difference between these two processes leads to Ribosome Drafting. Importantly, the length of RNA hairpin can either shorten or lengthen this time difference.
To model the mRNA-30S ribosome binding, we propose a two-step translation initiation process (Figure 8-7A), in which 30S ribosome first binds to mRNA through SD-antiSD hybridization and anchors mRNA to the P-site via tRNA^{Met}, and second, 30S ribosome loads the anchored mRNA into its narrow mRNA channel to complete the initiation step, followed by 50S ribosome recruitment. While the rate of 30S ribosome hybridization in the first step can be calculated directly from DNA sequence using our previously developed ribosome energy model (Chapter 2), the loading rate of an unstructured mRNA is assumed to be constant and sequence-independent (see Materials and methods). However, presence of inhibitory RNA hairpins slows down the loading process, according to their folding free energies. Further, we also assume that, before the loading process begins, ribosome-bound mRNAs can be a target for degradation at constant, sequence-independent rates.

We next employ Kinfold’s calculations, coupled to our four-state RNA folding model, to predict the folding kinetics of RNA hairpins directly from their sequence. Since the function of SD-sequestering RNA hairpins is different from the RNA hairpins at N-terminal coding section, we separate their folding kinetics into a 4x4 Markov model of mRNA folding, in which the two types of RNA hairpins can independently fold (Figure 8-7B). Transition rates of individual RNA hairpins are obtained using our four-state coarse-grained RNA folding model (Figure 8-6). Each of these 16 mRNA folding states can either start to degrade with constant rates ($\delta_1$, $\delta_2$, or $\delta_F$) (see Materials and method) or hybridize to 30S ribosome with four specific rates ($V_1$, $V_2$, $V_3$, and $V_4$), which depends on the 30S binding affinity ($\Delta G_{\text{complex}}$) and the folding free energy of inhibitory hairpins (see Materials and methods). After the 30S ribosome has hybridized to mRNA, mRNA loading starts with specific rate ($V_5$, $V_6$, $V_7$, and $V_8$), which depends on the folding free energy of N-terminal hairpins. However, before the loading process begins, mRNA-30S complex can also
be a target for degradation with specific constant rates ($\delta_{30S-U}$, $\delta_{30S-i}$, and $\delta_{30S-F}$) (see Materials and methods).

Finally, we monitor the 70S ribosome elongation process at a codon-based resolution, until the initiation becomes independent of elongation (Figure 8-8). For simplicity, we assume that 70S ribosome can translocate with a constant, sequence-independent rate. Since the longest RNA hairpin in our characterized mRNA sequences extends until the 16th codon, we track the 70S ribosome translocation steps until its P-site reaches the 22nd codon (see Materials and methods). Importantly, during the translocation of 70S ribosome from 11th codon to 22nd codons (13 elongation steps) new 30S ribosomes can bind to mRNA to initiate the next cycle of translation, according to the two-step ribosome binding process shown in Figure 8-7B. Notably, the probability of translation initiation in each of these 13 elongation states is coupled with probability of mRNA refolding, mRNA degradation, and ribosome elongation, which creates a large network of coupled interactions with 443 distinct states. Therefore, we develop a Markov model that incorporates all these 443 states and connects them using 29 transition rates to predict the non-equilibrium protein production rate. From these 29 transition rates, 12 of them are parameterized independently using our four-state RNA folding model, and the rest of them are global transitions rates that are calculated using 12 quantifiable variables (see Materials and methods). These 12 variables are parameterized by minimizing the error between Markov model’s steady-state protein production and the measured steady-state fluorescence level for 119 synthetic mRNA sequences studied in this work.
Figure 8-7. A non-equilibrium model of translation initiation. (A) The non-equilibrium model of translation initiation consists of several cellular processes: mRNA transcription, mRNA degradation, mRNA folding, two-step mRNA-30S ribosome binding, and 70S ribosome elongation. (B) A 4x4 Markov
model of mRNA refolding distinguishes between the RNA hairpins that sequester SD and start codon (horizontal pathways), and RNA hairpins that fold at N-terminal coding section (vertical pathways). Dashed lines are the six transitions rate ($K_1$ to $K_6$) obtained for individual hairpins (Figure 8-6). The structures in the intermediate folding states are arbitrary and are shown just for comparison. 30S ribosome hybridizes to each of these folding states, followed by loading of mRNA. The rates of 30S ribosome binding ($V_1$ to $V_8$) are calculated according to the states’ folding free energies (see Materials and methods).

**Markov states for co-locating the initiating and elongating ribosomes on mRNA**
Figure 8-8. Monitoring the elongation of 70S ribosomes during translation initiation. 89 distinct states of mRNA bound with 0, 1, or 2 ribosomes are shown. Numbers in red color specify the states in which translation initiation takes place. Numbers in black color show the states in which ribosomes elongate on mRNA. All the consecutive elongation states (eq. states 2 to 22) are connected using the constant translocation rate ($V_9$). Arrows pointing toward the red numbers mean that at that state, a translation initiation can take place. For example, from state 12, the elongating ribosome can translocate to state 13 and a new 30S ribosome can initiate at state 23. Once an elongating ribosome reaches the 22nd codon, it can continue to finish the protein production. For example, at state 22, the elongating ribosome can translocate to create a free unfolded mRNA at state 1, while continuing to finish the protein production. At the same time, a new 30S ribosome can also initiate translation at state 88 as well. Overall, protein production is generated from 13 distinct initiation states.

8.3.6. Non-equilibrium model more accurately predicts the rate of translation initiation than the equilibrium model

To demonstrate the accuracy of our non-equilibrium Markov model of translation initiation, we combined all the designed mRNA sequences characterized in this chapter and Chapters 5 and 7 (119 mRNA sequences in total). These mRNA sequences have diverse 30S binding rates, and contain carefully designed RNA hairpins with diverse sequences, folding free energies, folding kinetics, and folding transitions (52 RNA hairpins in total). The designed RNA hairpins either sequester SD and start codon, or overlap with the 30S ribosome footprint at the N-terminal coding section. Some of the N-terminal RNA hairpins were also located at wide distances from start codon.

By combining our non-equilibrium Markov model with a Genetic Algorithm optimization, we identified the 12 unknown variables (all based on Kinfold timescales) that
minimize the error between model predictions and experimental data (Supplementary_Data_8).

Overall, the non-equilibrium Markov model of translation initiation can predict the steady-state protein production rates within 4.1-fold error from measurements ($R^2 = 0.76$, $p = 7.3 \times 10^{-38}$) (Figure 8-9A), which is more accurate than the equilibrium-model of translation initiation (5.1-fold error, $R^2 = 0.69$, $p = 6.7 \times 10^{-32}$) (Figure 8-9B).

When the initiation region of mRNA sequence is unstructured, both the equilibrium and non-equilibrium models accurately predict the mRNA’s translation rate ($R^2 = 0.87$, $p = 7.3 \times 10^{-4}$, error = 2.7-fold and $R^2 = 0.89$, $p = 4.7 \times 10^{-4}$, error = 3.2-fold, for equilibrium and non-equilibrium models, respectively) (Figure 8-9CD). However, when RNA hairpins sequester the SD region, the non-equilibrium model more accurately predicts the translation rate ($R^2 = 0.43$, $p = 7.6 \times 10^{-5}$, error = 7.7-fold and $R^2 = 0.71$, $p = 6.1 \times 10^{-9}$, error = 4.1-fold for equilibrium and non-equilibrium models, respectively) (Figure 8-9EF). In particular, the effect of fast-folding versus slow-folding RNA hairpins on translation rate is more accurately predicted by non-equilibrium model. Similar results can be found when N-terminal RNA hairpins overlap with 30S ribosome footprint, where non-equilibrium model more accurately predicts the translation rate ($R^2 = 0.60$, $p = 6.0 \times 10^{-12}$, error = 5.0-fold and $R^2 = 0.69$, $p = 7.7 \times 10^{-15}$, error = 4.1-fold, for equilibrium and non-equilibrium models, respectively) (Figure 8-9GH). In addition, in this group of mRNA sequences, the non-equilibrium model well predicted the effect of both mRNA degradation and RNA folding kinetics on protein production rate (Figure 8-9G).
Figure 8-9. Non-equilibrium Markov model accurately predicts the rate of translation initiation. (A and B) Model predictions for all 119 mRNA sequences characterized in this chapter and Chapter 5 and 7 are compared to their measured fluorescence levels. Translation initiation in equilibrium model was calculated according to formula: $910 \times \exp(-0.45\Delta G_{\text{total}})$. (C and D) Model predictions versus
measurements for 8 mRNA sequences with unstructured initiation region (no-hairpin control mRNAs) (E and F) Model predictions versus measurements for 30 mRNA sequences with RNA hairpins sequestering SD and start codon, where RNA hairpins have diverse folding kinetics. (G and H) Model predictions versus measurements for 52 mRNA sequences with RNA hairpins located immediately after start codon (at position +4), inside the N-terminal coding section, and inhibit the loading process. RNA hairpins have diverse folding kinetics.

Further analysis of our identified parameters shows that structured mRNAs degrade faster than unstructured mRNAs ($\delta_U < \delta_F$ and $\delta_{30S-U} < \delta_{30S-F}$). The Boltzmann constant, $\beta$, is 0.41 mol/kcal, which is close to the previously measured value using the equilibrium model (0.45 mol/kcal) (21). When the mRNA is unstructured, mRNA loading takes place faster than mRNA-ribosome hybridization, suggesting that the loading of an unstructured mRNA is not rate-limiting, until $\Delta G_{\text{complex}}$ becomes more negative than -4.24 kcal/mol. Finally, 70S ribosome’s translocation step occurs much faster than 30S ribosome binding or mRNA loading, suggesting that elongation is not rate limiting step.

8.4. Discussion and conclusion

In this chapter, we designed 52 synthetic mRNA sequences to demonstrate whether Ribosome Drafting can control translation initiation for mRNAs with N-terminal RNA hairpins. Using an optimization algorithm coupled with Kinfold program, we carefully designed pairs of RNA hairpins with equivalent folding energies but disparate folding kinetics, and placed them immediately after start codon on mRFP1 fluorescent protein reporter (Figure 8-1). The measured steady-state fluorescence levels revealed that Ribosome Drafting mechanism can increase the translation rate of mRNAs with slow-folding RNA hairpins by up to 36-fold with respect to their
fast-folding RNA hairpin counterparts. Although, similar to SD-sequestering RNA hairpins, thermodynamics of N-terminal RNA hairpins was shown to control the impact of Ribosome Drafting on translation rate (Figure 8-3), this impact was much lower than SD-sequestering RNA hairpins (with up to 1110-fold increase in translation rate) (Figure 7-11). We argue that because translation initiation has two consecutive steps, mRNA-ribosome hybridization and mRNA loading, RNA hairpins that fold inside the coding section have more free time to refold before the loading process begins. This additional waiting time reduces the extent of non-equilibrium RNA folding, and consequently lowers the impact of Ribosome Drafting (Figure 8-4). We also observed a reduced fluorescence level for mRNAs with N-terminal RNA hairpins (Figure 8-3), which we relate them to their lower mRNA stability.

Another interesting observation was that mRNAs with fast-binding 30S ribosomes ($\Delta G_{\text{complex}} = -8.3 \text{ kcal/mol}$) and weakly stable N-terminal RNA hairpins ($\Delta G_{\text{refold}} = -4.6 \text{ kcal/mol}$) showed similar or higher fluorescence levels than the no-hairpin control mRNAs, as if the RNA hairpins did not form at all (Figure 8-3). We hypothesize that a fast-binding 30S ribosome gains a large momentum after hybridizing to SD, which can then unfold a weakly stable RNA hairpin inside the coding section without expending any unfolding energy penalty. However, further investigation is required to elucidate this mechanism.

We next designed specific RNA hairpins that exhibit two folding transitions (both a fast and a slow transition) with the purpose of studying how Ribosome Drafting controls their effect on translation rate. We found that these RNA hairpins behaved like fast-folding RNA hairpins, which were not predicted by our previously developed Ribosome Drafting model that assumes a single folding transition for RNA hairpins (Figure 8-5). These findings motivated us to develop a general model that can correctly predict RNA folding kinetics, independent of folding transitions.

We, therefore, developed an in silico approach to predict the RNA folding kinetics directly from DNA sequence. In this approach, we first calculate the folding time distribution and
the folding free energy landscape of a given RNA hairpin using 1000 Kinfold trajectories (Figure 8-6A). We then fit these Kinfold predictions to a simple coarse-grained RNA folding model consisting of four folding states: an unfolded RNA, two intermediate states with unknown structures and energies, and a fully folded RNA, all connected using six transition rates (Figure 8-6B). Finally, using an optimization algorithm, we parameterize all six transition rates and two unknown energy contents for intermediate states by minimizing the relative error between model predictions and Kinfold outputs (Figure 8-6C). Importantly, these transition rates all have Kinfold time units.

Having the four-state model of RNA folding, we then developed an integrated non-equilibrium model of translation initiation, which incorporates six molecular interactions that co-dependently function during initiation step: mRNA refolding, mRNA-ribosome hybridization, mRNA loading, ribosome elongation, mRNA degradation, and mRNA transcription (Figure 8-7). Importantly, since our designed RNA hairpins cover a wide range of codons at the N-terminal coding section (between 3 and 16 codons), the effect of ribosome elongation can be significant. A slow ribosome elongation can largely delay the formation of long RNA hairpins but not a short one. We enumerated all possible states that a 70S ribosome can be elongating on mRNA while a new 30S ribosome initiates the next cycle of translation. By incorporating additional states of mRNA refolding, mRNA-ribosome hybridization, and mRNA loading, we developed a Markov model of non-equilibrium translation initiation, consisting of 443 distinct states, connected using 29 transition rates (Figure 8-8). All these transition rates will have Kinfold time units. 12 of these transitions are already determined from our four-state model of RNA folding, leaving 17 transitions rates that are calculated using 12 unknown variables. We, therefore, determined these 12 variables using a Genetic Algorithm optimization that minimizes the relative error between model predicted protein production rate and the measured fluorescence levels, at steady-state, across 119 characterized mRNA sequences (Supplementary_Data_8).
Overall, the non-equilibrium model accurately predicted the translation initiation rate of 119 mRNA sequences with 4.1-fold average error (Figure 8-9), which was more accurate than the equilibrium model (5.1-fold average error). In contrast to the equilibrium model, the non-equilibrium model was able to predict the effect of fast-folding versus slow-folding RNA hairpins on translation rate (Figures 8-9E and 8-9G). The fitted variables that satisfied the model predictions demonstrated that the degradation rate of mRNAs with inhibitory structure is faster than mRNAs with unstructured initiation region. The loading of an unstructured mRNA is faster than mRNA-ribosome hybridization, but the ribosome translocation step is shown to be faster than mRNA loading. However, the ribosome translocation rate largely depends on the protein coding section. For example, all our mRNA sequences have mRFP1 fluorescent protein with a highly optimized coding section, explaining why the predicted ribosome translocation rate is fast. Changing the protein coding section will alter the ribosome translocation rate, but not mRNA loading or mRNA degradation rates. Similarly, by changing promoter DNA sequence, the mRNA transcription rate will change too.

The error in the non-equilibrium model predictions may be originated from different sources. First, the four-state model of RNA folding is parameterized based on the average of 1000 Kinfold trajectories. Therefore, any large deviation in Kinfold outputs can propagate into the non-equilibrium model of translation initiation, causing error in predicted protein production rates. Also the non-equilibrium model of translation initiation assumes that RNA folding starts only after 70S ribosome clears the entire sequence of RNA hairpin, which may not be true because RNA refolding can start even if only a portion of sequence is available. The rate of mRNA transcription is assumed to be constant, which can be further modified to include the co-transcriptional folding or co-transcriptional translation mechanisms into our non-equilibrium model of translation initiation. In this model, we only assume two steps for mRNA-ribosome binding. However, addition of other kinetics-driven interactions during initiation step, for
example the release of initiation factors or detachment of SD sequence from antiSD sequence (17, 201), can improve the model accuracy. Finally, incorporation of sequence-dependent mRNA degradation and ribosome translocation rates should further improve the accuracy of nonequilibrium model. Overall, our developed non-equilibrium model was the first step in building an integrated model of translation rate directly from DNA sequence. The findings in this work can also be used to study how different molecular interactions can cooperatively control protein synthesis rate under non-equilibrium conditions.
References


Appendix A

Supplementary Information for Chapter 2

A.1. Supplementary Figures

Figure A-1. Translation initiation begins with ribosomal platform’s binding to structured 5’ UTRs.

(A) The 16S ribosomal RNA of 30S subunit from wild type E. coli is shown with its three compartments. Six small ribosomal proteins create the ribosomal platform surface to bind to mRNAs. mRNA (black color)
passes across this surface and enters the mRNA channel of 16S rRNA (PDB entry 3J0U). (B) Total Gibbs free energy change during translation initiation, $\Delta G_{\text{total}}$, controls the translation initiation rate. (C) Previous model (21) inaccurately calculated $\Delta G_{\text{total}}$ of 5′ UTRs with structured standby sites. Measured fluorescence (light yellow bar) is compared to predicted translation initiation rate (blue bar). RBS-91 and RBS-44 are from Ref. (21).

**Figure A-2. Effect of structured 5′ UTR on mRNA stability and translation initiation rate.** (A and B) Three structured 5′ UTRs with constant distal binding site length and hairpin height ($D = 6$ nt and $H = 15$ nt), but different proximal binding site length ($P = 4$, $12$, and $24$ nt) were used to examine the effect of proximal binding site length on mRNA instability and translation initiation rate. The fluorescence fold reduction (dark red bar), with respect to an unstructured 5′ UTR, is compared to the mRNA level fold reduction (gradient light blue bar). (C and D) Three structured 5′ UTRs with constant proximal binding site
length and hairpin height (P = 4 nt and H = 15 nt) but different distal binding site length (D = 5, 12, and 20 nt) were used to test the effect of distal binding site length on mRNA instability and translation initiation rate. (D) The fluorescence fold reduction (dark red bar) is compared to the mRNA level fold reduction (gradient light blue bar). Fluorescence data are the result of three measurements in one day and mRNA level data are the result of two measurements in separate days.

Figure A-3. Biophysical model predicts the ribosomal platform’s binding free energy penalty to branched structured 5’ UTRs. (A and B) A standby site module with a two-branched hairpin is shown. The two-branched hairpin can bend over adjacent single stranded region in three different orientations. The available RNA surface area (shaded) depends on the hairpin’s bending orientation. (C and D) Schematic representation of a two-branched structure is shown. The effective hairpin height in each orientation (H₁, H₂, H₃) is depicted. A hairpin stem on the multi-branched loop is counted as an extra nucleotide (black dot). (E and F) Predicted ΔG_distortion using maximum hairpin heights (H = 15 nt) of five two-branched structures and one three-branched structure (Supplementary_Data_2) are compared to predictions using average
branched hairpin height (average errors are 1.40 and 0.38 kcal/mol, respectively; p-value is 0.017). In parts EF, ΔG_{distortion} numbers shown in secondary y-axis are calculated using Equation 2.3 in Chapter 2, and are directly related to the data according to Equation 2.4. The diagonal dashed line is also the predicted translation rate according to Equation 2.4, where ΔΔG_{other} is zero kcal/mol.

Figure A-4. Ribosomal platform unfolds mRNA structures based on hairpin sequence and energy. (A and C) The hypothetical hairpin with A:U rich closing base pairs shows a moderate slope for unfolding energy landscape. Two closing base pairs (bolded lower case) are predicted to be unfolded by the ribosomal platform in order to reach the minimum binding free energy penalty (5.11 kcal/mol). (B and D) The hypothetical hairpin with G:C rich closing base pairs has a steep unfolding energy landscape. No closing base pair is predicted to be unfolded. The fully folded state (A₅ = 6) has the minimum binding free energy penalty (8.96 kcal/mol). ΔG_{distortion} is dashed red line, ΔG_{unfolding} is dotted-dashed blue line, and ΔG_{standby} is solid green line.
Figure A-5. Examples of synthetic structured 5’ UTRs designed to characterize the interactions between ribosomal platform and structured 5’ UTR. (A) Designed structured 5’ UTRs with diverse sequences demonstrate the coupled trade-offs between ribosomal distortion and RNA unfolding. The predicted unpaired bases are shown in bold lower cases. (B) Synthetic 5’ UTRs with diverse structures were designed to examine the ability of biophysical model in predicting the ribosomal platform’s binding free energy penalty to the structured mRNAs. Sequence, size, and folding energy of all mRNA structures are listed in Supplementary_Data_2.
Figure A-6. Analysis of distribution of error in biophysical model predictions. (A) The probability plot for normal distribution of errors in predicted ΔG_{standby} (Anderson-Darling test with 95% confidence interval) shows that the error is normally distributed across 90% of the designed 5’ UTRs. However, 13 designed 5’ UTRs show deviation from normality, indicating some systematic errors in model prediction. Six of these designed 5’ UTRs are under-predicted by the biophysical model (shown by red color background in the Supplementary_Data_2), which are mostly affected by mRNA instability (See Figure A-2). Others are over-predicted by the biophysical model (shown by blue color background in the Supplementary_Data_2) (see also Discussion and conclusion section in Chapter 2). The red-dashed line joins the first and third quartiles of sample population. (B) A histogram shows that the biophysical model is capable of predicting ΔG_{standby} for most of designed 5’ UTRs. ΔΔG_{standby} = measured ΔG_{standby} - predicted ΔG_{standby}. 
A.2. Supplementary notes

A.2.1. Ribosomal platform interactions with structured standby sites control translation initiation rate

The ribosomal platform, which makes contact with structured mRNAs at its surface, is one of the three compartments of 16S ribosomal RNA, a scaffold for 30S ribosomal subunit. Six small ribosomal proteins at the platform surface create positively charged surface and bind non-specifically to the negatively charged phosphate backbone of mRNA via electrostatic interactions. In addition, the anti-SD of 16S ribosomal RNA, which is located on the platform surface, stabilizes the mRNA through hydrogen binding (Figure A-1A). Although about 30 nt of mRNA around the start codon has to be unfolded and fed into the mRNA channel to position the start codon at the P-site (22), the upstream 5’ UTR can remain folded and in contact with the ribosomal platform surface. The efficient binding of structured mRNAs to the ribosomal platform is a prerequisite for their translation. Therefore, understanding the interactions between ribosomal platform and structured mRNAs is necessary to accurately control their translation rate.

The previously developed biophysical model of translation initiation, RBS Calculator v1.0 (21), accounts for five molecular interactions between mRNA and 30S ribosomal subunit. The total Gibbs free energy change during translation initiation, \( \Delta G_{\text{total}} \), is the quantitative measure of its rate (Figure A-1B). The model accurately predicted the translation initiation rates of 132 mRNAs with different Shine-Dalgarno sequences, inhibitory mRNA structures, and 5’ coding sequences. However, v1.0 of the biophysical model was not able to accurately calculate the ribosome's binding free energy \( \Delta G_{\text{total}} \) for mRNAs that contain structures upstream of the Shine-Dalgarno sequence in the After-bound state. The predicted translation initiation rate of an example, RBS-91, was 26.25-fold greater than its measured fluorescence level (3387.86 au vs. 0 au).
129.08±39.93 au on the same proportional scale), equivalent to neglecting a 7.26 kcal/mol binding free energy penalty (Figure A-1C). In contrast, the previous model can accurately predict the translation initiation rates of mRNAs where mRNA structures overlap with the Shine-Dalgarno sequence, but they are primarily unstructured in the After-bound state (RBS-44 example shown in Figure A-1C).

The v1.0 of the biophysical model assumed that a small portion of standby site, which was defined as a 4 nt region upstream of the 16S rRNA binding, must be unstructured and that the energy to unfold this region would lower the translation initiation rate. However, based on this and other data, it was clear that this assumption was false, and that the interactions controlling the ribosomal platform's ability to bind standby sites were not correctly modeled. Therefore, it was important to develop a new biophysical model to accurately calculate the ribosomal platform's interactions with standby sites, denoted by $\Delta G_{\text{standby}}$.

A.2.2. The relationship between ribosomal platform binding and mRNA stability

We measured the fluorescent protein expression levels and mRNA levels from structured mRNAs with standby site modules containing either short, medium, and long distal or proximal binding site lengths, and compared these measurements to those from an unstructured mRNA (see Materials and methods section in Chapter 2). In all cases, the structured mRNAs contain a hairpin with a 15 nt height. We observed a 9.6-fold reduction in fluorescence for a short proximal binding site ($P = 4$ nt), compared to 5.4-fold reduction in mRNA level (Figure A-2B). Similarly, we observed a 10.1-fold reduction in fluorescence for a short distal binding site ($D = 5$ nt), compared to a 3.3-fold reduction in mRNA level (Figure A-2D). In contrast, fluorescence levels of a medium length proximal binding site ($P = 12$ nt) or a medium length distal binding site ($D = 12$ nt) were 1.3-fold and 2.5-fold lower than an unstructured mRNA, compared to a 2.4-fold and 1.3-
fold reduction in mRNA level, respectively. From this data, it appears that the hairpin itself does not significantly decrease the mRNA levels by destabilization. When shortening the proximal or distal binding site lengths, the fluorescent protein expression levels decrease more than the mRNA levels, showing that an inaccessible standby site module is primarily affecting translation rate. Though, there is likely a small amount of coupling between decreases in translation rate and decreases in mRNA stability, due to the protective effects of ribosomes actively elongating on a mRNA to prevent the binding of RNAses.

Long proximal (P = 24 nt) and distal (D = 20 nt) binding sites caused further reduction in fluorescence level (3.3- and 4.2- folds), comparable to the reduction in their mRNA levels (5.8- and 2.7- folds, respectively) (Figures A-2B and A-2D). These reductions in fluorescence level and mRNA levels cannot be explained by the proposed biophysical model of the ribosomal platform's interactions with standby sites. Instead, these changes are likely due to increased RNAsse binding to the long, unstructured proximal or distal binding sites, causing increased mRNA degradation and lower mRNA levels. The error in the biophysical model predictions is larger when long proximal or distal binding sites (greater than 15 nt) are utilized in mRNAs, due to the confounding effects of increased mRNA degradation (see Figure A-6).

A.2.3. Branched structured 5’ UTRs control the ribosomal platform’s binding according to their effective hairpin height

Branched mRNA structures can control the ribosomal platform’s distortion energy penalty differently than the normal mRNA structures. Because of the three dimensional orientations of the hairpin stems, we propose that a two-branched structure (Figure A-3A) can bend over the neighboring proximal or distal single stranded RNA regions in three different ways
It can bend over the single stranded RNA using its base hairpin stem together with one of its branched hairpin stems (stem I or II) (top and bottom configurations in Figure A-3B) or only by using the base hairpin stem (middle configuration in Figure A-3B). At each configuration, a different length of single stranded RNA is sequestered, resulting in different available RNA surface areas ($A_S$). We assume that the probability of each configuration is equal. Therefore, the effective $A_S$ for a two-branched structure is the average of three available RNA surface areas, according to formula: $A_S = 15 + P + D - H_{avg}$, where the average branched hairpin height, $H_{avg}$, is calculated as $H_{avg} = \frac{H_1 + H_2 + H_3}{3}$.

In order to calculate $H_1$, $H_2$, $H_3$, the branched hairpin stems (stems I or II) are counted as an extra nucleotide on the multi-branched loop (Figure A-3CD). In general, because $H_2$ is less than $H_1$ and $H_3$, the average branched hairpin height is shorter than the maximum hairpin height. This leads to a larger available RNA surface area and therefore lower ribosomal platform’s distortion energy penalty. Importantly, this proposed formulation of average branched hairpin height can be expanded to higher number of branched stems. For example, the average branched hairpin height of a three-branched structure is calculated as $H_{avg} = \frac{H_1 + H_2 + H_3 + H_4}{4}$, where calculation of individual hairpin heights follows the same procedure shown in Figure A-3CD.

We critically tested this proposed formulation, by predicting $\Delta G_{distortion}$ of five two-branched and one three-branched structures, with and without average branched hairpin height calculation. The average branched hairpin height calculation led to an improvement in accuracy of predicted $\Delta G_{distortion}$ (average error is 0.38 kcal/mol with average branched hairpin height compared to 1.40 kcal/mol with maximum hairpin height ($H = 15$ nt); p-value is 0.017 (Figure A-3EF).
Appendix B

Supplementary Information for Chapter 6

B.1. Supplementary Figures

**Figure B-1. Thermodynamic equilibrium in a simplified intracellular environment.** Cartoon shows a simplified cellular environment and the thermodynamic equilibrium between free mRNA (state 1 in Figure 6-1B), mRNA-ligand complex (state 3 in Figure 6-1B), and free ligand. Translation is initiated from free mRNA (state 1) and ligand-bound mRNA (state 3) but with different rates. The mRNA at both state 1 and state 3 has similar volume. The Gibbs free energy of each complex is shown. $\Delta \Delta G_{\text{mRNA}}$ and $\Delta G_{\text{ligand}}$ are calculated from Equations B.8 and B.9.
Figure B-2. Schematic of optimization algorithm. The inputs to the optimization algorithm are the aptamer’s sequence and structure and the protein coding sequence. Additional constraints such as restriction site sequences and positions can be defined by the user. The optimization starts with the randomized pre- and post-aptamer sequences to generate an initial pool of riboswitches. Next, riboswitches will be selected for random mutations based on predefined probabilities, followed by fitness evaluation using the biophysical model calculations. Only the riboswitches that meet the objective function requirements will be saved for further recombination step. The objective functions can be maximizing $AR_{\text{actual}}$, minimizing $\Delta G_{\text{mRNA}}$, or reaching targeted translation rates in OFF and ON states. Finally, the selected riboswitches will be sent back to the population for the next round of mutations. At the end of each cycle, a few riboswitches with the highest fitness values are saved as the elite members, and the cycles are repeated until the optimization converges to the optimum elite solutions.
Figure B.3. The biophysical model can predict the output of a combinatorial library. (A) Lynch et al. 2007 (63) generated four libraries with different sizes, in which the linker region between the theophylline binding aptamer and the SD region was randomized. The libraries were then screened using high-throughput methods (see supplementary notes in Appendix B.2.5 for more details). (B) According to our biophysical model calculations, the probability of finding a riboswitch with large predicted AR$_{max}$ dropped significantly across the entire library variants as the library size increased. (C) The number of experimentally identified riboswitches for each library is compared with the number of potential riboswitches that have large predicted maximum activation ratios (AR$_{max}$ > 10). (D) The in vivo $\beta$-galactosidase levels (Miller units) of 20 identified riboswitches from all four libraries (63) are correlated well with the predicted $\Delta G_{total}$ at both OFF state (white circles) and ON state (orange circle). Pearson R$^2$ = 0.64 and p-value = $7 \times 10^{-10}$. (E) The measured activation ratios of most of the 20 identified riboswitches are close to the predicted maximum activation ratio (AR$_{max}$). The dashed line is the $x=y$ line.
Figure B-4. The switching free energy of 52 designed riboswitches. The switching free energy ($\Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}$) of 52 designed riboswitches using theophylline, fluoride, and TMR aptamers are shown. Light blue bars correspond to metastable riboswitches and the light green bars are the stable riboswitches. The red asterisks indicate the two outliers in Figures 6-3CD (Theo-42 and Theo-44). The riboswitches are shown according to their order in the Supplementary_Data_6.
Figure B-5. The biophysical model generates riboswitches with diverse sequences. (A) The diversity of pre- and post-aptamer regions for 52 designed riboswitches using theophylline, fluoride, and TMR aptamers is shown. The plots are created by the WebLogo software (210). The last three nucleotides of the post-aptamer region is the ATG start codon. The expected SD region is shown. (B, C, and D) The diversity of pre- and post-aptamer for individual riboswitches with theophylline, TMR, and fluoride aptamers is shown. The first five nucleotides of aptamer are added to the end of pre-aptamer region and the last five nucleotides of aptamer are added to the beginning of post-aptamer. The last three nucleotide of post-aptamer is the ATG start codon. The expected SD region is also depicted. We also calculated the sequence entropy to be 63%, using the pre- and post-aptamer regions of all these 52 designed riboswitches according to formula:
where $p$ is the number of appearance of nucleotides A, T, C, or G in each position $i$ and $N$ is the total number of nucleotide positions.

![Diagram](image.png)

**Figure B-6. Secondary structure of a fluoride riboswitch in the OFF and ON states.** The predicted secondary structures of a designed Fluoride riboswitch (F-8) are depicted for both OFF and ON (constrained) states. (A) In the absence of ligand (OFF state), the free riboswitch mRNA folds with $\Delta G_{\text{mRNA,OFF}} = -46.1$, while ribosome binds to the mRNA with following energies: $\Delta G_{\text{mRNA-rRNA,OFF}} = -42.7$, $\Delta G_{\text{start,OFF}} = -1.2$, $\Delta G_{\text{spacing,OFF}} = 0.7$ (spacing = 7 nt), and $\Delta G_{\text{standby,OFF}} = 5.4$. Overall, the net energy change from the ribosome binding is $\Delta G_{\text{total,OFF}} = 8.3$. All the Gibbs free energies are in kcal/mol. (B) In the
presence of ligand (ON state), the mRNA folds to a constrained pseudoknotted structure, leading to different Gibbs free energies between ligand-bound mRNA and ribosome: $\Delta G_{mRNA,ON} = -29.9$, $\Delta G_{mRNA,rRNA,ON} = -36.6$, $\Delta G_{start,ON} = -1.2$, $\Delta G_{spacing,ON} = 0.7$ (spacing = 7 nt), and $\Delta G_{standby,ON} = 0$ (see supplementary notes in Appendix B.2.4). This increases the net energy release from the ribosome binding to mRNA ($\Delta G_{total,ON} = -7.24$), equivalent of 1090-fold maximum activation in translation initiation rate ($A R_{\text{max}} = \exp(-0.45 \times (-7.24 - 8.3)) = 1090$). However, largely positive $\Delta \Delta G_{mRNA} = 16.2$ ($\Delta \Delta G_{mRNA} + \Delta G_{ligand} = 10.2$) lowers the actual activation to 1.5-fold due to metastable binding of ligand to mRNA.

![Image](image.png)

Figure B-7. Secondary structure of a TMR riboswitch in the OFF and ON states. The predicted secondary structures of designed TMR-7 riboswitch are depicted for both OFF and ON (constrained) states.
(A) The calculated Gibbs free energies in the OFF state (no ligand) are: \( \Delta G_{\text{mRNA,OFF}} = -52.5 \), \( \Delta G_{\text{mRNA-rRNA,OFF}} = -53.9 \), \( \Delta G_{\text{start,OFF}} = -1.2 \), \( \Delta G_{\text{spacing,OFF}} = 0 \) (spacing = 5 nt), and \( \Delta G_{\text{standby,OFF}} = 7.8 \). Therefore, the total Gibbs free energy change between free mRNA and ribosome-bound mRNA is \( \Delta G_{\text{total,OFF}} = 5.2 \). All the Gibbs free energies are in kcal/mol. (B) Presence of ligand in the ON state changes the Gibbs free energies between ligand-bound mRNA and ribosome according to: \( \Delta G_{\text{mRNA,ON}} = -46.3 \), \( \Delta G_{\text{mRNA-rRNA,ON}} = -53.7 \), \( \Delta G_{\text{start,ON}} = -1.2 \), \( \Delta G_{\text{spacing,ON}} = 0 \) (spacing = 5 nt), and \( \Delta G_{\text{standby,ON}} = 0.2 \), which leads to much stronger ribosome binding rate (\( \Delta G_{\text{total,ON}} = -8.4 \)). Therefore, the maximum activation ratio is \( AR_{\text{max}} = \exp(-0.45 \times (-8.4-5.2)) = 456 \)-fold. Although the binding of ligand to mRNA is stable (\( \Delta \Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}} = -4.0 \)), the TMR toxicity limits the amount of ligand availability, and therefore reduces the actual activation ratio to \( AR_{\text{actual}} = 6.5 \)-fold.

Figure B-8. The designed riboswitches showed continuous activation during growth. (A) The fluorescence per cell (FLPC, Equation 6.1 in Materials and methods section in Chapter 6), measured per time during the serial dilution experiment, is shown for Theo-44 theophylline riboswitch (top) and a no-
aptamer control (bottom) upon addition of 2 mM Theophylline. (B) The FLPC values for the F-4 fluoride riboswitch (top) and a no-aptamer control (bottom) during the serial dilution experiment and at the presence of 150 mM NaF is shown. The first and second dilutions are marked. The shaded regions are the standard deviation for 2 separate measurements. Averages are the bold lines.

Figure B-9. Predicted AR$_{max}$ versus measured AR. The predicted maximum activation ratios (AR$_{max}$) are shown against the measured AR for 67 riboswitches using theophylline, fluoride, and TMR binding aptamers. The color representations and ligand concentrations are similar to those in Figures 6-3CD. Pearson R$^2$ = 0.002 and p-value = 0.71, N = 67. The dashed line is x=y line.
Figure B-10. Predicted AR_{conc} versus measured AR for stable and metastable riboswitches. The predicted concentration-limited activation ratios (AR_{conc}) are shown against the measured AR for 67 riboswitches using theophylline, fluoride, and TMR binding aptamers. The riboswitches are divided into (A) stable riboswitches (with ΔΔG_mRNA + ΔG_{ligand} ≤ 0) and (B) metastable riboswitches (with ΔΔG_mRNA + ΔG_{ligand} > 0). The color representations and ligand concentrations are similar to those in Figure 6-3CD.

The dashed lines are x=y line. (A) When riboswitches are stable, the model calculated AR_{conc} is sufficient to predict their function. Spearman R = 0.72, p-value = 2.5x10^{-5}, N = 27; Pearson R^2 = 0.62, p-value = 3.5x10^{-6}, N = 25. The two outliers are Theo-15 (green circle) and Theo-42 (brown circle). (B) When the riboswitches have metastable ligand binding, their measured AR largely deviates from the predicted AR_{conc}. Spearman R = 0.27, p-value = 0.09, N = 40; Pearson R^2 = 0.04, p-value = 0.22, N = 40.
Figure B-11. Error distribution analysis for model predictions. (top row) The probability plot for normal distribution of errors in predicted translation initiation rates (OFF and ON) and $A_{R_{\text{actual}}}$ showed that the errors in $A_{R_{\text{actual}}}$ are normally distributed, while the errors in individual translation initiation rates are less normal. Overall, there are 8 outliers out of 67 riboswitches that their error in $A_{R_{\text{actual}}}$ is not normal; three of them have over-predicted $A_{R_{\text{actual}}}$ and five of them are under-predicted. The red-dashed line joins the first and third quartiles of sample population. (bottom row) The histograms show the distribution of error, the red lines are the corresponding Gaussian distributions. The mean ($\mu_{\log}$) and standard deviation ($\delta_{\log}$) for the log-normal distributions are: 0.14 and 2.07 (for translation initiation rate OFF), 0.09 and 1.74 (for translation initiation rate ON), and -0.05 and 1.27 (for $A_{R_{\text{actual}}}$). Therefore, the mean ($\mu$) and standard deviations ($\delta$) in normal scale are: 1.15 and 7.94 (for translation initiation rate OFF), 1.10 and 5.16 (for translation initiation rate ON), and 0.96 and 2.44 (for $A_{R_{\text{actual}}}$), which are calculated according to formula:

$$\mu = \exp(\mu_{\log}) \quad \text{and} \quad \delta = \exp(\mu_{\log}) \times \left[\exp(\delta_{\log}) - 1\right].$$
Figure B-12. Predicted versus measured translation initiation rates and activation ratios for individual aptamers. (top row) The predicted translation initiation rates for individual aptamers (45 theophylline riboswitches, 12 fluoride riboswitches, and 10 TMR riboswitches) are compared to their measured fluorescence or luminescence levels in both OFF and ON states. All the luminescence levels are shown as one tenth of the measured values (arbitrary units). The color code is similar to Figure 6-3E. The dashed line is x=y line. Spearman R = 0.65, p-value = 3.2x10^{-12}, N = 90; Pearson R^2 = 0.40, p-value = 3.0x10^{-11}, N = 90 (theophylline riboswitches). Spearman R = 0.21, p-value = 0.32, N = 24; Pearson R^2 = 0.09, p-value = 0.15, N = 24 (fluoride riboswitches). Spearman R = 0.66, p-value = 0.002, N = 20; Pearson R^2 = 0.43, p-value = 0.002, N = 20 (TMR riboswitches). (bottom row) The predicted AR_{actual} is compared to the measured AR for individual aptamers. The black and yellow circles in theophylline plot are the over-predicted outliers (Theo-41, Theo-42, and Theo-44) and under-predicted outliers (Theo-13, Theo-14, Theo-15, Theo-26, and Theo-27), respectively. The dashed line is x=y line. Spearman R = 0.39, p-value = 0.009, N = 45; Pearson R^2 = 0.55, p-value = 1.8x10^{-7}, N = 37 (theophylline riboswitches). Spearman R = -0.18, p-
value = 0.57, N = 12; Pearson $R^2 = 0.009$, p-value = 0.77, N = 12 (fluoride riboswitches). Spearman R = 0.53, p-value = 0.12, N = 10; Pearson $R^2 = 0.62$, p-value = 0.007, N = 10 (TMR riboswitches).

**Figure B-13. Correlation between errors in predicted translation initiation rate ON and translation initiation rate OFF.** Error in predicted translation initiation rate (OFF) is highly correlated with the error in the predicted translation initiation rate (ON) for 88% of the riboswitches. Spearman R = 0.78, p-value $< 10^{-30}$, N = 67; Pearson $R^2 = 0.87$, p-value = $3.6 \times 10^{-27}$, N = 59 (without 8 outliers). The color code is similar to **Figure 6-3CD**. The diagonal dashed line is x=y line. The blue and green circles in the top-left box are the under-predicted theophylline riboswitches with measured AR $> 150$-fold (Theo-13, Theo-14, Theo-15, Theo-26, and Theo-27). All these five riboswitches have lower OFF expression than their predicted translation initiation rate OFF. The three brown circles in the bottom-right box are theophylline riboswitches that are over-predicted by the model (Theo41, Theo-42, and Theo-44). All these three riboswitches have higher OFF expression than their predicted translation initiation rate OFF.
Figure B-14. **Model errors versus individual Gibbs free energy terms.** (top row) Errors in predicted activation ratios are compared to the difference in individual Gibbs free energy terms between OFF and ON states as calculated by the biophysical model for all 67 riboswitches: $\Delta \Delta G_X = \Delta G_{X,\text{ON}} - \Delta G_{X,\text{OFF}}$. These plots show that $\Delta G_{\text{mRNA}}$, $\Delta G_{\text{mRNA-rRNA}}$, and $\Delta G_{\text{standby}}$ are changing greatly between OFF and ON states, although they don’t show any correlation with the error in activation ratio. The three circles shown in the boxes are the over-predicted theophylline riboswitches (Theo-41, Theo-42, and Theo-44). The orange highlighted box shows that these three outliers have large positive $\Delta \Delta G_{\text{mRNA}}$, which could be responsible for their lower-than-predicted ON expression. (three bottom rows) The first row shows the errors in predicted activation ratios against the calculated Gibbs free energy terms in the OFF state (blue circles) and ON state (red circles). The boxes indicate the three over-predicted theophylline riboswitches (Theo-41, Theo-42, and Theo-44). The second row shows the errors in predicted TIR_{OFF} (OFF state translation initiation rate) with respect to the calculated Gibbs free energy terms in the OFF state. The third row shows the errors in predicted TIR_{ON} (ON state translation initiation rate) with respect to the Gibbs free energy
terms in the ON state. Overall, there is no significant correlation between individual Gibbs free energy terms and error in AR_{actual}, TIR_{OFF}, or TIR_{ON}. However, the highlighted orange box in ΔG_{SD-antiSD} plot shows that a largely negative ΔG_{SD-antiSD} could be responsible for the higher-than-predicted OFF expression for the three outlier riboswitches.

Figure B-15. The measured activities of all 62 designed riboswitches. (A) in vivo expression levels at (white circles) zero and (orange circles) 2 mM theophylline are shown for 24 theophylline riboswitches that control luciferase expression by using either TCT8-4 aptamer (left section) or its mutated version (right section). Activation ratios are shown (gray bars). These riboswitches are transcribed using a Tac promoter
on a pUC19 vector. Riboswitches are (M) metastable if their switching free energy exceeds 0 kcal/mol. Otherwise, their ligand binding is stable (S). (B) in vivo fluorescence levels at (white circles) zero and (orange circles) 2 mM theophylline are shown for 6 theophylline riboswitches (with mutated aptamer), controlling mRFP1 expression. These riboswitches are transcribed using a J23100 promoter on a pFTV1 vector. A designed control theophylline riboswitch with minimal predicted activation is shown in black bar. The red asterisks indicate outliers in Figures 6-3CD (Theo-42 and Theo-44). (C) in vivo fluorescence levels at (white circles) zero and (orange circles) 150 mM fluoride and at (white circles) zero and (orange circles) 1 mM DNT are shown for 12 fluoride riboswitches and 3 DNT riboswitches, both controlling mRFP1 expression. The J23100 and AEB-3 promoters drive the transcription of fluoride and DNT riboswitches on a pFTV1 vector, respectively. (D) cell-free transcription-translation expression levels are shown for 5 dopamine riboswitches and 2 thyroxine (T4) riboswitches, both controlling luciferase expression. These riboswitches are transcribed using a tac promoter on a pFTV1 vector. The activity of dopamine riboswitches were measured under 5 mM ascorbic acid and with (white circles) zero and (orange circles) 1 mM dopamine. The thyroxine riboswitches were characterized under 100 mM NaCl, 1.5 mM NaOH, and with (white circles) zero and (orange circles) 150 uM thyroxine. (E) Expression levels are shown for 10 TMR riboswitches in in vivo and cell-free transcription-translation assays within the (white circles) absence and (orange circles) presence of 20 µM and 30 µM TMR, respectively. These riboswitches control the luciferase expression and are transcribed using a tac promoter on a pUC19 vector. (F) The in vivo measured expression levels of TMR riboswitches were compared to their measurements using cell-free transcription-translation assays, showing proportionality. The top two characterized riboswitch variants for each aptamer are shown in Figure 6-4A. The riboswitch measurements in all the sub-figures are labeled according to their order in the Supplemental_Data_6.
Figure B-16. *Secondary structure of a dopamine binding aptamer.* Secondary structure of a dopamine binding aptamer (dopa2/c.1) that was used to generate three dopamine riboswitches is shown here. A proposed pseudoknot base pairing is depicted. A shorter RNA aptamer (dopa1.3/C.3) was also used to generate two additional dopamine riboswitches and is shown in Figure 6-3B in the Chapter 6.

Figure B-17. *Relative gene expression from different promoters and plasmid copy numbers.* (A) The expression of luciferase on four different plasmids with increasing copy numbers is shown. The values are the average OFF state luminescence from three riboswitches in Figure 6-5C and a no-aptamer control. The numbers are normalized with respect to pUC19. In all four plasmids, tac promoter drives the transcription process. (B) The relative transcription rates of three promoters are shown. The values are the average OFF state fluorescence from five riboswitches in Figure 6-5E. The transcription rates are normalized with respect to the J23100 promoter.
Figure B-18. Comparison of riboswitch growth rate and expression levels at different mRNA levels.

(A) The specific growth rates of empty E. coli DH10B cell, the cells with plasmids harboring the no-aptamer control mRNA, and the cells with theophylline and fluoride riboswitches controlling mRFP1 production are shown. The black bar is the empty cell, the white bars are mRNAs with weak AEB-3 promoter, and grey bars are mRNAs with strong LmrA promoter. Overall, cells grow slower in 150 mM fluoride than 2 mM theophylline regardless of the riboswitch mRNA. The LmrA promoter decreased the cell’s growth rate by only 13% for Theo-44 theophylline riboswitch, while decreasing it by 20% for no-aptamer control mRNA. Similarly, the cell’s growth rate using LmrA promoter was reduced by 48% and 20% for F-6 and F-11 fluoride riboswitches, while the no-aptamer control mRNA reduced the growth rate by 30%. (B) The expression levels of fluoride and theophylline riboswitches together with the no-aptamer control is shown using AEB-3 (white bars) and LmrA (grey bars) promoters. In all cases, the expression
levels of riboswitch mRNAs are much lower than the no-aptamer control mRNAs, providing evidence that cells with riboswitch mRNAs have not reached the maximum translational capacity.

Figure B-19. Accepted and unaccepted types of pseudoknot base pairings in the biophysical model.

(A) A standard pseudoknot base pairing is depicted. The RNA structure will be divided into two canonical base pairings sub-structures (red and blue lines). (B) A non-standard pseudoknot base pairing (green lines) is shown. The RNA structure cannot be divided into separate canonical sub-structures. The energy of this type of pseudoknot base pairing is not calculated by the model.

B.2. Supplementary notes

B.2.1. Biophysical model of riboswitches to control translation initiation

The biophysical model of 30S platform’s binding to the structured 5’ UTR (12) allows us to predict the effect of any structured 5’ UTR on the translation initiation rate. The core component of this model is the ability to predict the translation initiation rate directly from the RNA folding shape and geometry. Importantly, the ligand-bound RNA aptamers are geometrically similar to the RNA structures that appear in the 5’ UTR and theoretically should control the translation initiation rate in the similar fashion. This special feature of the biophysical model helped us to develop the automated design of synthetic riboswitches to control the translation initiation rate directly from the RNA aptamers. Notably the automated design does not
add any additional parameters to the model. The only input for the model is the structural constraints of RNA aptamers after it is bound to the ligand.

The biophysical model of riboswitches assumes a two state process, an “OFF state” and an “ON state”. In the “OFF state” when the ligand is not present inside the cell, mRNA which carries the RNA aptamer sequence in its 5’ UTR, folds into its minimum free energy $\Delta G_{\text{mRNA,OFF}}$. Therefore, the 30S subunit binds to the free mRNA with a total Gibbs free energy change $\Delta G_{\text{total,OFF}}$ (Figure 6-2A) and initiates translation with a rate according to Equation B.1.

$$r_{\text{OFF}} \propto \exp(-\beta \Delta G_{\text{total,OFF}})$$  \hspace{1cm} (B.1)

In the ON state, however, it is assumed that the excess ligand is present inside the cell and all mRNA molecules are stably bound to ligand. Thus, mRNA folds into its free energy state $\Delta G_{\text{mRNA,ON}}$, which is subject to structural constraint of the RNA aptamer. This will pose a constraint on 30S subunit binding to the ligand-bound mRNA, where it should compete with the ligand-bound RNA aptamer for proper assembly on the mRNA. This constrained binding requires new binding free energy $\Delta G_{\text{total,ON}}$ and results in new translation initiation rate, according to Equation B.2. In the ON state, in addition to $\Delta G_{\text{mRNA,ON}}$, the $\Delta G_{\text{mRNA-rRNA,ON}}$, $\Delta G_{\text{spacing,ON}}$, and $\Delta G_{\text{standby,ON}}$ can also be perturbed by the structural constraint of RNA aptamer (Figure 6-2B).

$$r_{\text{ON}} \propto \exp(-\beta \Delta G_{\text{total,ON}})$$  \hspace{1cm} (B.2)

The $r_{\text{ON}}$ calculated from Equation B.2 is the maximum predicted translation initiation rate for the ON state, because it assumes that there are excess number of ligand inside the cell and all the mRNA molecules exist in the stable ligand-bound state. The $in \ vivo$ measured expression levels (at both OFF and ON states) of the previously designed theophylline riboswitches (163) were well-correlated with the predicted Gibbs free energies according to Equations B.1 and B.2 (Figures 6-2CDE), demonstrating the model’s accuracy at the stable switching and excess ligand condition.
However, the riboswitch activity decreases as ligand concentration drops. Here, by employing a thermodynamic minimization approach ([Appendix B.2.2]), we can predict the ON state translation initiation rate, $r_{ON,conc}$, as a function of total mRNA and ligand concentrations according to Equation B.3. This formulation assumes that mRNA and ligand have reached the equilibrium, and their binding is stable.

$$r_{ON,conc} \propto \frac{C_{mRNA,1}}{C_{mRNA, total}} \exp(-\beta \Delta G_{total, OFF}) + \frac{C_{mRNA,3}}{C_{mRNA, total}} \exp(-\beta \Delta G_{total, ON})$$

Equation B.3

Since the ligand and mRNA molecules have very different molecular sizes (V) and occupy different excluded volumes inside the cell, we sought to rewrite $r_{ON,conc}$ as a function of volume fraction of mRNA and ligand inside the cell. Therefore, by multiplying both numerators and denominators in Equation B.3 by $V_{mRNA} / (V_s * C_s)$ we can simplify Equation B.3 to Equation B.4, in which $v_i$ is the volume fraction of species $i$ and is calculated from formula: $v_i = (V_i * C_i) / (V_s * C_s)$. “s” stands for solvent, which is assumed to be water.

$$r_{ON,conc} \propto \frac{V_{mRNA,1}}{V_{mRNA, total}} \exp(-\beta \Delta G_{total, OFF}) + \frac{V_{mRNA,3}}{V_{mRNA, total}} \exp(-\beta \Delta G_{total, ON})$$

Equation B.4

Using the molar conservation of ligand, we can find the following relationship for the volume fraction of mRNA:ligand complex (state 3 in Figure 6-1B) as a function of the specific volume and the volume fraction of total and free ligand molecules (Equation B.5).

$$V_{mRNA,3} = \frac{V_{mRNA}}{V_{ligand}} (V_{ligand, total} - V_{ligand, free})$$

Equation B.5

By substituting Equation B.5 into Equation B.4, we can simplify the $r_{ON,conc}$ calculation to Equation B.6. In this equation, $\Delta G_{total,OFF}$ and $\Delta G_{total, ON}$ are calculated using the biophysical model given the sequence of mRNA and structural constraint of aptamer, while all the volume...
fractions at equilibrium are calculated using the thermodynamic minimization of the box given the specific volume of mRNA and ligand.

\[
r_{\text{ON,conc}} \propto \frac{V_{\text{mRNA,1}}}{V_{\text{mRNA,total}}} \exp(-\beta \Delta G_{\text{total,OFF}}) + \frac{V_{\text{mRNA}}}{V_{\text{ligand}}} \left[ \frac{V_{\text{ligand,free}} - V_{\text{ligand,total}}}{V_{\text{mRNA,total}}} \right] \exp(-\beta \Delta G_{\text{total,ON}})
\]

(B.6)

Finally, the concentration-limited riboswitch’s activation ratio \(AR_{\text{conc}}\) (\(AR_{\text{conc}} = r_{\text{ON,conc}}/r_{\text{OFF}}\)) can be calculated at any given mRNA and ligand concentrations, according to Equation B.7, in which \(AR_{\text{max}}\) is a function of \(\Delta \Delta G_{\text{total}}\), which is the difference between \(\Delta G_{\text{total,OFF}}\) and \(\Delta G_{\text{total,ON}}\) (\(AR_{\text{max}} = \exp(-\beta \Delta \Delta G_{\text{total}})\)). This equation is also presented as Equation 6.2 in Chapter 6. The measured activation ratio of TMR riboswitches at limited ligand concentration (20 uM TMR) were well-correlated with the model predictions according to Equations B.7 (Figure B-10A), whereas the model predictions with excess ligand assumption failed to do so (Figure B-9).

\[
AR_{\text{conc}} = \frac{V_{\text{mRNA,1}}}{V_{\text{mRNA,total}}} + \frac{V_{\text{mRNA}}}{V_{\text{ligand}}} \left[ \frac{V_{\text{ligand,free}} - V_{\text{ligand,total}}}{V_{\text{mRNA,total}}} \right] AR_{\text{max}}
\]

(B.7)

The biophysical model of riboswitch is a thermodynamic-based model that predicts the interactions between ligand, mRNA, and ribosome at steady-state. Currently, it does not consider the kinetics of mRNA structural switching or dynamics of mRNA-ligand binding. Also other factors including the transport of ligand across the cell membrane, its diffusion inside the cytosolic environment, and the ligand’s toxicity effect on the cell growth have not been considered in this model.

The thermodynamic calculations for \(AR_{\text{max}}\) and \(AR_{\text{conc}}\) are under the assumption that, at steady-state, the binding of mRNA molecules to the ligands is stable. During the ligand binding event, the mRNA folding conformation does switch from its minimum free energy at the OFF
state to a more positive free energy at the ON state, resulting in a positive energy penalty $\Delta \Delta G_{mRNA}$, according to Equation B.8.

$$\Delta \Delta G_{mRNA} = \Delta G_{mRNA,ON} - \Delta G_{mRNA,OFF}$$ (B.8)

However, this energy penalty can be compensated by the free energy input from the binding of ligand to the mRNA, $\Delta G_{ligand}$ (negative value). $\Delta G_{ligand}$ can be calculated from the dissociation constant of ligand from RNA aptamer according to the Equation B.9.

$$\Delta G_{ligand} = RT \ln \left( K_D \right)$$ (B.9)

$R$ is the universal gas constant and $T$ is the temperature in Kelvin. $K_D$ is the ligand’s dissociation constant. A tightly binding ligand to RNA aptamer results in more negative $\Delta G_{ligand}$.

When $\Delta G_{ligand}$ can fully compensate the free energy penalty $\Delta \Delta G_{mRNA}$, the RNA folding switch can occur at thermodynamic equilibrium and the binding would be stable. In contrast, a weak ligand binding with a less negative $\Delta G_{ligand}$ cannot compensate a large free energy penalty, and consequently the transition from the OFF state to the ON state would not be stable. We define this condition as a metastable switching condition, where the chance of remaining bound to ligand is a function of the mRNA’s switching free energy. We calculate the switching free energy by summing $\Delta \Delta G_{mRNA}$ and $\Delta G_{ligand}$. When the switching free energy $(\Delta \Delta G_{mRNA} + \Delta G_{ligand})$ is negative, the mRNA-ligand binding is stable, and when the summation is positive the binding becomes metastable.

When a riboswitch is metastable, we assume that the metastable mRNA-ligand complex will partition into two distinct, but transitioning states: an unbound mRNA and a ligand-bound mRNA. Therefore, using the “mass-action kinetic” formulation and applying the mRNA conservation, the fraction of unbound and ligand-bound mRNAs can be written as Equations B.10 and B.11.
Based on the biophysical model, the actual translation initiation rate in the ON state for a metastable riboswitch can be calculated using three states: free mRNA, unbound mRNA from the metastable complex, and ligand-bound mRNA from the metastable complex, according to Equation B.12. Here, the concentration of free mRNA and metastable complex are $C_{\text{mRNA,1}}$ and $C_{\text{mRNA,3}}$, respectively.

$$r_{\text{ON,actual}} \propto \frac{C_{\text{mRNA,1}}}{C_{\text{mRNA, total}}} (r_{\text{OFF}}) + \frac{C_{\text{mRNA,3}}}{C_{\text{mRNA, total}}} \text{fraction}_{\text{unbound}} (r_{\text{OFF}}) + \frac{C_{\text{mRNA,3}}}{C_{\text{mRNA, total}}} \text{fraction}_{\text{mRNA:ligand}} (r_{\text{ON}})$$

(B.12)

With some mathematical rearrangement, Equation B.12 will reduce to Equation B.13, and then to Equation B.14.

$$r_{\text{ON,actual}} \propto \frac{r_{\text{OFF}} + \frac{C_{\text{mRNA,1}}}{C_{\text{mRNA, total}}} r_{\text{OFF}} + \frac{C_{\text{mRNA,3}}}{C_{\text{mRNA, total}}} r_{\text{ON}}}{1 + \exp(-\beta(\Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))} \exp(-\beta(\Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))$$

(B.13)

$$r_{\text{ON,actual}} = \frac{r_{\text{OFF}} + r_{\text{ON,conc}}}{1 + \exp(-\beta(\Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))}$$

(B.14)

The $r_{\text{OFF}}$ is calculated from Equation B.1 and $r_{\text{ON,conc}}$ is calculated according to Equation B.6. Finally, the actual activation ratio “AR$_{\text{actual}}$” is calculated as the ratio of $r_{\text{ON,actual}}$ and $r_{\text{OFF}}$ (Equation B.15).
\[
\Delta G = \Delta \exp(1) \Delta G
\]

\[
\Delta G = \Delta \exp(AR_1) \Delta G
\]

\[
\Delta G = \frac{1 + AR_{\text{conc}} \exp(-\beta(\Delta \Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))}{1 + \exp(-\beta(\Delta \Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))}
\]

Equation B.15

Again, with some mathematical rearrangements, the Equation B.15 will be simplified to Equation B.16, which is also presented as Equation 6.3 in Chapter 6.

\[
AR_{\text{actual}} = 1 + \frac{\exp(-\beta(\Delta \Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))}{1 + \exp(-\beta(\Delta \Delta G_{\text{mRNA}} + \Delta G_{\text{ligand}}))} (AR_{\text{conc}} - 1)
\]

Equation B.16

B.2.2. Thermodynamic minimization model to capture the effect of ligand and mRNA concentration on riboswitch’s function

The biophysical model of riboswitches assumes excess ligand for binding to mRNA in the ON state. However, it is possible that the concentration of ligands is not in excess and even lower than the mRNA level. This would affect the model predictions for riboswitch’s function. Here we used a thermodynamic minimization approach to predict the translation initiation rate for diverse mRNAs and ligand levels inside the cell.

A pool of total mRNA transcripts and total ligands that co-exist inside the cell can partition into three different configurations: free mRNA (state 1 in Figure 6-1B), free ligand, and ligand-bound mRNA complex (state 3 in Figure 6-1B) (Figure B-1). Each of these three configurations has specific concentration \(C\), volume \(V\), and Gibbs free energy \(\Delta G\). We simplify the intracellular environment into a water-filled, constant-volume box containing the three above species. For simplicity we also assume that the free ribosomes inside the cell are in excess and their number remain constant when binding to the riboswitches. Therefore, the concentration of each species can be estimated using a thermodynamic minimization approach, in which the total Gibbs free energy of the box, \(\Delta G_{\text{box}}\) (Equation B.17), is minimized subject to total mRNA and
ligand molar conservation (211). We should mention that, when Dirks et al. 2007 (211) calculated the mole fraction of each species inside the box, they assumed that the species in the box have similar sizes. That assumption is correct when dealing with multiple strands of RNA where all have roughly similar sizes. However, in the case of ligand and mRNA binding, the size of ligand can be 1000-fold smaller than the mRNA. Thus, the similar size assumption would not be correct. Therefore, we modified the original derivation in order to incorporate the volume of species in the \( \Delta G_{\text{box}} \) formulation.

\[
\Delta G_{\text{box}} = \Delta G_{\text{ref}} + \frac{1}{\beta} \sum_{i=1}^{N} v_i \left( \log v_i - \log Q_{v,i} - 1 \right) \tag{B.17}
\]

\( \Delta G_{\text{ref}} \) is the Gibbs free energy of reference and \( \beta \) is the apparent Boltzmann coefficient from (21). \( v_i \) is the volume fraction of each species \( i \) and is related to the molar fraction of specie \( i \), \( x_i \), according to Equation B.18.

\[
v_i = \frac{V_i}{V_S} x_i \tag{B.18}
\]

\( V_i \) is the volume of specie \( i \) and \( V_S \) is the volume of solvent. Here we assumed that the solvent is water and as the reference we set the \( V_S \) to be 1. Thus, by dividing \( V_i \) to \( V_S \), we normalize the volume of species to the volume of water molecules. The molar fractions \( x_i \) is written as Equation B.19, where \( C_S \) is the molar concentration of solvent (water molecules), which is 55.14 mol/lit at 37 °C.

\[
x_i = \frac{C_i}{C_S} \tag{B.19}
\]

In Equation B.17, \( Q_{v,i} \) is the adjusted partition function of specie \( i \) for its volume, and is calculated according to Equation B.20. \( Q_i \) is the partition function of specie \( i \), which is calculated according to Equation B.21, where \( \Delta G_i \) is the Gibbs free energy for specie \( i \) (Figure B-1).
\[ Q_{v,j} = Q_{1} \left( \frac{1}{V_{i}} \right) \]  \hspace{1cm} (B.20)

\[ Q_{i} = \exp \left( -\beta \Delta G_{i} \right) \]  \hspace{1cm} (B.21)

The total mRNA and ligand molar conservation can be written as:

\[ x_{\text{mRNA, total}} = x_{\text{mRNA, 1}} + x_{\text{mRNA, 3}} \]  \hspace{1cm} (B.22)

\[ x_{\text{ligand,total}} = x_{\text{ligand,free}} + x_{\text{mRNA, 3}} \]  \hspace{1cm} (B.23)

Finally, the 30S ribosomal subunit can bind to the mRNA and initiate translation from two separate configurations; the free mRNA (state 1) using a binding Gibbs free energy \( \Delta G_{\text{total,OFF}} \), and ligand-bound mRNA (state 3) using a binding Gibbs free energy \( \Delta G_{\text{total,ON}} \). Therefore, the translation initiation rate per total mRNA level in the presence of ligand can be calculated according to Equation B.6.

\section*{B.2.3. Riboswitch Calculator: automated design of synthetic riboswitches}

The Riboswitch Calculator offers two design modes for the automated design of riboswitches to control translation initiation: a reverse engineering mode and a forward engineering mode. For both two modes, four inputs are necessary for the calculations: the RNA aptamer’s sequence, the RNA aptamer’s structural constraint upon binding to the ligand (in the “dot-bracket” format), the binding free energy \( \Delta G_{\text{ligand}} \) or its equivalent dissociation constant \( K_{D} \) of the ligand to RNA aptamer, and the protein coding section for the gene of interest. Here, we describe these two modes in detail.

Reverse engineering mode: this mode of the Riboswitch Calculator evaluates the performance of a given riboswitch sequence by calculating the following energetic terms: \( \Delta G_{\text{total,OFF}} \) for the OFF state, \( \Delta G_{\text{total,ON}} \) of the ON state by incorporating the structural constraint
of the RNA aptamer, the $\Delta G_{\text{ligand}}$ calculated from the $K_D$ of ligand to RNA aptamer (Equation B.9), and $\Delta \Delta G_{\text{mRNA}}$ according to Equation B.8. Finally the translation initiation rate at OFF and ON states (Equations B.1 and B.2) and as a result the riboswitch’s maximum activation ratio ($AR_{\text{max}}$) will be calculated. It also predicts the concentration-limited activation ratio ($AR_{\text{conc}}$) and actual activation ratios ($AR_{\text{actual}}$) using the given mRNA and ligand concentration and the calculated $\Delta \Delta G_{\text{mRNA}}$ (Equations B.7 and B.16). This mode of Riboswitch Calculator has been validated using 15 synthetic theophylline riboswitches that controlled the translation of luciferase gene (163) (Figure 6-2CDE).

Forward engineering mode: this mode of the Riboswitch Calculator is able to design riboswitch sequences from any given RNA aptamer sequence in order to control translation initiation of a gene of interest with large dynamic range. To design the riboswitch, the input RNA aptamer sequence will be located within the 5’ UTR of the gene of interest and the sequences upstream (pre-aptamer) and downstream (post-aptamer) of the RNA aptamer in the 5’ UTR will be carefully optimized to achieve the design objectives. For cloning purposes, proper restriction sites can be inserted in the designed sequences. Importantly, since the DNA sequence space to design the pre- and post-aptamer regions is substantially large (Figure 6-3B), we use an optimization algorithm to search for the functional riboswitches (Figure B-2). The optimization algorithm starts with an initial sequence for pre-aptamer and post-aptamer regions and continues with rounds of mutations (substitution, insertion, and deletion, each with a pre-defined chance) until reaching the design objectives. At each mutation, the core biophysical model of the Riboswitch Calculator evaluates the $\Delta G_{\text{total,OFF}}$, $\Delta G_{\text{total,ON}}$, $\Delta G_{\text{ligand}}$, and $\Delta \Delta G_{\text{mRNA}}$ by applying the structural constraint of the input RNA aptamer. Mutations that meet the specified objective function requirements such as minimizing the $\Delta \Delta G_{\text{mRNA}}$, maximizing the dynamic range of riboswitch’s function ($AR_{\text{max}}$), or reaching a user-defined OFF or ON expression targets, will be
selected for next rounds. This mode of Riboswitch Calculator has been validated using 62
designed riboswitches that bind to theophylline, TMR, fluoride, DNT, dopamine, and thyroxine,
and control the translation of luciferase and mRFP1 reporter proteins (Figures 6-3, 6-4 and 6-5).

B.2.4. Incorporation of pseudoknot structural constraint into Riboswitch Calculator

The biophysical model of structured 5’ UTR was parameterized using canonical
secondary structures. Although many SELEX based RNA aptamers have canonical secondary
structures, most of natural RNA aptamers contain pseudoknot base pairing (212), such as add-A
and xpt-G aptamer (213), met-A aptamer (214), met-H aptamer (168), and c-di-GMP binding 84-
Cd aptamer (215). In order to generalize the Riboswitch Calculator to all types of RNA aptamers,
we developed an automated method to calculate the folding energy of the pseudoknot structures
based on their given structural constraint. The algorithm takes the pseudoknot structural
constraint in the form of “pseudoknots dot and bracket”. It calculates the pseudoknot structure
folding energy according to Equation B.24.

\[
\Delta G_{PK} = \Delta G_{\text{assembly}} + \Delta G_{\text{coaxial}} + \Delta G_{\text{hyb}} - T \Delta S_{PK} \tag{B.24}
\]

\(\Delta G_{\text{assembly}}\) is the free energy released from initiating the pseudoknot base pairing between
two hairpins and is set to 1.3 kcal/mol, according to Cao and Chen (171, 216). \(\Delta G_{\text{coaxial}}\) is coaxial
stacking energy between two stems that have 0 or 1 nt distance. The coaxial stacking energy is a
function of closing base pairs and their neighboring base pairs. The related energies for coaxial
stacking energy calculations, according to Cao and Chen (216), were taken from the DotKnots
algorithm (172, 217). \(\Delta G_{\text{hyb}}\) is the free energy contribution of all the base pairs in the pseudoknot
structures. To calculate this base pairing energy, the pseudoknot structure will be automatically
divided into two canonical structures (Figure B-19A), where their folding energies are calculated,
according to Equation B.25, using Vienna RNA suite, version 1.8.5. It is important to note that our model only calculates the folding energy of the standard pseudoknot base pairings, as the energy parameters for the non-standard pseudoknot base pairings are not available yet (Figure B-19B).

\[ \Delta G_{\text{hyb}} = \Delta G_{\text{stem}} - \Delta G_{\text{loop}} \]  

(B.25)

In this equation, \( \Delta G_{\text{stem}} \) is the folding energy of the RNA hairpin and \( \Delta G_{\text{loop}} \) is the energy penalty for closing the loop (as a result of entropy loss), which is equivalent of the free energy of hairpin loop’s closing base pair. On the other hand, the effect of entropy change upon pseudoknot formation is included within \( \Delta S_{\text{PK}} \) (Equation B.24). A virtual bond model has been proposed by Cao and Chen (171, 216, 218), in which the entropy of pseudoknot formation is related to the length of hairpin stem and hairpin loops within the pseudoknot structure. Therefore, our algorithm first determines the effective length of hairpin stem and hairpin loops. Next, \( \Delta S_{\text{PK}} \) is calculated using the entropy functions provided by Cao and Chen for different types of pseudoknots, such as linear-loop (171, 216) and loop-loop (218), and also by using the fitted entropy functions available in DotKnots algorithm (172, 217). \( T \) in Equation B.24 is temperature in kelvin unit.

Overall, our automatic method is capable of calculating the folding energy of standard pseudoknot structures just by using their structural constraints. These pseudoknots can contain bulges or mismatches in their hairpin stems, or can have extra hairpin stems on their loop or on their hairpin stems.
B.2.5. Combinatorial library and high throughput screening method

Lynch et al. 2007 (63) generated four different combinatorial libraries (N$_5$, N$_6$, N$_7$, and N$_8$), where they randomized the linker region between theophylline binding aptamer and the SD region of IS10-LacZ fusion protein. Using high-throughput screening, they identified a handful of riboswitches with wide range of activation ratios (Figure B-3A). Using our developed biophysical model, we sought to predict the function of all the individual members of these libraries and identify which riboswitch is silent and which one is functional. Therefore, we performed our calculations for all the riboswitch variants in the N$_5$, N$_6$, N$_7$, and N$_8$ libraries. Interestingly, we found that the chance of finding a functional riboswitch from these libraries dropped dramatically as the library size increased (Figure B-3B). Moreover, our predictions show that the number of predicted functional riboswitches largely exceeded the number of experimentally identified ones as the library size increased, demonstrating a technological challenges for the high-throughput screening methods (Figure B-3C). Finally, the in vivo measured β-galactosidase levels (Miller units) at both OFF (zero theophylline) and ON (2 mM theophylline) states, for 20 identified and reported riboswitches in these libraries, were correlated with the model predicted ΔG$_{\text{total}}$ values (Figure B-3D). The biophysical model was able to accurately predict the maximum activation ratio AR$_{\text{max}}$ for most of these riboswitches (Figure B-3E).

B.2.6. Volume and intracellular concentration of mRNA and ligands

The average volume of E. coli cells is estimated to be 10$^{-15}$ liters. As the water molecules occupy 75% of the cellular volume (219), the free volume inside the cell is approximately 10$^{-16}$ liters. The volume of one molecule of water is 3x10$^{-26}$ liters. Therefore, there are roughly 10$^9$
water molecules (solvents) inside the cell, which is much greater than the number of mRNA or ligand inside the cell \((C_S >> C_i)\). The size of the luciferase mRNA (around 1800 nucleotides) is almost the same size as the 16S ribosomal RNA. To estimate this mRNA’s molar volume, we use as a reference the 30S ribosome, which has a volume of about \(3 \times 10^{-21}\) liters. We then estimate that luciferase mRNA is as large as \(10^{-21}\) liters, or 10,000 fold larger than water molecules \((V_{\text{luciferase}}/V_{\text{water}} = 10000)\). Similarly, since the mRFP1 mRNA is almost half size of the luciferase gene (900 nucleotides), we estimate that the molar volume of mRFP1 mRNA is around \(5 \times 10^{-22}\) liters \((V_{\text{mRFP1}}/V_{\text{water}} = 5000)\).

Theophylline has density and molecular weight of 0.63 g/ml and 180.2 g/mol, respectively. The volume of one molecule of theophylline is around \(5 \times 10^{-25}\) liters, which is around 10-fold larger than water molecules \((V_{\text{theophylline}}/V_{\text{water}} = 10)\). Similarly, as TMR has two additional aromatic rings compared to theophylline, we assume that its molar volume is twice as big as theophylline and 20-fold larger than water molecules \((V_{\text{TMR}}/V_{\text{water}} = 20)\). Finally, we assume that the elemental fluoride ion has the same molar volume as water molecule \((V_{\text{fluoride}}/V_{\text{water}} = 1)\). These estimates can be found in Supplementary_Data_6.

The luciferase gene was transcribed from pUC19 vector and mRFP1 from pFTV1 vectors. The pUC19 vector has around 75 copies at 37 °C (220) and pFTV1 (ColE1 based) has between 40-60 copies inside E. coli Top10 or DH10B (221). The tac promoter drives the production of luciferase gene on pUC19, while J23100 promoter, which is slightly weaker than tac promoter, drives the transcription of mRFP1 gene on pFTV1. However, the longer length of luciferase gene compared to mRFP1 may lead to a lower transcription efficiency for luciferase gene due to higher chance of RNA polymerase dissociation at longer distances. Therefore, we estimate that there are around 5,000 and 10,000 mRNA molecules inside the cell carrying the luciferase and mRFP1 genes, respectively. These estimates are equivalent of ~70 and ~170 copies
of mRNA per pUC19 and pFTV1 plasmid DNAs, respectively, which fall in the range of mRNA copies generated from other natural promoters across genome (222, 223). Transferring the luciferase expression cassette (which includes the tac promoter, 5’ UTR, and luciferase gene) to other plasmid vectors with different copy numbers (pBAC, p15A, and pFTV1) results in different range of mRNA levels. Based on our measurements, we estimated that mRNA molecules for luciferase gene are present in around 170, 3730, and 3130 copies under pBAC, p15A, and pFTV1 plasmids, respectively (Figure B-17A). In addition, the mRNA levels for mRFP1 gene are estimated to be 1,000 and 50,000 molecules when using either the AEB-3 or LmrA promoters on pFTV1 plasmid, respectively (Figure B-17B). Finally, we assume that the intracellular ligand concentration is almost 100-fold less than the extracellular ligand concentrations. For the case of theophylline, this is slightly higher than its reported intracellular concentration (224). However, for different ligands, this estimated constant ratio can be different, depending on their partition coefficient value across the cell membrane.

B.2.7. The mRNA and ligand level comparison between cell-free and in vivo systems for mRNA dose response experiments

For the mRNA dose response assays (Figure 6-5B), 30 uM TMR in a 10 ul in vitro S30 extract reaction is equivalent of around 2x10^{14} TMR molecules. Addition of a given concentration of DNA (A ng/ul) to the test tube results in 2xAx10^9 DNA molecules inside the reaction (the DNA plasmid has around 4000 base pairs and the molecular weight of one DNA base pair is 660 g/mol). If we assume that the transcription process within the E. coli S30 extract produces on average 10 mRNA transcripts per DNA template, we will have 2xAx10^{10} mRNA molecules inside the reaction mixture. In our cell-free transcription-translation assay, A varied from 0.1 to 100.
Therefore, the ratio of total ligand \((2 \times 10^{14})\) to total mRNA \((2 \times 10^{10})\) varies from \(10^5\) to \(10^2\). Therefore, we can approximate that in all cases the ligand is present in excess, and the number of free ligand is almost equal to the total ligand level \(C_{\text{ligand,total}} \sim C_{\text{ligand,free}}\). Using the mass action kinetic formulation and the dissociation constant for TMR aptamer, we obtain the ratio of free mRNA to the ligand bound mRNA, according to Equation B.26.

\[
K_D = \frac{C_{\text{ligand, free}}}{C_{\text{mRNA, free}}} \approx \frac{C_{\text{ligand, total}}}{C_{\text{mRNA, -ligand}}} \tag{B.26}
\]

With 30 \(\mu\)M total TMR and \(K_D\) of 60 nM, the concentration of mRNA-ligand complex is 500-fold larger than free mRNA, hence \(C_{\text{mRNA,total}} \sim C_{\text{mRNA-ligand}}\). This explains the increase in activation ratio as the mRNA level increases (Figure 6-5A, dilute volume). It is noteworthy that addition of mRNA molecules does not reduce the free volume in the test tube (Figure 6-5A, dilute volume).

In contrast to the cell-free transcription-translation reaction, the mRNA and ligand molecules inside the cell are present in very limited amounts. As explained in previous section, the number of transcribed mRNA molecules from pUC19 plasmid is roughly 5,000 molecules. In addition, assuming that the intracellular concentration of TMR is around 0.2 \(\mu\)M, there are roughly 120 molecule of TMR present inside the cell, which is in the same order of magnitude as the mRNA molecules. Therefore, to determine the fraction of ligand-bound mRNAs we need to use thermodynamic minimization approach shown in Equation B.17. Interestingly, since the total volume of cell is limited, increasing the number of mRNA molecules reduces the free volume, which further impacts the mRNA-ligand binding fractions (Figure 6-5A, crowded volume). Overall, our thermodynamic minimization calculation is able to predict the fraction of free-mRNA, free-ligand, and mRNA-ligand complex as the total mRNA level increases. More
importantly, this model can correctly predict the decrease in activation ratio as the mRNA level increases (Figure 6-5C).

B.2.8. Designing a “perfect” riboswitch

By combining the biophysical model of riboswitch function together with the optimization algorithm, we investigated whether we could design a so-called “perfect” riboswitch that used a best-possible, hypothetical aptamer to bind a ligand and maximally regulate gene expression. Using this design, we investigated how the maximum available ligand concentration and the aptamer’s binding affinity determined the best-possible sensing scenarios. We chose a short 22 nt hypothetical aptamer sequence (Figure 6-6A) and highly optimized its pre-aptamer, post-aptamer, and coding section to generate a “perfect” riboswitch that exhibited an extremely large activation ratio \( \text{AR}_{\text{max}} = 16000 \), but with a large switching free energy, \( \Delta \Delta G_{\text{mRNA}} = 15.2 \text{ kcal/mol} \). Using Equation 6.3 in Chapter 6, we predicted the \textit{in vivo} actual activation ratio \( \text{AR}_{\text{actual}} \) for this riboswitch at a wide range of \( \Delta G_{\text{ligand}} \) (from pico-molar to milli-molar ligand binding affinities) and maximum ligand concentrations, and by assuming \( V_{\text{mRNA}}/V_{\text{water}} = 10000 \), \( V_{\text{ligand}}/V_{\text{water}} = 20 \), and total mRNA concentration = 0.16 \( \mu \text{M} \) (equivalent of 50 mRNA molecules inside the cell). We found that riboswitch activation was extremely high when using a high-affinity aptamer and when the ligand is plentiful (\( \text{AR}_{\text{actual}} > 1000\text{-fold} \)). When picomolar ligand levels must be sensed, a “perfect” riboswitch will activate expression by 20-fold using a high-affinity aptamer, and 3-fold using a very low-affinity aptamer. These calculations illustrate the outer bounds of riboswitch sensing regardless of the aptamer sequence or structure (Figure 6-6B).
B.2.9. Examples of data analysis (DNT, dopamine, and thyroxine riboswitches)

Riboswitch-specific activation ratios were calculated identically for all riboswitches. Here, we provide specific example calculations for the DNT, dopamine, and thyroxine riboswitches to illustrate the procedure.

DNT riboswitches and no-aptamer controls were characterized in bacterial assays where the media was supplemented with 1% (v/v) DMSO and either 0 or 1 mM DNT. The addition of DNT non-specifically increased reporter expression. The no-aptamer control's mRFP1 fluorescence was 1765 au without DNT and 8510 au with 1 mM DNT, yielding an activation ratio of 4.82-fold. Under the same buffer and ligand conditions, the DNT-3 riboswitch's mRFP1 fluorescence was 22 au without DNT and 1176 au with 1 mM DNT, yielding an activation ratio of 53.31. In our data analysis, we eliminate the effects of non-specific expression changes by dividing the ON/OFF activation ratio of the riboswitch by the ON/OFF activation ratio of the no-aptamer control. The riboswitch-specific activation ratio for DNT-3 is therefore 53.31 / 4.82 = 11.10.

The dopamine riboswitches and no-aptamer control were characterized in cell-free transcription-translation assays supplemented with 5 mM ascorbic acid. The no-aptamer control's luciferase expression level (luminescence) was 124680 au without dopamine and 76340 au with 1 mM dopamine, resulting in an activation ratio of 0.61. The Dopa-5 riboswitch's luciferase expression level was 10794 without dopamine and 12898 au with 1 mM dopamine, yielding an activation ratio of 1.19. The riboswitch-specific activation ratio for Dopa-5 was therefore 1.19 / 0.61 = 1.95.

The thyroxine riboswitches and no-aptamer control were characterized in cell-free transcription-translation assays supplemented with 1.5 mM NaOH and 100 mM NaCl. The no-aptamer control's luciferase expression level was 24659 au without thyroxine and 9104 au with
150 μM thyroxine, yielding a 0.37-fold activation ratio. The T4-2 thyroxine riboswitch's luciferase expression level was 1773 au without thyroxine and 1543 au with 150 μM thyroxine, yielding an activation ratio of 0.87. The riboswitch-specific activation ratio for T4-2 was therefore $0.87 / 0.37 = 2.35$. 
VITA

Amin Espah Borujeni

Education

<table>
<thead>
<tr>
<th>Institute</th>
<th>Field of study</th>
<th>Degree</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pennsylvania State University, USA</td>
<td>Chemical Engineering</td>
<td>PhD</td>
<td>2009-2015</td>
</tr>
<tr>
<td>Delft University of Technology, The Netherlands</td>
<td>Biochemical Engineering</td>
<td>MSc</td>
<td>2007-2009</td>
</tr>
<tr>
<td>Sharif University of Technology, Iran</td>
<td>Chemical Engineering</td>
<td>BSc</td>
<td>2003-2007</td>
</tr>
</tbody>
</table>

Honors and Awards

- NAR Breakthrough Article, Nucleic Acids Research journal, November 2015
- Best Graduate Student Paper Award, Chemical Engineering, Penn State, September 2014
- Young Researcher Travel Award, Synthetic Biology 5.0 conference, Stanford, USA, June 2011
- Cum Laude, M.Sc. degree, Delft University of Technology, The Netherlands, July 2009
- Huygens Scholarship, Delft University of Technology, The Netherlands, September 2008
- Brilliant Talent Recognition, B.Sc. degree, Sharif University of Technology, Iran, July 2007
- Ranked top 1% among more than 100,000 students participated in the national entrance exam for undergraduate studies, Iran, July 2003

Publications

(Note: Publication name was changed to Espah Borujeni A. since 2013.)

1) Espah Borujeni A., Farasat I., Smith A., Huso W., Lundgren N., Salis HM. “Precise Quantification of Translation Inhibition by RNA Structures Overlapping with the Ribosome Footprint at N-terminal Coding Sections.”, *in progress*

2) Espah Borujeni A., Salis HM. “Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism.”, *under review*

