The Dissertation Committee for Alberto Jorge Vazquez Anderson Certifies that this is the approved version of the following dissertation:

Insights into RNA design from novel molecular tools

Committee:

Lydia M. Contreras, Supervisor

Rick Russell

Hal Alper

Pengyu Ren

George Georgiou

Insights into RNA design from novel molecular tools

by

Alberto Jorge Vazquez Anderson, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2016

Dedication

To God for the countless gifts throughout my life and specially for inspiring me to study His enthralling creation at the molecular level.

To my beloved wife, for her unconditional support and love throughout this journey.

To my son and daughter, for they confer meaning to everything I do.

To my parents, for I owe them everything.

To my countrymen, those countless immigrants for they work tirelessly to build this nation and, who face absurd hate and discrimination due to their lack of privilege

Acknowledgements

First I want to thank Dr. Lydia M. Contreras for her unrestricted support, invaluable advice and crucial teachings. I will always be grateful for the opportunity. Second, I want to show appreciation to the Contreras group specially Mia, Kevin V., Kevin B. and Steve for their friendship, the philosophical and highly productive discussions, the hard work and the amazing conversations.

Insights into RNA design from novel molecular tools

Alberto Jorge Vazquez Anderson, PhD

The University of Texas at Austin, 2016

Supervisor: Lydia M. Contreras

RNA, previously recognized merely as a messenger of genetic information, has been recently rediscovered as a versatile molecule with a central role in cellular regulation. These regulatory functions are enabled by its specific chemical makeup that allows it to fold into intricate and flexible structures. In stark contrast with DNA, RNA forms a variety of structural motifs that serve as efficient *points of contact* in molecular recognition. It is therefore clear, that dynamic RNA structures dictate the binding availability of interfaces that play important roles in molecular regulation inside living cells. As such, the need for tools that can accurately capture and predict RNA structure *in vivo* continues to be essential to understand RNA function. To this end, my dissertation focuses on the development of molecular tools to predict and characterize accessible RNA interfaces in their native environment. First, I established the usefulness of a fluorescence-based *in vivo* oligonucleotide hybridization approach to identify accessible interfaces by characterizing numerous RNA regions in several biologically relevant molecules in *E. coli*. I then described these RNA interactions using a biophysical model based on thermodynamic principles and incorporating large sets of data collected using this fluorescence-based system. This approach displayed improved prediction capabilities of RNA accessibility compared to un-optimized versions without incorporation of *in vivo* data. Finally, I detailed the development and application

of a high throughput tool for the large-scale characterization of accessible interfaces within native RNAs in a single experiment. In this approach, *in vivo* oligonucleotide hybridization was coupled to transcriptional elongation control to allow analysis via next generation sequencing. This tool was used to obtain complete landscapes of functional structure for 72 regulatory molecules in a single experiment (>1000 regions). Altogether the results of this high throughput approach revealed a pattern indicating that RNA-RNA interaction sites are either highly accessible or highly protected, suggesting their binding status (e.g. actively bound or unbound). In addition, within bacterial small RNAs, our approached revealed the role of the global regulator Hfq as universal structural relaxer. The compendium of these tools provides a unique and fundamental perspective in the study of functional RNA structure, namely, the identification of dynamic structures. Furthermore, the information provided by these approaches significantly aids in the design of synthetic RNAs for a variety of purposes, including gene expression control. In my time at the University of Texas at Austin, I participated in a total of seven scientific articles: as a leading author in four works (two published and listed below and, two more near publication) and as a collaborating author in three others (one published, one in review and one more in preparation).

1. J. Vazquez-Anderson, and L.M. Contreras, "Regulatory RNAs: Charming Gene Management Styles for Synthetic Biology Applications" *RNA Biology*, 10(12):1778-97 (2013).
2. S. Sowa, J. Vazquez-Anderson, C. Clark, R. De la Peña, K. Dunn, E. Fung, M. Khoury, and L.M.Contreras, "Exploiting post-transcriptional regulation to probe RNA structures in vivo via fluorescence," *Nucleic Acids Research,* 43(2):e13 (2015).

Table of Contents

List of Tables xii

List of Figures xiii

Chapter One 1

Introduction and background 1

1.1 Introduction 1

1.2 Current techniques to characterize RNA structure and the iRS^3 3

1.3 Current approaches to predict hybridization efficacy and the inTherAcc approach 5

1.4 current progress in understanding functional structure in regulatory RNAs and the INTERFACE approach 11

1.5 Summary of research objectives and accomplishments 14

Chapter Two 17

Exploiting post-transcription regulation to probe RNA structures in vivo via fluorescence 17

2. 1 Introduction 17

2.2 Results 18

2.2.1 Molecular design and optimization of iRS^3 18

2.2.2 iRS^3 fluorescence is specific to the interaction between probe and target RNA 20

2.2.3 iRS^3 can discriminate between accessible and protected regions along the group I intron 25

2.2.4 Assaying a probe library along the group I intron 27

2.2.5 iRS^3 can discriminate between group I intron mutants 33

2.3 Discussion 35

2.4 Material and Methods 40

2.4.1 Plasmids and Strains 40

2.4.2 Computational Analysis of Probes 40

2.4.3 Flow cytometry 41

2.4.4 In vitro binding Assays 42
2.4.5 Northern Blot Analysis of iRS^3 transcript.....43
2.4.6 In vivo dimethyl sulfate footprinting 43
2.4.6.1 Primers fluorescent labeling 43
2.4.6.2 In vivo dimethyl sulfate treatment 44
2.4.6.3 Capillary Electrophoresis.....44

Chapter Three.....45

Optimization of a biophysical model using large scale *in vivo* antisense RNA hybridization data in bacteria displays improved prediction capabilities 45
3.1 Introduction.....45
3.2 Results.....46
3.2.1 Description of asRNA hybridization efficacy by a thermodynamic model that includes a regional measure of interaction availability46
3.2.2 Model optimization using in vivo experimental profiling of asRNA hybridization efficacy 51
3.2.3 The inTherAcc Model Proves Effective in Predicting High asRNA Hybridization Regions in Other RNA targets 58
3.2.4 inTherAcc Aids in Prediction of Target mRNAs 64
3.4 Discussion and conclusions 66
3.4 Methods.....69
3.4.1 Plasmids and strains 69
3.4.2 Selection of target RNAs 71
3.4.3 Fluorescence Measurements and Calculations of asRNA hybridization using the in vivo RNA Structural Sensing System (iRS^3).....72
3.4.4 In vivo DMS footprinting and calculation of regional availability ($\boldsymbol{\theta}$) 73
3.4.5 Derivation of the accessibility-based thermodynamic model 74
3.4.6 Calculation of free energy of hybridization ($\Delta \boldsymbol{GasT}$) 74
3.4.7 Calculation of free energy of the target region ($\Delta \boldsymbol{GTf}$).....75
3.4.8 Calculation of regional availability.....75
3.4.9 Calculation of free energy of folding for the asRNA ($\Delta \boldsymbol{Gasf}$) ..76

3.4.10 Model optimization via regression analysis using experimental hybridization data....76
3.4.11 Selection of target regions for evaluation of model prediction power 77
3.4.12 Statistical Evaluation of Model Prediction Power 79
3.4. 13 Evaluation of Prediction of sRNA-mRNA Binding Regions ..79
3.4.14 Strains and culture conditions for MS2 pull-downs 81
3.4.15 RNA Preparation for Evaluation of Zms4 and Zms6 mRNA Targets 81
3.4.16 Purification of MS2-MBP fusion proteins 82
3.4.17 Affinity purification of MS2-MBP fusion proteins 83
3.4.18 Transcriptomics data analysis 83

Chapter Four 85

High throughput in vivo sensing of accessible interfaces in a large ensemble of small RNAs reveals Hfq as a universal structural relaxer 85
4.1 Introduction 85
4.2 Results 86
4.2.1 Harnessing transcriptional regulation for high throughput characterization of RNA accessible interfaces 86
4.2.2 Validating molecular features governing the ability of INTERFACE to capture regional accessibility 91
4.2.3 Large-scale characterization of accessible interfaces in native regulatory RNAs aided by machine learning reveals potential functional regions....93
4.2.4 Hfq facilitates sRNA-mRNA interaction by releasing target sRNA structure and increasing accessibility....100
4.3 Discussion 104
4.4. METHODS 106
4.4.1 Plasmids and strains 106
4.4.1.2 Synthesis of constructs....107
4.4.2 INTERFACE experiments 108
4.4.2.3 Total RNA extraction....109

4.4.4 Computation selection of accessible interfaces 109
4.4.4.1 Estimation of binding potential using a biophysical model109
4.4.4.2 Machine learning algorithm 110
4.4.4.3 Computational simulations to test for algorithm performance 112
4.4.5 Synthesis of DNA libraries for next generation sequencing.....112
4.4.6 Illumina sequencing of DNA libraries 113
4.4.7 Computational processing pipeline of sequencing results 113
4.4.8 Calculation of relative accessibility 114
4.4.9 Estimating Hfq-dependency class from accessibility changes between parent and Hfq-deficient strains 115
4.4.10 Sequence motif discovery and search 115
Chapter Five 117
Conclusions and perspectives 117
Appendices 121
Appendix A: Supplementary data for Chapter Two 121
Supplementary figures for Chapter two 121
Appendix B: Supplementary data for Chapter Three 125
Supplementary figures Chapter Three 125
Supplementary tables Chapter Three 132
Appendix C: Supplementary data for Chapter Four 154
Supplementary figures for Chapter Four 154
Supplementary tables for Chapter Four 230
References 254

List of Tables

Table 4.1. INTERFACE reveals differential dependency of Hfq for all sRNAs analyzed and correlates strongly with pull down data obtained from the literature. 103

Table B.1. List of asRNAs used in this study 132

Table B.2. List of target molecules and cloning strategy used in this study. 145

Table B.3. Estimated coefficients and statistical measures of goodness of fit for the regressions of inTher and inTherAcc models, which were optimized using in vivo data. 148

Table B.4. Summary of performance results for the prediction of mRNA targets for Z. mobilis. 149

Table C.1. List of sRNAs in this study and relevant information on sRNA-mRNA binding sites, stress-related responses and Hfq-dependence 230

List of Figures

Figure 2.1. Fundamentals of the *in vivo* RNA Structural Sensing System (iRS^3). 19

Figure 2.2. GI intron *Tetrahymena* Ribozyme Model System......21

Figure 2.3. Fluorescence shifts result from specific interactions between the reporter and the *trans* target RNA......23

Figure 2.4. Pilot test reveals that fluorescence assay can detect relative levels of accessibility. 26

Figure 2.5. iRS3 can capture difference in accessibility along the length of the wild-type group I intron. 29

Figure 2.6. iRS^3 differs from *in vivo* DMS footprinting findings in the P3 domain of the gI intron......32

Figure 2.7. iRS^3 can detect different levels of structural disruption in intron variants......34

Figure 2.8. Oligonucleotide vs. small-molecule *in vivo* structural probing......38

Figure 3.1. Proposed accessibility-based mechanism of anti-sense hybridization in living cells. 48

Figure 3.2. Structural target availability. 50

Figure 3.3. asRNA hybridization map as measured by *in vivo* oligonucleotide hybridization. 54

Figure 3.4. Relative significance of each term in the (A) inTherAcc and (B) inTher models......56

Figure 3.5. Improvement in performance for *in vivo* optimized models underscores the influence of intracellular factors......58

Figure 3.6. Experimental evaluation of hybridization efficacy in four RNAs shows inTherAcc model prediction accuracy comparable to that of benchmark IntaRNA. 61

Figure 3.7. Regression analysis on experimental versus inTherAcc-(top), inTher-(center) and IntaRNA-(bottom) predicted hybridization efficacy for (A) 2-MS2 and (B) gII intron. 63

Figure 3.8. inTherAcc aids in prediction of mRNA targets for *Z. mobilis* (A) Zms4 and (B) Zms6. 66

Figure 4.1. Modular engineering of a synthetic transcriptional control for high throughput characterization of RNA accessible interfaces.....88

Figure 4.2. INTERFACE experimental workflow. 90

Figure 4.3. INTERFACE allows for high-throughput characterization of functional structure. 93

Figure 4.4. Characterizing the acessosome for a large ensemble of sRNAs. ..95

Figure 4.5. Large-scale characterization of accessible interfaces in native regulatory RNAs reveals functional regions.....99

Figure 4.6. INTERFACE analysis reveals structural changes upon protein binding in small RNAs.....101

Figure 4.7. The accessosome sits at the "core" of the structure-function relationship 106

Figure A.1. General Methodology for iRS^3 experimental design. 121

Figure A.2. Northern Analysis of iRS^3 reporters show that reporter transcript levels do not correlate with fluorescence output. 122

Figure A.3. All probes bind in vitro to the denatured gI intron in the context of total RNA extract. 123

Figure A.4. Test for binding affinity bias. 124

Figure B.1. Experimental plasmids and Golden Gate cloning procedure for synthesizing asRNAs of interest......125

Figure B.2. Engineering the iRS^3 for characterizing native transcripts. 126

Figure B.3. DMS reactivity data for group I intron. 127

Figure B.4. Local versus regional base pairing probability-DMS correlations show no observable correlation for the group I intron at the local level.....128

Figure B.5. Undesirable correlation between folding energy of asRNA and hybridization efficacy 129

Figure B.6. Linear regression residuals for inTher (A) and inTherAcc (B).130

Figure B.7. Detailed results on prediction performance benchmark study .131

Figure C.1. INTERFACE plasmids (left) for heterologously expressed target RNAs and (right) for native target RNAs......154

Figure C.2. INTERFACE traces for each region characterized in the group I intron......155

Figure C.3. INTERFACE accessibility heat maps for each sRNA molecule analyzed. Red is accessible, gray is in the middle and blue is inaccessible......158

Chapter One

Introduction and background

1.1 Introduction

Ever since RNA functions (other than the transmission of genetic information) were discovered in the 1980's (Cech, Zaug et al. 1981), RNA has been at the center of novel research aimed at understanding and exploiting its versatility to interact with other molecules. Its capacity to exert regulation through molecular interactions has contributed to most of the interest seen in RNA in the last three decades (Vazquez-Anderson and Contreras 2013). RNA's unique chemical makeup enables its intrinsic ability to fold into intricate structures and adopt shapes suitable for intermolecular interactions. Therefore, characterizing, understanding and predicting RNA structure remains at the heart of RNA research.

From the experimental point of view, RNA structure has been extensively studied *in vitro* using a variety of techniques that include X-ray crystallography, NMR and, chemical and enzymatic probing (Tijerina, Mohr et al. 2007, Scott and Hennig 2008, Edwards, Garst et al. 2009). As a consequence of the realization that RNA folding is influenced by several factors present in the cellular environment (Schroeder, Grossberger et al. 2002, Leamy, Assmann et al. 2016), researchers developed approaches to study RNA structure inside living cells such as *in vivo* DMS footprinting (Tijerina, Mohr et al. 2007) and *in vivo* SHAPE (Spitale, Crisalli et al. 2013). The field reached its most recent development stage upon achieving *in vivo* high throughput characterization of RNA structure (Lorenz, Wolfinger et al. 2016, Silverman, Berkowitz et al. 2016). Despite the value that these approaches have in understanding RNA function, they represent only a piece of the puzzle in comprehending RNA intermolecular interactions. For this reason,

several experimental approaches, both low and high-throughput, have been developed to characterize binding sites and interacting partners, shedding light into the mechanisms of RNA regulation (Li, Song et al. 2014, Holmqvist, Wright et al. 2016, Nguyen, Cao et al. 2016). Nevertheless, a common feature of all the approaches mentioned above is the complexity that comes along when planning and executing experiments.

Consequently, many research groups have endeavored to create *in silico* predictive approaches aimed at both predicting RNA structure and RNA intermolecular interactions (Gorodkin and Ruzzo 2014). Secondary structure prediction approaches can be classified into two main groups: thermodynamics-based and stochastic-based. For predictive approaches aimed at predicting intermolecular interactions, two classes are also considered: concatenation and accessibility-based approaches (Backofen 2014). For simplicity, thermodynamic and accessibility-based approaches are often applied to forecast RNA function from RNA structure predictions, particularly regulation of gene expression. However, computational tools only have a limited scope and accuracy (Mathews, Sabina et al. 1999).

The information provided by the approaches above has just started to be instrumental in understanding and manipulating natural RNA systems often by designing and engineering synthetic RNA molecules. These synthetic approaches have been applied mainly to gene expression control to degrees that include multiplex fine-tuning of gene expression for metabolic engineering purposes (Vazquez-Anderson and Contreras 2013). Specifically, bacterial small RNAs (sRNAs) have seen an enhanced interest since they bear this multiplex capacity (Vazquez-Anderson and Contreras 2013). sRNAs are versatile regulators with potential to harness full regulatory networks for the production of phenotypes of interest with a conceivable impact that spans from production of biotechnological compounds to bacterial virulence control.

In order to increase the success rate of these applications, it is necessary to understand, characterize and predict the ability of RNA to interact with other molecules, namely, its *functional structure*. There is currently no approach aimed at understanding the capacity of RNA to interact with another molecule. We have termed this ability *structural accessibility:* a direct measure of the availability of a given RNA region within a target molecule to interact and establish binding with another molecule (also known as hybridization efficacy). Throughout my PhD, I have engaged in pursuit for the development of tools to characterize and predict structural accessibility as a measure of RNA functional structure.

The subsequent section introduces the state of the research in three key areas of my work: (1) current techniques to characterize RNA structure, (2) current approaches to predict hybridization efficacy and (3) current progress in understanding functional structure in regulatory RNAs. Each section also introduces each molecular tool in more detail.

1.2 CURRENT TECHNIQUES TO CHARACTERIZE RNA STRUCTURE AND THE IRS[3]

Biotechnological applications of RNA have exploded in recent years, amplifying the need to develop tools to better understand RNA folding dynamics and structural changes. RNA structures have been extensively studied using a variety of *in vitro* techniques (Tijerina, Mohr et al. 2007, Scott and Hennig 2008, Edwards, Garst et al. 2009). Prominent among these techniques is the use of chemical or enzymatic modifications (e.g. DMS, hydroxyl radicals, metal ions, RNase, S1 nuclease mapping) to map RNA structures (Wurst, Vournakis et al. 1978, Shcherbakova and Brenowitz 2008, Wan, Suh et al. 2010). While these methods have provided valuable structural data, most do not provide information on RNA folding dynamics or structural changes *in vivo*.

RNA folding is influenced by a complex cellular milieu that is difficult to replicate *in vitro* (Emerick and Woodson 1993, Zemora and Waldsich 2010). Cellular factors such as the speed and directionality of transcription, metabolite levels, RNA localization and other bimolecular interactions can all have a significant impact on the acquisition of native RNA structures (Emerick and Woodson 1993, Zhang, Ramsay et al. 1995, Schroeder, Grossberger et al. 2002, Zemora and Waldsich 2010). With these considerations in mind, a few groups have developed protocols based on chemical modification for characterizing RNA structures *in vivo* (Wells, Hughes et al. 2000, Lindell, Romby et al. 2002, Kertesz, Wan et al. 2010, Spitale, Crisalli et al. 2013). One of the most recent examples of *in vivo* RNA structural probing is Selective 2'-Hydroxyl acylation Analyzed by Primer Extension (SHAPE) in living cells (Spitale, Crisalli et al. 2013). This technique adapts traditional chemical probing to an *in vivo* setting.

A common feature of chemical probing techniques is that they rely on non-targeted modification of RNA molecules. Since the chemical probe does not modify a unique sequence within the molecule, chemical footprinting methods are less likely to detect transient differences within specific regions that hallmark rare folding intermediates (Wan, Kertesz et al. 2011). These rare folding intermediates can have alternative functions when compared to the final structure and can be important to understand RNA folding pathways (Lai, Proctor et al. 2013, Grohman, Gorelick et al. 2014). In order to detect these intermediates *in vivo*, a more targeted approach would be required.

In this work, we demonstrate the novel ***i**n vivo* **R**NA **S**tructural **S**ensing **S**ystem (iRS^3) for probing RNA structures *in vivo*. Our design exploits the ability of a previously-designed, well-studied riboregulator to control green fluorescent protein (GFP) expression post-transcriptionally (Isaacs, Dwyer et al. 2004, Vazquez-Anderson and

Contreras 2013). The fundamental premise of this approach is that highly structured areas are physically blocked from binding the designed structural reporter. In contrast, "open" regions that do not participate in any intra- or inter-molecular contacts or that are simply not hindered by the topology of the molecule will be more readily available to bind to the reporter. Hereby, we present the iRS^3 as a useful tool, not based on chemical modifications, capable of probing RNA structure in living cells.

To demonstrate the value of this system, we use the iRS^3 to explore the structural organization of the Tetrahymena group I intron (gI intron). This gI intron is a well-studied (~400 nt) catalytic RNA (Kruger, Grabowski et al. 1982, Koduvayur and Woodson 2004, Wan, Suh et al. 2010) that has been structurally characterized by a variety of different *in vitro* techniques (Cech, Damberger et al. 1994, Kieft and Tinoco 1997, Golden, Gooding et al. 1998, Russell, Zhuang et al. 2002). The gI intron has also been confirmed to be catalytically active when expressed heterologously in *E. coli* (Waring, Ray et al. 1985, Zhang, Ramsay et al. 1995). We establish the ability of our system to distinguish between the wild type and two mutant introns and to identify some of the most accessible regions of each intron. When compared to all available DMS and hydroxyl radical footprinting data (including our own *in vivo* DMS data), results from our iRS^3 probing revealed a higher potential to detect low abundance folding intermediates. As such, the iRS^3 methodology complements other *in vivo* and *in vitro* probing methods based on small molecule accessibility.

1.3 CURRENT APPROACHES TO PREDICT HYBRIDIZATION EFFICACY AND THE INTHERACC APPROACH

In vivo RNA targeting via antisense base pairing provides an efficient mechanism to characterize RNA interactions as well as to post-transcriptionally regulate gene

expression. In the native cellular environment, sequence-specific antisense RNAs (asRNAs) are ubiquitous in natural gene regulatory mechanisms, ranging from bacterial small RNAs (sRNAs) (both cis- and trans-encoded) (Georg and Hess 2011, Vazquez-Anderson and Contreras 2013, Cho, Haning et al. 2015) and circular RNAs (Memczak, Jens et al. 2013) to more complex eukaryotic systems such as the RNA interference (RNAi) pathway. Likewise, the common use of affinity-based purification assays to characterize *in vivo* RNA interactions (i.e. pulldown of a target RNA and its interacting partners from cellular extracts) relies on targeting the RNA of interest with an immobilized bait molecule, often an antisense RNA (Srisawat and Engelke 2002, Faoro and Ataide 2014). Furthermore, the simplicity and universality of nucleic acid Watson-Crick complementarity makes antisense nucleic acids highly attractive for controlling gene expression (Coleman, Green et al. 1984, Chan, Lim et al. 2006, Bennett and Swayze 2010, Vazquez-Anderson and Contreras 2013, Haning, Cho et al. 2014) in biotechnological applications such as bacterial cellular engineering (Nakashima and Tamura 2009, Yoo, Na et al. 2013, Nakashima and Miyazaki 2014, Chae, Kim et al. 2015). Given the broad utility of RNA targeting via antisense binding, recent efforts to design effective synthetic antisense RNAs (asRNAs) in bacteria have become more systematic, mimicking mechanisms of natural non-coding RNAs that downregulate their cognate messenger RNAs (mRNAs) by base-pairing, reviewed in (Vazquez-Anderson and Contreras 2013, Chaudhary, Na et al. 2015, Cho, Haning et al. 2015). A more recent study in bacteria provided general guidelines for the design of asRNAs using large sets of gene-repression data (Hoynes-O'Connor and Moon 2016). However, there remains a significant challenge in the asRNA applications described above: the design of *effective* antisense oligonucleotides for sequence-specific targeting of RNA *in situ* (Faoro and Ataide 2014, Cho, Haning et al. 2015). This is particularly true in bacterial systems, since

most design models for asRNAs have been developed in the context of more complex organisms.

Rational efforts to design asRNA have traditionally been aided by algorithms that predict RNA-RNA interactions (reviewed exhaustively in (Backofen 2014, Lorenz, Wolfinger et al. 2016)). These approaches are numerous and span from simple and fast surveying methods such as GUUGLe (Gerlach and Giegerich 2006)and Blast (Altschul, Gish et al. 1990) that score potential target regions within an RNA of interest using the sole criterion of complementarity, to more sophisticated approaches that use energy-based algorithms to predict joint secondary structures (Lorenz, Wolfinger et al. 2016). Complementarity-based approaches have been followed by several methods that display varying degrees of accuracy and sophistication: from (i) those neglecting intramolecular structure (e.g. RNAduplex(Gruber, Lorenz et al. 2008), RNAhybrid(Rehmsmeier, Steffen et al. 2004), TargetRNA(Tjaden, Goodwin et al. 2006) and RNAplex(Tafer and Hofacker 2008)) to (ii) those considering only one interaction site and intramolecular structure (e.g. Nupack(Dirks, Bois et al. 2007), RNAup(Muckstein, Tafer et al. 2006), AccessFold(DiChiacchio, Sloma et al. 2016) and IntaRNA(Busch, Richter et al. 2008)), or even to (iii) those highly computationally complex tools that predict several interactions sites (e.g. IRIS(Pervouchine 2004)) and the joint secondary structure using the energy partition function (e.g. PiRNA(Chitsaz, Salari et al. 2009) and RIP(Huang, Qin et al. 2009)). In contrast, accessibility-based approaches (e.g. RNAup(Muckstein, Tafer et al. 2006) and IntaRNA(Busch, Richter et al. 2008)) have been developed as comparatively simpler tools for prediction of RNA-RNA interactions, as they assume both interacting partners must be unfolded (i.e. accessible) prior to binding (Backofen 2014). In this context, accessibility is defined as the property of a given potential interaction site to be free of intramolecular base-pairs. Target accessibility has been

generally introduced in predictive algorithms as an energy penalty estimated from the ensemble of possible target structures with the corresponding target region unpaired. The specific role of target accessibility in the asRNA hybridization has been extensively studied with a particular focus on miRNAs and siRNAs (Ding and Lawrence 2001, Ding, Chan et al. 2004, Muckstein, Tafer et al. 2006, Lu and Mathews 2008, Tafer and Hofacker 2008, Bernhart, Muckstein et al. 2011, Tafer 2014). However, to our knowledge very few works have shed light on how accessibility plays a role in antisense hybridization within living bacteria (Vickers, Wyatt et al. 2000). Furthermore, there is limited accuracy of the aforementioned structure prediction approaches (e.g. high false positive rate (Backofen 2014) and limited accuracy (70% for molecules up to 500 nt and as low as 40% for longer RNAs) (Mathews, Burkard et al. 1999)) due to simplifications in the energy model that overlook intracellular factors that affect hybridization. This underscores the need for more *realistic* approaches that account for the *in vivo* environment, incorporating the influence of differences across domains of life, binding factors, ionic strength and molecular crowding (Leamy, Assmann et al. 2016).

Hereby, we propose a novel approach to predict and evaluate hybridization efficacy in bacteria that features the inclusion of large sets of experimental data collected *in vivo*. This model uniquely considers a *regional* availability factor. *Regional* characteristics of the target RNA have long been implicated in asRNA efficacy. For instance, Zhao and Lemke proposed a criterion that at least 4 highly accessible nucleotides are necessary for the initiation of asRNA- target RNA binding based on investigating correlations between predicted structure and asRNA efficacy (Zhao and Lemke 1998). In addition, established mechanisms of RNA molecular recognition, such as the existence of an intermediate step in which a few nucleotides interact to initiate the binding, termed *seeding interaction*, or even the existence of recognition sequences that

act as first "points of contact" such as the YUNR motif (Lucks, Qi et al. 2011) further supports this notion of "regionality" (Rodrigo, Landrain et al. 2012) . To derive the corresponding model, we start from a common thermodynamic framework used in accessibility-based approaches (Backofen 2014) that considers the overall change of free energy of Gibbs ($\Delta G_{overall}$) in the reaction system, a predictor of asRNA binding (19). Lastly, we introduce a novel consideration of target accessibility as a combination of the local unfolding of the target region based on the minimum free energy structure and a measure of the availability of a region (cohesive stretch of nucleotides) to be a "point of contact" in an intermolecular interaction based on the ensemble of suboptimal structures. The latter measure, the availability factor, is estimated from the average of base-pairing probabilities over the length of the target region. Hereafter, we refer to this predictive approach as the **in** vivo optimized **Ther**modynamic **Acc**essibility-adjusted model, **inTherAcc**.

The inTherAcc model was developed using large data sets describing *in vivo* hybridization efficacy of asRNAs targeting approximately 80 regions (i.e. a continuous stretch of 8-27 nucleotides) within 3 well-studied RNA molecules: the autocatalytic group I (gI) intron from *Tetrahymena*, the global small RNA regulator CsrB, and the glutamate tRNA in *E. coli*. Statistical interactions among the predictive parameters, discovered via regression analysis, were investigated and a subset of experimental data collected (29 asRNAs) was utilized for model *optimization*. Experimental characterization of asRNA hybridization efficacy for model optimization purposes was performed using a previously published *in vivo* RNA probing assay that measures asRNA-target RNA hybridization via fluorescence: ***i**n vivo* **R**NA **S**tructural **S**ensing **S**ystem-(IRS^3) (Sowa, Vazquez-Anderson et al. 2015).

Following model optimization with experimentally collected data, the model was used to predict asRNA hybridization efficacy of numerous regions in the RNA 2-MS2 phage coat protein transcript (2-MS2), glgC 5' UTR (glgC), group II intron (gII), and the Spinach II (SpII) RNAs. Collection of experimental data on hybridization efficacy in these regions allowed evaluation of the optimized model. The performance of our model was benchmarked against its *in vivo*-optimized thermodynamic-only version (which lacks the availability term) and the computational tool IntaRNA (Busch, Richter et al. 2008), an accessibility-based approach that also considers a *regional* adjustment by incorporating the existence of a user-definable seed that has been proven useful in predicting RNA-RNA interactions in bacteria.

Lastly, since bridging the gap between the discovery of sRNAs, and the identification of their corresponding target mRNAs remains a significant challenge, we proposed using inTherAcc to predict sRNA-mRNA binding regions. To this end, we compared the ability of the optimized biophysical model coupled to BLAST (Altschul, Gish et al. 1990) to predict mRNA targets of recently discovered *Z. mobilis* sRNAs, Zms4 and Zms6 (Cho, Lei et al. 2014). Experimental confirmation using RIP-seq data validated the ability of inTherAcc to identify regions within these sRNAs that likely interact with mRNAs via base-pairing complementation and, thereby, the identity of these potential targets. Furthermore, comparison of our results to IntaRNA forecasts, suggests beneficial complementarity between the prediction approaches. Finally, the demonstration of inTherAcc utility in another bacterial species underscores its broad applicability.

Our results demonstrate that inclusion of the cellular milieu significantly improves the prediction of regional hybridization efficacy. The *in vivo* optimization also demonstrated that interplay between the availability factor and energetics of binding

exerts a strong influence on asRNA hybridization efficacy. Altogether, our results show the competence of the proposed biophysical model to assist in determining effective accessible regions to aid asRNA design and sRNA target prediction within a broad range of target RNAs in bacterial systems.

1.4 CURRENT PROGRESS IN UNDERSTANDING FUNCTIONAL STRUCTURE IN REGULATORY RNAS AND THE INTERFACE APPROACH

Recent discoveries have revealed that RNA plays a central role in gene regulation through RNA-RNA (and RNA-protein) interactions(Cruz and Westhof , Sharp , Buratti, Muro et al. 2004, Kozak 2005, Vazquez-Anderson and Contreras 2013, Gu, Xu et al. 2014). Bacterial small RNAs (sRNAs) constitute a distinctive class of RNAs that do not encode proteins but possess intrinsic roles in cellular regulation. When faced with environmental stress (e.g. pH, temperature, osmolarity, nutrient availability), organisms take advantage of the regulatory mechanisms of sRNAs to switch on and off several metabolic pathways, often simultaneously(Wassarman 2002). Most well characterized sRNAs to date control gene expression by binding messenger RNAs (mRNAs). However, several aspects have challenged sRNA characterization studies, rendering the number of sRNAs that are mechanistically understood relatively small (<20 srnaTarBase) relative to the number of sRNAs that continue to be discovered (e.g. >100 in *E. coli* (Li, Huang et al. 2013, Wang, Liu et al. 2015), >500 in *V. cholera* (Liu, Livny et al. 2009)). To name a few, some of these challenges include: (i) relative small sizes (10-25 nt (Peer and Margalit 2011) with 8-9(Gottesman 2004) complementary nt cognate regions(Vazquez-Anderson and Contreras 2013), (ii) ability to bind multiple targets, (iii) in trans expression of their targets, and (iv) differential and complex dependency on the Hfq chaperone for target binding. Indeed, the generality of the role of the Hfq chaperone amongst bacteria remains obscure. Thus far, it has been hypothesized that the Hfq

chaperone acts upon sRNAs via two mechanisms: (1) Hfq binds to both sRNA and target mRNA to facilitate their interaction by unwinding their structural restrictions or (2) Hfq increases local concentrations of sRNA and respective target mRNA to enhance binding(Ishikawa, Otaka et al. 2012). However, these hypotheses have emerged from limited number of sRNAs (~20) that have been extensively analyzed in *E.coli* in the context of Hfq and their target mRNAs (Table C.1).

At the molecular level, RNAs (i.e. sRNAs) regulate gene expression through "accessible" interfaces (hereby defined as regions within an RNA that are able to establish intermolecular interactions with target molecules). Analysis of 107 well-characterized sRNA-mRNA pairs (for 21 sRNAs) shows that many sRNAs bear binding domains that harbor interactions with multiple mRNAs linked to specific regulation during recovery from a particular stress response. These binding domains are expected to be more accessible for intermolecular interactions than random regions per structural prediction analyses(Peer and Margalit 2011). However, given the small set of sRNA-mRNA pairs that have been studied in detail in the context of the intracellular environment (relative to thousands recently proposed(Melamed, Peer et al. 2016)), questions regarding the universality of the observed patterns with regards to regions that are accessible for interactions arise. Likewise, questions regarding the universal nature of the role of Hfq in mediating sRNA-mRNA interactions arise. The enigma concerning the complex role and mechanisms of Hfq in sRNA biology is further sustained by the inability to identify functional Hfq homologues in several gram-positive bacterial species(Romby and Charpentier 2010, Haning, Cho et al. 2014).

To date, many *in vivo* chemical and enzymatic probing methods gauge the accessibility of a region by evaluating the level of "protection" of individual nucleotides within that region and have recently been expanded to genome-wide approaches via

RNA-seq coupling(Silverman, Berkowitz et al. 2016, Strobel, Watters et al. 2016). While this measure of local accessibility reveals important information on secondary structure, it often does not correlate strongly with the more regional-level accessibility that provides additional information about the dynamics of potential functional regions in cells. In recognition of the importance of regional-level accessibility, recent efforts have aimed to understand the ability of RNA to interact with other RNAs and proteins by unveiling the "interactome"(Li, Song et al. 2014, Holmqvist, Wright et al. 2016, Nguyen, Cao et al. 2016), albeit limited to the study of only one class of interaction since these studies require targeting towards one protein as the binding partner. As such, these methods lack the capability to answer global fundamental questions about RNAs such as: (1) What are regions likely to be interacting sites i.e. contributors to the "accessosome"? (2) How do binding partners and cellular conditions affect accessible regions within the structure? and (3) How might the functional structures change in response to stress?

Motivated by the interest in surveying the regional accessibility of a large collection of sRNAs in cells to capture global trends and differences that could add to our understanding of their regulatory mechanisms *in vivo*, in this work we developed INTERFACE, **in** vivo **T**ranscriptional **E**longation analyzed by **R**NA-seq for **F**unctional **A**ccessosome **C**haracterization in a single **E**xperiment. This method is capable of profiling (in one step) regional accessibilities for a large collection of RNAs. Specifically, we employ a combined machine learning and synthetic biology approach to design a large number of oligonucleotides for *in vivo* antisense hybridization that can interrogate all local regions within an RNA landscape. It is worth noting that the notion of correlating *in vivo* hybridization to structural accessibility has been previously demonstrated(Sowa, Vazquez-Anderson et al. 2015), albeit using low-throughput fluorescence-based assays. Importantly, as formerly established, *in vivo* antisense

hybridization bears an ability to detect sites involved in intermolecular interactions and to identify short-lived transient states, often relevant in regulation(Sowa, Vazquez-Anderson et al. 2015)(Vazquez-Anderson J, Mihailovic M, in Review, 2016). As part of this work, we analyzed a collection of 72 trans-encoded experimentally verified sRNAs in *E. coli,* using ~1000 antisense oligonucleotide probes to achieve 100% coverage of all sRNA regions. , A major finding of our studies is the global presence of patterns of extreme accessibility (i.e. low or high) in regions harboring known mRNA-binding sites in a way correlates with the level of "usage" of the region depending on cellular conditions. In addition to extreme accessibility, we unveiled a characteristic strong sequence motif (similar to the YUNR ubiquitous RNA-RNA recognition motif (Franch, Petersen et al. 1999)) and a significant enrichment of the most 5' two thirds of the sRNA molecule as potential predictors of sRNA-mRNA binding sites. We also showed important sensitivity of INTERFACE to capture Hfq influence on accessibility patterns of sRNA as indicated by the strong correlation of our proposed Hfq-dependency per sRNA with quantitative data previously reported in the literature. Finally, we revealed Hfq as a strong universal structural relaxer for the specific subset of sRNAs that depend on the chaperone for regulatory activity, known as Hfq-dependent sRNAs.

1.5 SUMMARY OF RESEARCH OBJECTIVES AND ACCOMPLISHMENTS

The following chapters embody a compendium of the research that I performed at the University of Texas at Austin, collected into three main works that have been published or are near publication.

Chapter 2 is a description of the development of our flagship approach to characterize structural accessibility: the in vivo RNA Structural Sensing System (iRS^3). In this work I closely collaborated with Steven Sowa to prove the concept of the *in vivo*

oligonucleotide hybridization tool. This tool exploits post-transcriptional regulation to probe RNA structure *in vivo*. Briefly it consists of a variable sequence termed probe that is complementary to a region within a target RNA that we wish to characterize that in turn controls the translation of green fluorescent protein, functioning as a reporter. We showed that the fluorescence signal results specifically from the interaction between the probe and the target region. Next we studied several regions within the *Tetrahymena* group I intron (gI intron) as well as two other gI intron mutants using the iRS^3. Collectively our results suggest that the iRS^3 is a direct measure of structural accessibility and that it differs from other chemical probing methods in that it can capture dynamic regions.

In chapter 3 I recount the construction of a biophysical approach to predict hybridization efficacy, performed in collaboration especially with Mia Mihailovic and Kevin Baldridge. The novelty of this work partially lies in incorporating target accessibility as a linear combination of: (1) regional energy cost for disruption of the binding site and (2) considering the base-pairing probabilities of the global structure as a regional availability factor (as a measure of the ability of a region to establish binding with other molecules), termed pseudo collision probability factor, in the equilibrium derivation that in turn accounts for potentially neglected *in vivo* interactions. Arguably the most important contribution however, was model optimization using extensive *in vivo* data collected for the gI intron, CsrB and glutamate tRNA. The resulting model allows incorporation of binding factors, which were not previously accounted for, in the regional pseudo collision probability factor. Next, the model was exhaustively tested using standard cross-validation schemes and a set of four new RNA molecules whose results were compared to an appropriate benchmark. Altogether our results suggest that the performance of our biophysical approach, the in vivo-optimized Thermodynamic

Accessibility-adjusted model (inTherAcc), is at least comparable to benchmark IntaRNA. Lastly we observed an advantage over the benchmark as per linear fits of our predictions versus the experimental results in the cases of complex large RNA molecules such as group II intron.

In Chapter 4, I present the high throughput tool INTERFACE: **in** vivo **T**ranscriptional **E**longation analyzed by **R**NA-seq for **F**unctional **A**ccessosome **C**haracterization in a single **E**xperiment. In this work, with help from Mia Mihailovic, I coupled *in vivo* oligonucleotide hybridization to transcriptional elongation for the high-throughput characterization of accessible interfaces via RNA-seq. First, I demonstrate that transcriptional elongation control correlates to structural accessibility. Next, I showcase the scope of the approach by characterizing about 1000 regions within 73 regulatory RNAs. To this end, I first coupled a version of the inTherAcc model to a machine-learning algorithm developed by a collaborator in Princeton, with the purpose of selecting likely accessible regions. I show that using this selection scheme represents an advantage over random selection (pure exploration) and only using the model (pure exploitation). With these suggestions I used INTERFACE to fully characterize the sequence universe of the small RNA regulatory network in *E. coli* and in an Hfq-deficient strain of *E. coli*. The results showed that many functional regions are identified as highly or lowly accessible. Finally this work is the first to show global patterns of accessibility for functional sites and the global effect of Hfq on sRNAs as a structural relaxer. We plan to file a patent application for this work.

Chapter Two

*Exploiting post-transcription regulation to probe RNA structures in vivo via fluorescence

*This work was published in (Sowa, Vazquez-Anderson et al. 2015)

2.1 INTRODUCTION

While RNA structures have been extensively characterized *in vitro*, very few techniques exist to probe RNA structures inside cells. Here, we have exploited mechanisms of post-transcriptional regulation to synthesize fluorescence-based probes that assay RNA structures *in vivo*. Our probing system involves the co-expression of two constructs: (1) a target RNA and (2) a reporter containing a probe complementary to a region in the target RNA attached to an RBS-sequestering hairpin and fused to a sequence encoding the green fluorescent protein (GFP). When a region of the target RNA is accessible, the area can interact with its complementary probe, resulting in fluorescence. By using this system, we observed varied patterns of structural accessibility along the length of the Tetrahymena group I intron. We performed *in vivo* DMS footprinting which, along with previous footprinting studies, helped to explain our probing results. Additionally, this novel approach represents a valuable tool to differentiate between RNA variants and to detect structural changes caused by subtle mutations. Our results capture some differences from traditional footprinting assays that could suggest that probing *in vivo* via oligonucleotide hybridization facilitates the detection of folding intermediates. Importantly, our data indicates that intracellular oligonucleotide probing can be a powerful complement to existing RNA structural probing methods.

* In this work I am a leading author contributing to 50% of all research done in collaboration with Steve Sowa.

2.2 RESULTS

2.2.1 Molecular design and optimization of iRS^3

Our design is an alteration of a previously published, highly-controllable riboregulator that inhibits the synthesis of GFP in the presence of a regulatory hairpin and promotes the synthesis of GFP in the absence of the same hairpin (Isaacs, Dwyer et al. 2004). As illustrated in **Figure 2.1**, this new structural reporter is comprised of 5 segments at the RNA level: (1) a specific 15-18 nucleotide sequence (probe) that is complementary to a specific region of the target RNA sequence, (2) a *cis*-blocking (CB) region complementary to the Ribosome Binding Site (RBS), (3) a linker region (LR) in between the RBS and its complement, (4) a region containing an RBS, and (5) a region encoding GFP. The probe, which is the only variable feature of this system, acts as a sensor for a specific region within the target RNA. The RBS, LR, and the CB form a stable hairpin. The stability of this hairpin is controlled by the expression of a separate molecule (e.g. the Tetrahymena gI intron) in *trans* that can base-pair to the probe and destabilize the hairpin, most likely by steric hindrance and/or by subsequent structural reconfigurations upon probe binding. In addition, once the hairpin is opened, the iRS^3 transcript could be further stabilized by interactions with the ribosome; similar protection effects have been reported (Contreras, Huang et al. 2013). Thus, the hairpin acts as an adaptor converting the extent of the RNA-probe interaction into fluorescence readout.

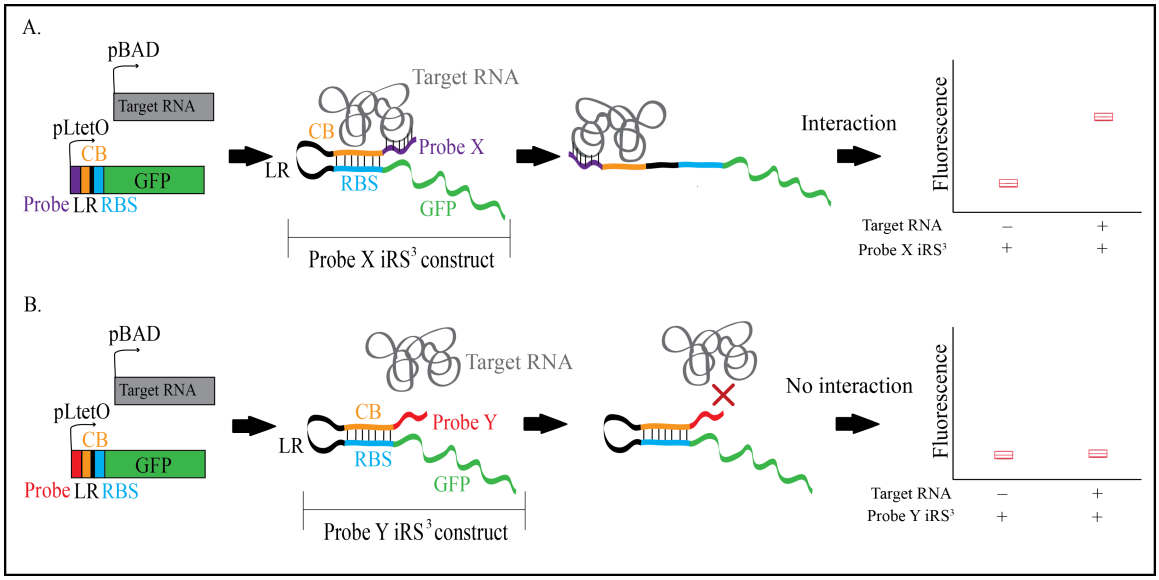


Figure 2.1. Fundamentals of the *in vivo* RNA Structural Sensing System (iRS^3).

(A) Accessible region on target RNA. Our system is expressed on a plasmid using two promoters, pBAD and PLtetO. The RNA reporter construct (Probe X iRS^3 construct) incorporates a GFP transcript (green) whose translation is inhibited due to ribosome binding site (RBS, blue) sequestration by a cis-blocking region (CB, orange) that is connected to the RBS through a flexible linker region (LR, black). If the probe (purple) targets an accessible region on the target RNA, an interaction will occur causing the hairpin loop to open, exposing the RBS, and lead to GFP expression. **(B) Inaccessible region on target RNA.** If the probe (red) targets an inaccessible region on the target RNA, there will be reduced interaction between the intron and the probe, the hairpin loop will not open and a negligible increase in fluorescence will be observed compared to non-induced levels.

We initially tested two designs for the riboregulator. The first design contained a NotI restriction site between the probe and the CB-RBS hairpin, which was intended to simplify cloning. However, we did not observe a significant shift in fluorescence upon

induction of the intron (data not shown). As a result, we created a second design which contained the probe immediately adjacent to the CB-RBS hairpin. This probe design allowed us to detect a shift in fluorescence upon induction of the target and was therefore used for the remainder of our analyses.

To determine an appropriate length for the complementary probes, we carried out binding predictions (see Methods: Computational Analysis of Probes) and discovered that 15-18mers gave a more stable bound complex than 8-12mers (data not shown). Therefore, we chose to continue our studies with 15-18mers given their higher specificity, stronger binding, and their ability to provide more intron coverage for our initial studies. Although we proceeded with 15-18mers, our computational predictions suggested that 8-12mers could provide sufficient, albeit weaker, binding to be used in these studies to increase the structural resolution of the system. From these preliminary studies, we developed a general methodology for designing the iRS3 system to target different RNAs (Figure A.1).

2.2.2 iRS^3 fluorescence is specific to the interaction between probe and target RNA

After determining an appropriate design for the iRS^3 (**Figure 2.1**), we then built several controls to verify that the fluorescence observed was specific to the recognition and binding of the probe to the target gI intron. All controls were built in the context of Probe 1(**Table A.1 in (Sowa, Vazquez-Anderson et al. 2015)**). We hypothesized that a shift in fluorescence would result from the 5' target region binding to its complementary probe-iRS^3 reporter. The binding location of Probe 1 on the intron (and all other probes used) is shown in **Figure 2.2**.

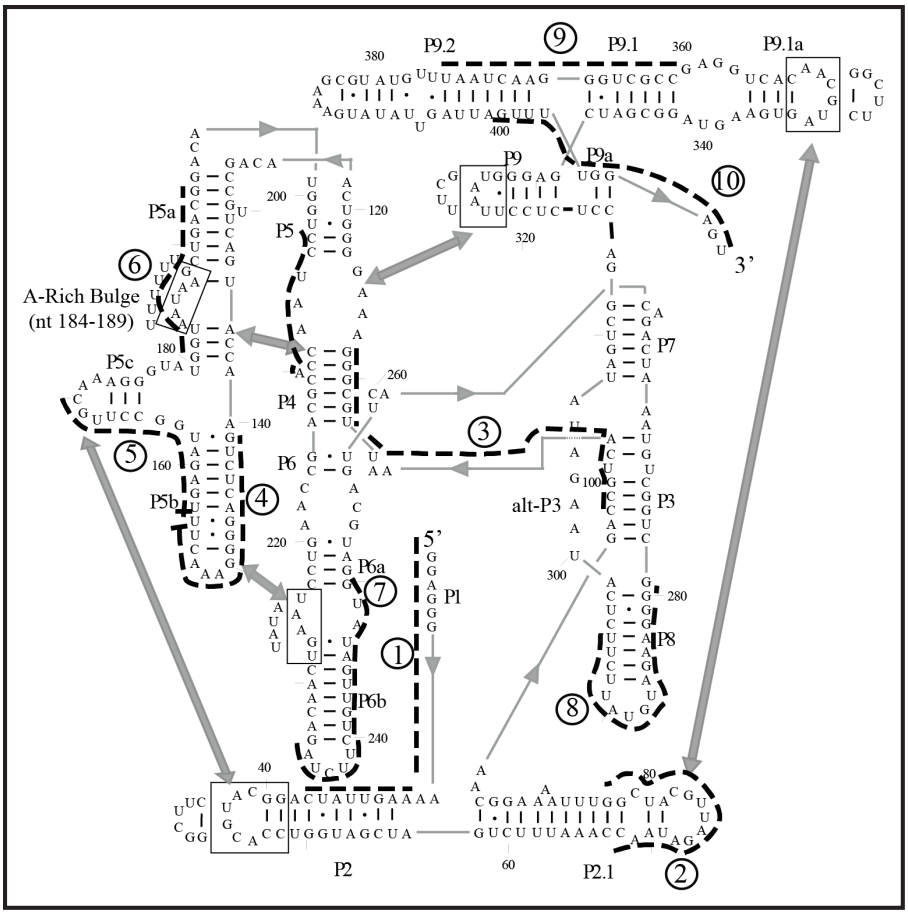


Figure 2.2. GI intron *Tetrahymena* Ribozyme Model System.

This figure depicts the sequence and secondary structure of the *Tetrahymena* wild-type gI intron. Probe numbers are circled and indicated next to black dashed lines, which show the regions of the molecule being targeted by these complementary probes. Structural domains are indicated by the letter "P" followed by a number. The dashes between the nucleotides on opposing sides of each stem loop indicate complementary base pairing, with the dots signifying G-U wobble base pairing. The quintuple mutant contains a total of five mutated regions (black boxes) and the letters outside of the boxes represent the new mutant sequences. These mutations abolish five key tertiary contacts within the group I intron molecule shown as thick gray arrows. The A-rich bulge mutant has the same A-rich bulge mutation as the quintuple mutant, but contains no other mutations.

Using Probe 1, we tested if a fluorescence shift was specific to the intracellular presence of both the target gI intron (expressed by the pBAD promoter) and the iRS[3] transcript (expressed by the pLtetO promoter). For these experiments, we inoculated cells harboring the plasmid construct containing the target gI intron RNA and the iRS[3] construct. We conducted flow cytometry assays under inducing (presence of intron) and non-inducing (absence of intron) conditions and ran each experiment in at least quadruplicates. For all experiments, we defined a fluorescence shift as the difference in fluorescence between means of induced and the non-induced replicates, five hours after inducing the appropriate samples. As illustrated in **Figure 2.3**, significantly more fluorescence is observed only when the reporter transcript is expressed in the presence of the target gI intron. When the Probe 1-iRS[3] reporter and the wild-type intron constructs were co-expressed in a ΔaraC knockout strain (where the pBAD promoter cannot be activated) no appreciable shift was detected (**Figure 2.3B**). This experiment demonstrated that induction of intron expression is required to achieve a significant fluorescence shift.

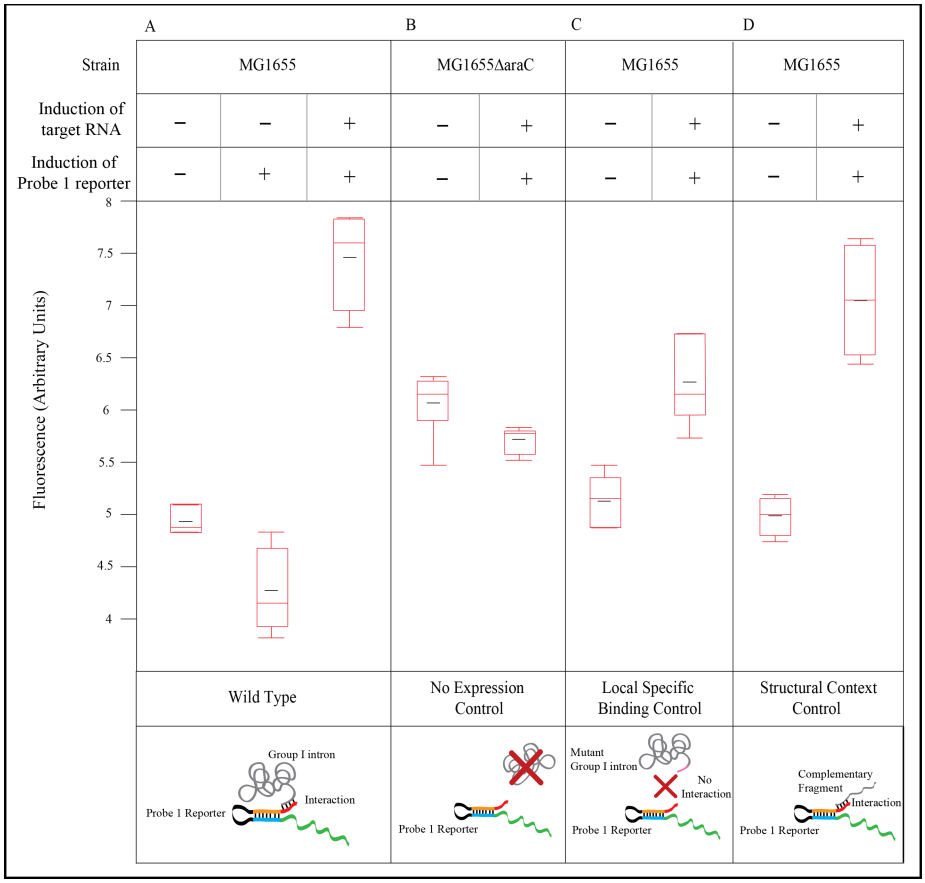


Figure 2.3. Fluorescence shifts result from specific interactions between the reporter and the *trans* target RNA.

Each sample contains the Probe 1 IRS[3] construct co-expressed with a different target RNA. (A) The wild type sample shows the interaction between wild-type intron and the IRS[3]-Probe 1 construct. (B) The wild-type intron and the IRS[3]-Probe 1 transformed into a ΔaraC knockout strain to impair expression of the group I intron. (C) The IRS[3]-Probe 1 construct with an intron where 9/16 nucleotides in the target region are mutated, such that Probe 1 is no longer as specific to its target region. (D) The IRS[3]-Probe 1 construct co-expressed with a shorter target RNA that contains a complementary sequence to Probe 1. The box plots represent the 75% quartile (upper red line), the median (middle red line), and the 25% quartile (lower red line) of the median fluorescence of at least quadruplicate samples. Whiskers above and below the box plot indicate the furthest data point that is within 1.5x the interquartile range from the box.

We then tested if fluorescence was specific to the binding of the Probe 1-iRS[3] reporter to the targeted 5' end region of the intron. For this experiment, we designed a mutant gI intron where the local binding region to Probe 1 was altered by mutating a 9 base pair stretch in the center of the 16-mer target area ($GGGAAAAGT_{25\text{-}33}$ $\rightarrow CCCTTTTCG_{25\text{-}33}$) coupled with compensatory mutations ($GCTA_{54\text{-}57}\rightarrow$ $TGAT_{54\text{-}57}$) to preserve the native secondary structure. As shown in **Figure 2.3C**, mutating the target region on the intron resulted in a smaller shift in fluorescence upon induction relative to the wild type gI intron-Probe 1 system. These results indicated that the interaction between the wild-type gI intron and the Probe 1 reporter occurs specifically within the targeted region and, that it was necessary for a significant shift in fluorescence. The residual shift shown by the mutant intron can be explained by the few remaining nucleotides of complementarity that are located at each end of the mutated sequence.

Lastly, we tested the dependency of observing a fluorescence shift on the structural context of the targeted area. For this experiment, we built a much smaller transcript (~100 nt compared to ~400 nt) that mimicked the gI intron by containing the complementary region to Probe 1. The strong interaction we observed suggests that even the smaller target RNA was sufficient to destabilize the hairpin (**Figure 2.3D**), as long as the binding sequence was specific and present. In this way, we demonstrated that the probe can still bind outside of the molecular context provided by the gI intron. Importantly, these results implied that the use of this system could be extended beyond large molecules such as the gI intron.

2.2.3 iRS^3 can discriminate between accessible and protected regions along the group I intron

After identifying a sensitive molecular design that led to fluorescence in the presence of specific binding between the Probe 1 reporter and an accessible region of the gI intron, we tested if the iRS^3 could capture contrast in structural accessibility along the target gI intron. For these experiments, we designed nine additional constructs with different probing regions along the gI intron (**Figure 2.2**, **Supplementary Tables 1, 3 in (Sowa, Vazquez-Anderson et al. 2015)**), resulting in 33% sequence coverage of the gI intron. The probes were numbered sequentially according to their position in the primary sequence of the intron.

We then tested the ability of the iRS^3 system to report on different regions expected have a wider range of accessibility based on *in vitro* results (Zarrinkar and Williamson 1994, Tijerina, Mohr et al. 2007). We focused on the 3' end (corresponding to Probe 10), a more structurally hindered region relevant to the folding of the gI intron catalytic core (Zarrinkar and Williamson 1996) (corresponding to Probe 9), and the P5a domain important to the activation of the ribozyme (Ikawa, Yoshimura et al. 2002) (corresponding to Probe 6). Once we confirmed that each of the four complementary probes could bind the intron *in vitro* (**Figure 2.4A**), we incorporated each probe into the in vivo iRS^3 reporter. As shown in **Figure 2.4B**, using the iRS^3 system, we observed a differentiated pattern where Probes 1 and 6 showed similar fluorescence shift, Probe 10 showed the largest shift in fluorescence and Probe 9 the lowest shift in fluorescence upon intron induction. We also demonstrated using northern blotting analysis that the fluorescence observed from these probes does not correlate with detected levels of the probe-iRS^3 mRNA transcript (**Figure A.2**). These results suggested that the iRS^3 system

could detect differences in accessibility between target regions by differential shifts in fluorescence.

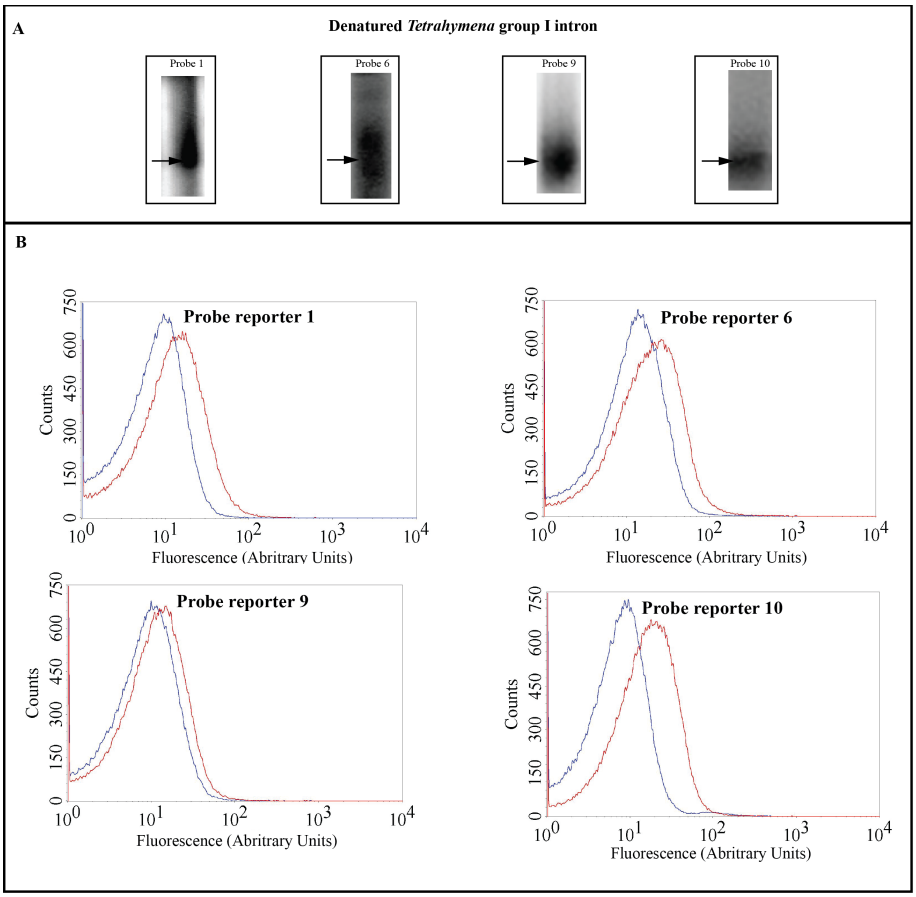


Figure 2.4. Pilot test reveals that fluorescence assay can detect relative levels of accessibility.

(A) *In vitro* binding assays show that the probes can bind to the gI intron. Total cellular RNA containing the gI intron was extracted from cells, denatured and then hybridized to the corresponding P^{32} labeled probe. The resulting hybridized mixture was loaded onto a native polyacrylamide gel, dried, and imaged. Black arrows point to the location of the gI intron band. (B) Representative flow cytometry curves from cells harboring the gI intron and the IRS^3 construct non-induced (blue) induced (red) samples after 5 hours.

It is important to note that prior to testing the different reporters, we performed a computational analysis using the NUPACK (Zadeh, Steenberg et al. 2011) software suite to test for self-dimerizations or hairpin formation (see Methods: Computational Analysis of Probes). We also used NUPACK to test for the strand-to-strand binding affinity between the iRS^3 reporters and the target wild type intron. All probes designed and tested in this study were predicted to bind with the target region on the intron without many side interactions (data not shown), albeit with relatively different strength. As a result, we expected that each iRS^3 reporter probe needed to be individually normalized. That is, as confirmed experimentally, each probe affected the intrinsic stability of the hairpin in the iRS^3 construct differently and lead to varying baseline levels of fluorescence. Differential baseline levels of fluorescence observed between probes can also be explained by the presence of a low amount of gI intron (confirmed by northern blotting analysis, data not shown), even under non-inducing conditions.

2.2.4 Assaying a probe library along the group I intron

After confirming that the iRS^3 could discriminate between different levels of accessibility along the target RNA by displaying differential levels of fluorescence, we tested if fluorescence shifts represented a good measure of accessibility across a wide range of probes. We confirmed the binding capabilities of all probes to the gI intron by conducting *in vitro* binding assays in which we hybridized 5'P^{32} labeled probes to denatured total RNA extracted from *E. coli* cells over-expressing the gI intron (**Figure A.3**). After we saw that the probes could bind to the denatured intron, we incorporated the remainder of our designed probes into the structure sensing system and co-expressed each construct with the gI intron in cells. According to previous *in vitro* studies: (i)

Probes 3, 4, 5, and 6 target important regions for the catalytic activity of the intron (Strobel and Shetty 1997, Naito, Shiraishi et al. 1998), (ii) Probes 2, 5 and 6 target key tertiary contacts (Das, Kwok et al. 2003), (iii) Probe 7 targets the P6a and P6b domains of the gI intron (Wan, Suh et al. 2010), and (iv) Probe 3 (domains P3 and P4) targets a heavily base-paired region likely in the interior of the molecule and thus relatively hindered (Zarrinkar and Williamson 1994). By selecting these key areas, we anticipated a wide representation of accessible and inaccessible regions as well as biologically-relevant areas within the intron.

As shown in **Figure 2.5A**, we observed meaningful differences in the accessibility of the ten regions probed. We calculated statistical error using the standard error of the mean (SEM) for the fluorescence shift, as propagated from the SEM of multiple determinations (≥ 4) of fluorescence at non-inducing and inducing conditions. We determined that accessibility of regions was statistically different from each other when the means of our observations differed by at least two standard errors. Using this highly stringent metric, we concluded that Probes 3 and 7 were significantly more exposed than Probes 2, 4, 5, and 9. This observation supported that the iRS^3 system can discriminate between exposed and protected regions. We then categorized each region as exposed, protected, or in between based on comparing each fluorescence shift to the median accessibility of all regions (**Figure 2.5B**). To determine if the inherent thermodynamic properties of each probe could cause the observed fluorescence patterns, we plotted the minimum free energy (MFE) of the bound complex between the probe sequence and the specific target region versus the normalized fluorescence shift. As can be seen from **Figure A.4**, there is no significant bias in the binding affinity of the probes (note that the slope of the trend line and the R^2 value approximate to zero and, points are randomly distributed around the trend line).

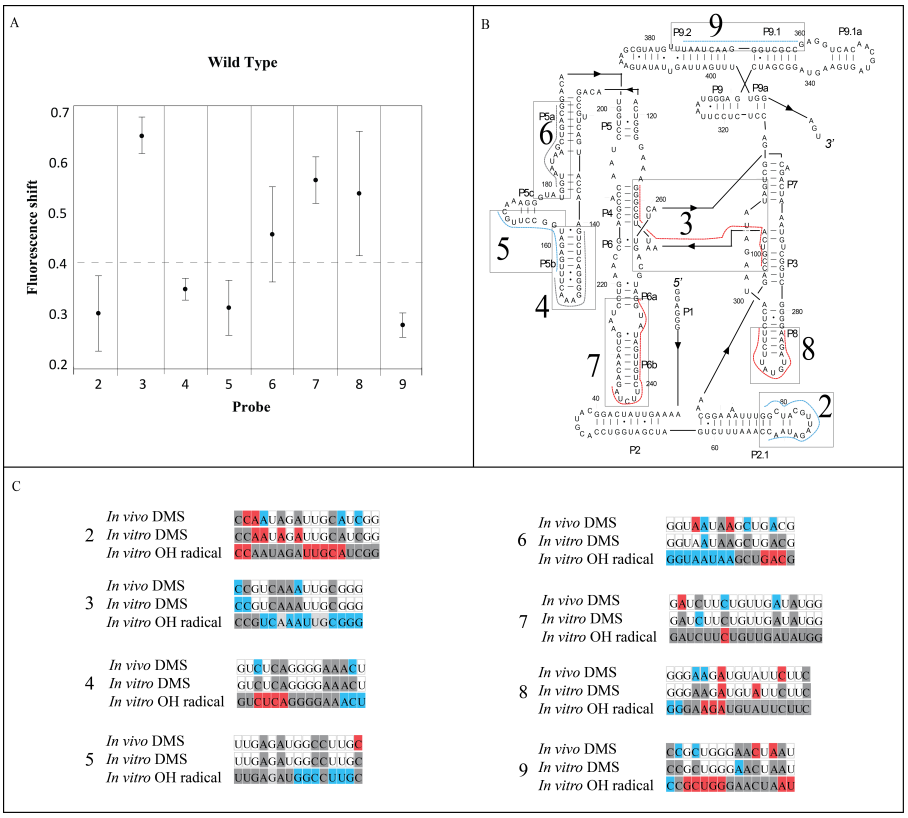


Figure 2.5. iRS3 can capture difference in accessibility along the length of the wild-type group I intron.

(A) Fluorescence shifts resulting from probing different regions along the length of the group I intron. (B) Map of the wild-type Tetrahymena gI intron marked with the relative accessibility for regions as determined by comparison of their fluorescence shifts. The data is separated into two general categories with regions showing protection (blue) or higher accessibility (red). Regions whose fluorescence shift mean falls within one standard deviation of the median in Panel A were considered in the in the middle and marked in gray. (C) Footprinting data for regions assayed by probes. *In vivo* DMS data, performed as part of this work, was discretized. The *in vitro* DMS and hydroxyl radical footprinting data was adapted from (Russell, Das et al. 2006) using an equivalent discretization scheme. All chemical reactivities were normalized by the global average reactivity.

To explain the general accessibility of each target region, we conducted an *in vivo* DMS footprinting analysis covering approximately 92% of the entire intron (except about 35 nucleotides at the 3' and 5' ends of the intron). Additionally, we compared the iRS^3 results to structural studies of the group I intron reported in the literature (Lehnert, Jaeger et al. 1996, Russell, Das et al. 2006, Tijerina, Mohr et al. 2007, Wan, Suh et al. 2010). A particularly useful study was performed by Russell and colleagues when they explored the structure of the group I intron using DMS footprinting and hydroxyl radical footprinting *in vitro* (Russell, Das et al. 2006). We found reliable consistency between *in vivo* and *in vitro* DMS footprinting of the group I intron structure when looking at the overall protection level of each area targeted by our different probes (**Figure 2.5C, Figure A.5** in (Sowa, Vazquez-Anderson et al. 2015)). In general, our iRS^3 reporters show some agreement with the *in vitro* hydroxyl radical footprinting (Russell, Das et al. 2006) and DMS footprinting data (**Figure 2.5**). Specifically, the region targeted by Probe 5 appear to be protected, Probes 4 and 6 targeted regions appear to be moderately protected and exposed respectively and, Probes 7 and 8 regions appear to be more exposed when doing a qualitative assessment of all three footprinting patterns (**Figure 2.5C**). Overall, we noted two major differences when comparing **Figures 5B** and **5C**.

First, in general the footprinting results generally estimate an overall higher exposure level (**Figure 2.5C**) for the region targeted by Probe 9 than iRS^3 determinations (**Figure 2.5A**). This difference is likely due to the region being a stable and heavily based paired helix (31) potentially more difficult to be disrupted by oligonucleotide hybridization than modified by chemical probes. The discrepancy between hydroxyl radical footprinting and iRS^3 is reasonably logical given that hydroxyl radical footprinting cleaves the RNA phosphodiester backbone and this cleavage is less

influenced by base pairing in the intron structure (1,39). While DMS footprinting identified a few exposed nucleotides in the probe 9 target region, this region has enough protected and undetermined nucleotides to preclude drawing conclusions about the global region's accessibility from *in vivo* DMS. These findings strongly suggest that the iRS^3 is a measure of global accessibility that may provide different information, specifically, how available an entire region is to form base pairing interactions.

Second, iRS^3 accessibility results for the region assayed by Probe 3 appear to contradict all footprinting studies in general (**Figure 2.5 and Figure 2.6A**). We reasoned that the greater accessibility of this region to the IRS^3 reporter could reflect an ability of the probe to interact with folding intermediates in which the complementary segment of the intron is exposed. This hypothesis was supported by previous findings that transitions from some folding intermediates to the native form require transient disruption of the long-range P3 base pairs (Mitchell, Jarmoskaite et al. 2013, Mitchell III and Russell 2014). Upon P3 disruption, the 5´ strand of P3 is expected to be accessible to probe 3, while the 3´ strand most likely forms the alternative base pairs alt P3 (Pan and Woodson 1998, Russell, Das et al. 2006). To test whether the accessibility of probe 3 depends on exposure of the P3 region, we split up the targeted area into two shorter target sequences: P3 (nt 95-104 targeted by Probe 3a, 10 nucleotides) and P4 (nt 104-112 targeted by Probe 3b, 9 nucleotides) (**Figure 2.6B**). After demonstrating that these shorter probes bound to the group I intron *in vitro* (**Figure A.3**), we incorporated the probes into the iRS^3 construct. Interestingly, Probe 3a showed an even higher fluorescence shift than Probe 3 while Probe 3b displayed no fluorescence shift (**Figure 2.6C**). We conclude that indeed, the high accessibility to probe 3 likely arises from interaction with the 5´ strand of P3, probably because the probe is able to interact with and trap partially-folded intermediate

(see Discussion). These results also confirmed our ability to obtain higher structural resolution in our system when using shorter probes.

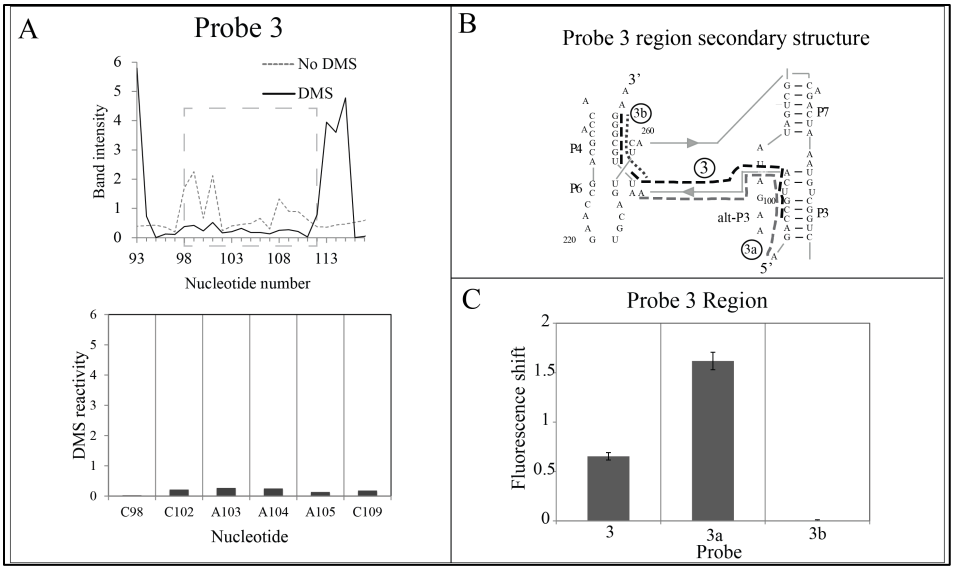


Figure 2.6. iRS[3] differs from *in vivo* DMS footprinting findings in the P3 domain of the gI intron.

Panel (A) shows a relatively low DMS reactivity for the 5' strands of domains P3 and P4 in the gI intron suggesting an overall protected area. The upper plot is a representative footprinting pattern where the dashed box represents the region of interest. The lower plot is a subtraction between "DMS" and "no DMS" band intensities averaged for two independent determinations. Panel (B) illustrates the Probe 3 target area, where dash lines indicate the target segments of each probe and the letter "P" followed by a number indicates the different structural domains. Panel (C) shows the fluorescence shift for Probe 3 (domains P3 and P4) and shorter versions Probe 3a (domain P3) and Probe 3b (domain P4).

2.2.5 iRS^3 can discriminate between group I intron mutants

We also examined if our system could be used to assay structural differences between RNA variants. For these experiments, we compared the wild type gI intron to two intron variants, the quintuple mutant and the A-rich bulge mutant (Naito, Shiraishi et al. 1998). The quintuple mutant contains mutations in five critical tertiary contacts (**Figure 2.2**) that are known to be highly disruptive to the tertiary structure and catalytic activity of the intron (Das, Kwok et al. 2003). Based on this, we hypothesized that the quintuple mutant would exhibit significant differences in accessibility using the IRS^3 relative to the wild-type intron (Das, Kwok et al. 2003). Given that most of the targeted regions of the intron remained unaltered (at the primary sequence level) in the quintuple mutant, we used the same library of probe reporters designed for the wild type gI intron. As for the A-rich bulge mutant, it is a milder variant than the quintuple mutant as it disrupts only one tertiary contact (P5a, **Figure 2.2**).

As shown in **Figure 2.7**, expression of the quintuple mutant results in significantly different shifts in fluorescence (marked with asterisks) relative to the wild type intron for Probes 1, 2, and 7 (**Table A.4 in** (Sowa, Vazquez-Anderson et al. 2015)). From these results, we learned that a couple of areas of the intron become more accessible to oligonucleotides *in vivo* (e.g. domains P1 (Probe 1) and P6ab (Probe 7)), while others become more protected as a result of the quintuple mutations (e.g. domain P2.1, corresponding to Probe 2). Furthermore, the increase in fluorescence observed in the quintuple mutant relative to the wild type intron (~30%) indicates the potential of capturing increased molecular accessibility in the quintuple mutant that results from its lack of tertiary structure relative to the wild type intron (Benz-Moy and Herschlag 2011). On the other hand, the A-rich bulge mutant shows potential differences mostly around the area of mutations, domain P5ac (Probes 4 and 5). The P5abc domain is known to be

important to the stabilization and catalytic activity of the intron (Zarrinkar and Williamson 1994), and it is plausible that mutations affect tertiary contacts in this local area. It is important to note that the 9/10 probes were designed to target regions in the mutants that contain the same sequence as the wild type intron.

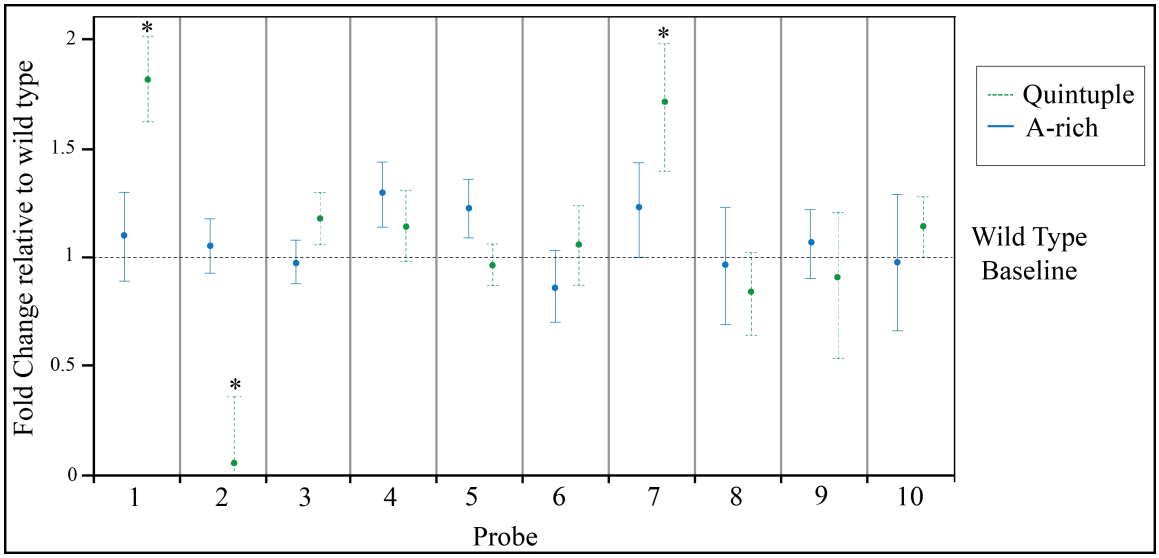


Figure 2.7. iRS[3] can detect different levels of structural disruption in intron variants.

Structural differences between the wild type gI intron, quintuple mutant, and A-rich bulge mutants were measured by calculating the change of the fluorescence shift from the wild type (Fold Change = Fluorescence shift for mutant using Probe X/ Fluorescence shift for wild type intron using Probe X). Standard error of the mean (SEM) for the fold change was calculated as propagated from the SEM for the fluorescence shifts. Differences in accessibility (compared to the wild type) for a given region were considered statistically significant (marked with asterisks) if the mean differed by at least two standard errors from the wild type baseline (indicated as a dashed line).

The area targeted by Probe 3 displayed particular differences in accessibility between the wild type intron and quintuple mutant. Specifically, Probe 3 showed an increased fluorescence shift in the quintuple mutant relative to the wild type intron indicating higher accessibility in this region. This was expected given that Probe 3 targets domain P3-P4, which does not properly base-pair in the quintuple mutant due to the absence of required long range interactions (Zarrinkar and Williamson 1994, Pan and Woodson 1999, Rangan, Masquida et al. 2003, Woodson 2005). Likewise, we captured differences between the wild type intron and the quintuple mutant via IRS^3 fluorescence shifts in the P6b region (targeted by Probe 7). We expected that the P6ab area would be more accessible in the quintuple mutant because this domain normally contributes to a tertiary contact that has been mutated in the quintuple mutant (Wan, Suh et al. 2010). Interestingly, for Probe 7, the A-rich bulge mutant shows a milder difference with respect to the wild type than the one shown for the quintuple mutant. This finding supports the iRS^3 sensitivity to discriminate more subtle structural differences.

2.3 DISCUSSION

In this work, we have combined the traditional idea of using nucleotide accessibility as a measure of RNA structure (Zarrinkar and Williamson 1994, Tijerina, Mohr et al. 2007) with a genetically encoded biosensor to sense that availability. The novelty of our approach lies in the creative implementation of oligo-hybridization probing directly in living cells. In this work, we demonstrated the potential of the iRS^3 to be used as a powerful tool in the study of RNA structures *in vivo*. First, we showed the ability to capture differential structural accessibilities (as defined by base pairing interactions) with high specificity within various local regions throughout the *Tetrahymena* gI intron. We also established the ability of the iRS^3 to capture structural

differences between two variants of the intron. Finally we have also showcased the ability of the iRS[3] to sense milder mutations as it is the case of the A-rich bulge mutant with respect to the quintuple mutant.

Despite the fact that most of the regions of the gI intron exhibited similar structural behaviors *in vivo* and *in vitro*, for this stable model intron, we discovered some regions that behave differently when probed using oligonucleotides in living cells. The most unambiguous example of regions that behave differently when probed with the iRS[3] *in vivo* was the region P3 (assayed by Probe 3), which appeared significantly more accessible by oligonucleotide hybridization *in vivo*, relative to all three standard techniques: *in vitro* hydroxyl radical footprinting and, *in vitro* and *in vivo* DMS footprinting techniques. This difference could exist because the misfolded RNA is not at a high enough concentration to be detected by classical primer extension, but is detectable through oligonucleotide probing.

Figure 2.8 illustrates the fundamental differences in using oligonucleotide probes vs. small molecules for *in vivo* RNA structural probing that can explain increased sensitivity to the detection of low abundance intermediates in our iRS[3] approach. We illustrate the simplest case of a two state folding equilibrium system to represent the dynamics of folding in a region (Region X) presumed to be mostly protected (State 1). However, as expected given dynamic folding equilibrium, other structural conformations are also observed. These structures may be more exposed (e.g. State 2), but appear at a lower frequency (indicated by the larger equilibrium arrow pointing to State 1, **Figure 2.8A**). Based on these dynamics, the potential of modifying Region X (during its less favorable but more accessible equilibrium states) by a small molecule like DMS is rather low (**Figure 2.8B**). This is due to the intrinsic single-hit kinetics of these chemical probing approaches that result in, at most, one modification per molecule. These

modifications can happen at any cytidine or adenosine of the molecule not specifically within the target region. In contrast, an oligonucleotide that has strong complementarity for a target region can bind with high specificity to its target. This oligonucleotide-target interaction can lock less-favorable structures in that state, shifting the equilibrium towards less abundant conformations and giving a signal for exposure that is larger than the exposure of the region in the absence of probe, when exposure is considered as a fraction of the RNA population (**Figure 2.8C**). We suggest that this process underlies the ability of iRS^3 to capture the presence of low level folding intermediates that are accessible to probe 3, whereas these intermediates are not observed in the steady state by DMS or hydroxyl radical footprinting methods (Russell, Das et al. 2006, Wan, Suh et al. 2010). The ability of the iRS^3 to capture folding equilibrium intermediates of importance to the RNA folding pathway makes it a useful complement to current *in vivo* small-molecule based structural probing approaches.

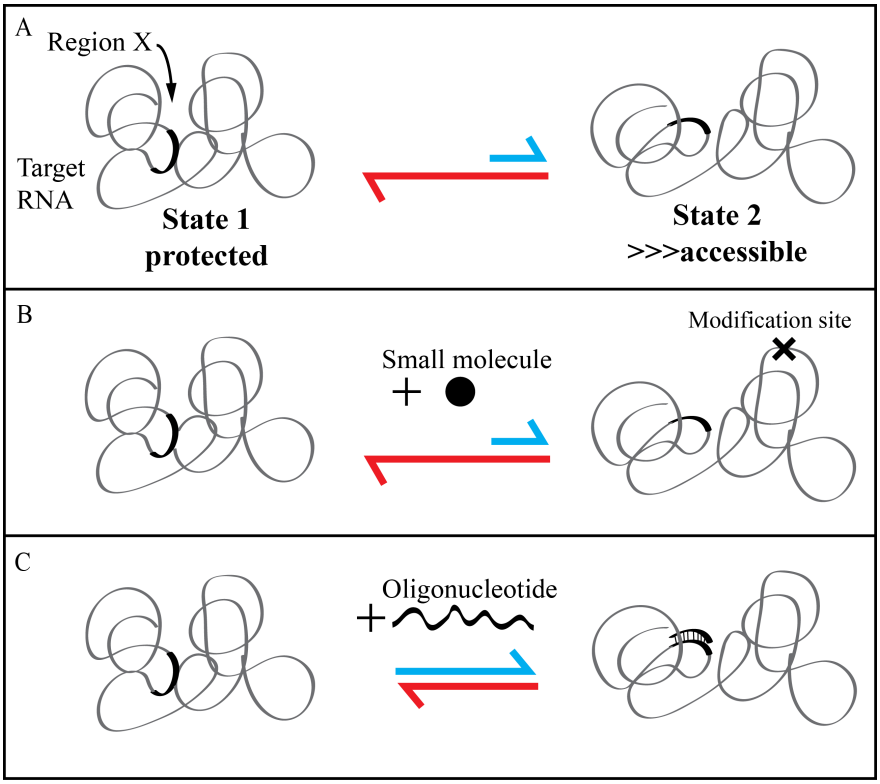


Figure 2.8. Oligonucleotide vs. small-molecule *in vivo* structural probing.

(A) Equilibrium between two conformations of the target RNA. The conformation on the left (State 1) exhibiting relative protection for Region X is more favorable. The less favorable alternative conformation (State 2) on the left exhibits increased exposure of Region X. The equilibrium arrows indicate the relative abundance of each state. (B) The addition of a small-molecule to react with the target RNA results in single-hit kinetics where, on average, at most one specific A or C nucleotide is modified per molecule ("X" representing site of modification); the figure illustrates the low likeliness of modifying nucleotides within Region X. Equilibrium shown is same as described above. (C) An oligonucleotide has an increased probability to hit and bind Region X as it bares full base-pairing complementarity to Region X. Additionally, the oligonucleotide can benefit from capturing the less-favorable State 2 by locking the target RNA at the more exposed conformation, shifting the equilibrium towards the less favorable alternative conformation (illustrated by the longer blue arrow, relative to panels A and B).

We believe that the structural information gained from the iRS^3 can enrich *in vivo* RNA structural studies for a couple of reasons. First, the ability of the iRS^3 to capture folding equilibrium intermediates of the RNA folding pathway makes it a useful complement to current structural probing approaches. Second, iRS^3 probing provides a fundamentally different measure of structural accessibility than chemical methods, namely that accessibility of a region can be interpreted as the availability of a given region to participate in interactions with other macromolecules (i.e. DNA, RNA). Given these capabilities, potential applications of the iRS^3 include detecting alternative structural conformations, and observing the structural effects of tuning RNA-RNA interfaces using mutations.

Another major appeal of this *in vivo* oligonucleotide-based structural probing approach is the ability to design the central riboregulator of the system to sense any target RNA in a rational way. Simple manipulation of the iRS^3 plasmid (e.g. making plasmid compatible for golden gate cloning) should make the system highly amenable for rapid insertion of any probe of choice. Although this system could be incorporated in other organisms beyond *E. coli,* further studies are required to determine if the iRS^{3}system can detect RNAs at native levels. Given these advantages, our method has the potential to provide a relatively easy-to-use platform to capture dynamic structural changes in a wide range of RNAs within living systems.

2.4 Material and Methods

2.4.1 Plasmids and Strains

To build our new *in vivo* reporter system, we modified the intron-harboring plasmids above to contain probes complementary to the gI intron. The pZER21αγ12aG plasmid incorporates two promoters to drive gene expression, the pBAD promoter which expressed the gI intron variants, and the pLtetO promoter which expressed the hairpin-GFP reporter. We designed complementary probes to be inserted immediately upstream of the hairpin-GFP reporter, creating a probe-hairpin-GFP reporter. The cloning strategies for all plasmids in this study are described in **Table A.3 in (Sowa, Vazquez-Anderson et al. 2015)**.

The primary sequence of all constructs was confirmed by using primers E and F (**Table A.2 in** (Sowa, Vazquez-Anderson et al. 2015)) in sequencing reactions performed by the University of Texas core facility. We provide the sequence of the wild type intron Probe 1 reporter (WTI Probe 1 reporter) in **Supplementary Text File 1 (Sowa, Vazquez-Anderson et al. 2015)** and the plasmid is available upon request. The probes (15-18 and 9-10 nucleotides) were designed to be fully complementary to the gI intron (target RNA) and to give good coverage of regions representing the entire length of the intron (**Table A.1 in (Sowa, Vazquez-Anderson et al. 2015)**). All probes were also purchased as oligonucleotide primers (25 nmol, standard desalting) from Integrated DNA Technologies (Coralville, IA).

2.4.2 Computational Analysis of Probes

NUPACK (Zadeh, Steenberg et al. 2011) was used to estimate Minimum Free Energies (MFE) of the probe-reporter (EcoRI-Probe-CB-LP-RBS-+35 nucleotides) and

to test for the strand-to-strand binding affinity between the probe reporters and the gI intron. The following user input settings were used: RNA nucleic acid type at 25ºC, 1 strand species and 1 strand max complex size for probe-reporters (structure for probe-reporter was predicted to test for base pairing interactions of probes with downstream GFP coding sequence), and 2 strand species and 2 strand max complex size for probe-reporters and introns. The concentration of the hairpin-GFP reporters and introns were assumed to be equimolar (2 μM) in order to test for the relative level of binding of the intron-hairpin complex. NUPACK was also used to calculate the binding energy (ΔMFE) using the same parameters above described.

2.4.3 Flow cytometry

For all experiments, the fluorescence output of each probe was measured using quadruplicate samples. Cells were grown overnight in Luria-Bertani medium (Benton-Dickenson and Company, Sparks, MD) and 10 mg/mL kanamycin (Amersco, Solon, OH), seeded into 20 mL of LB plus 100 μL kanamycin (10 mg/mL stock), and cultured for 2 hours. The remainder of the experiment was carried out under two conditions: (i) samples induced with 800 μL of 20% arabinose (final concentration 0.8%) and 20 μL of anhydrotetracycline (aTc) (final concentration 100 ng/μL), (ii) non-induced samples, where neither of the inducers were added. Five hours after induction, we sampled 100 μL to measure the optical density and an additional 100 μL were pelleted and re-suspended in 1x PBS (Amersco) for flow cytometry. The flow cytometry data was collected with a Benton Dickinson FACSCalibur flow cytometer with a 488 nm argon laser and 530 nm FL1 logarithmic amplifier. Sample data was collected using CellQuest Pro (Benton-Dickenson and Company) with a user define gate. Fluorescent measurements were collected from ~150,000 cells and analyzed using Microsoft Excel and JMP, a statistical

software package. The medians of the populations for non-inducing and inducing conditions were normalized by the average fluorescence for all probes of a given intron variant.

To generate the graphs in Figure 2.5, we determined shifts by calculating differences in fluorescence between the average of median values for the induced and non-induced samples of the same probe five hours after induction of appropriate samples. To quantify the error associated with these shifts, we calculated the standard error of the mean (SEM) as propagated from the original data points. The dashed line represents the median of the fluorescence shift means for all probes.

2.4.4 In vitro binding Assays

An *in vitro* binding assay was performed to demonstrate the ability for probes to bind to intron expressed in cellular extracts. A 20 μM working solution of each oligonucleotide probe (Integrated DNA Technologies) was prepared and the probes were then radiolabeled with P^{32} ATPγ using a reaction mixture containing 1.5 μL of T4 polynucleotide kinase (NEB), 2μL 10x polynucleotide kinase buffer (NEB), 1 μL of the 20 μM working solution, 13.5 μL of double-distilled H_2O, and lastly 1.5 μL P^{32} ATPγ (PerkinElmer Inc). The mixture was then incubated at 37°C for one hour.

In vivo samples of the gI intron were harvested and purified using the techniques described in previously published methods (Cho, Lei et al. 2014). RNA from these sources was suspended in 7μl of buffered solution (50 mM KCl (Avantor Performance Materials Inc) and 80mM MOPS (pH 7.0 Amresco)) and denatured at 95˚C for 2 minutes. After denaturing, 10μL of the radioactive probes were immediately added to all samples, and hybridization occurred at 37˚C for 30 minutes. 2x RNA loading dye (NEB) was

added to all samples and to the ladder to get a final concentration of 10% per volume, and then nuclease free H_2O (Ambion) was added to make the sample volume consistent.

The samples were loaded onto a 6% native polyacrylamide gel that was run at two watts for 24 hours. The gel was then carefully placed on blotting paper (VWR), loosely covered with saran wrap, and left to dry for 3-4 hours at 70°C in a vacuum dryer (BioRad). Upon removal from the dryer, the gel was exposed to a phosphor screen (GE Healthcare) for 3-4 hours at 4°C. Following exposure, the phosphor screen was imaged using a Typhoon Phosphorimager.

2.4.5 Northern Blot Analysis of iRS^3 transcript

To determine the steady state levels of iRS^3 transcript, RNA was extracted as per protocol described in (Cho, Lei et al. 2014) from cells expressing the iRS^3 reporter at 2.5 and 5 hours after induction. The RNA was run down an agarose/formaldehyde gel and blotted using previously described methods (Contreras, Huang et al. 2013)(see **Table A.1** in (Sowa, Vazquez-Anderson et al. 2015) for 16S rRNA and iRS^3 transcript probes).

2.4.6 In vivo dimethyl sulfate footprinting

2.4.6.1 Primers fluorescent labeling

5' amine modified primers (DMS primers K, L and M in Table A.2 in (Sowa, Vazquez-Anderson et al. 2015)) were fluorescently labeled according to previously published protocols (17). For the labeling reaction 1 mL of the purified amine primer (25 mg/mL), 1.2 mL of distilled water, 15 mL of Borax buffer (0.1 M) and 3 mL of NHS-Dye (IRDye® 650 Infrared Dye, Li-Cor) were mixed and incubated in the dark for 3-4 h. Finally, the primers were gel-purified and re-dissolved in 60 mL of nuclease-free water. Their concentrations were estimated using 260 nm absorbance and the extinction coefficients provided by IDT for each primer.

2.4.6.2 In vivo dimethyl sulfate treatment

Cells containing wild-type intron (WTI) plasmid (**Table A.3 in (Sowa, Vazquez-Anderson et al. 2015)**) were grown overnight at 37°C in 5 mL of Luria-Bertani medium (LB). The main culture was induced with 4 mL of 20% arabinose (final concentration 0.8%) at an OD_{600} between 0.15-0.3, and the culture was left to grow for 5 hours at 37°C. Samples were then treated DMS and prepared as described in (Waldsich, Grossberger et al. 2002) and total RNA was extracted from cells using previously described methods (Cho, Lei et al. 2014). After extraction, 4 μg of total RNA were reverse-transcribed using Superscript III RT (Invitrogen) as per manufacturer's instructions.

2.4.6.3 Capillary Electrophoresis

A Capillary Electrophoresis (CE) system (Beckman Coulter A26572 GenomeLab™ GeXP Genetic Analysis System) was used to separate the DMS treated fragments. Each cDNA sample obtained above was mixed with 1 μl of a DNA size standard 600 ladder (GenomeLab Beckman Coulter 608095) and nuclease-free water was added to a final volume of 30 μl in a conical 96-well plate. A drop of mineral oil (GenomeLab Beckman Coulter) was added to prevent evaporation. Another flat bottom 96-well plate was prepared by adding 6-8 drops of separation buffer (GenomeLab Beckman Coulter 608012) to as many wells as cDNA samples are to be run. The samples were separated in the CE system using the following parameters: temperature pre-set to 60°C, denaturation at 90°C for 150 s, injection at 2.0 kV for 30 s and separation at 3.0 kV for 90 min. Lastly the data obtained were analyzed using Capillary Automated Footprinting Analysis (CAFA) (Mitra, Shcherbakova et al. 2008). The CE traces obtained were aligned to the ladder peaks using CAFA. Then, using CAFA, the fit data were filtered and normalized using the "no-DMS" control. All samples were run by technical and biological duplicates.

Chapter Three

Optimization of a biophysical model using large scale *in vivo* antisense RNA hybridization data in bacteria displays improved prediction capabilities

**Article in review in Nucleic Acids Research*

3.1 INTRODUCTION

Current approaches to design efficient asRNAs rely primarily on a thermodynamic understanding of RNA-RNA interactions. However, these approaches depend on structure predictions and have a limited accuracy, arguably due to overlooking factors present in the cellular environment. In this work, we incorporate *in vivo* factors that influence asRNA-RNA hybridization in a biophysical model using large-scale experimental hybridization data. These data are comprised of asRNA hybridization efficacy to 80 regions in three model RNAs: *Tetrahymena* gI intron, *csrB*, and glutamate tRNA. A novelty of this work has been the use of an *in vivo* experimental technique to readily assay "hybridizable" regions within a target RNA. Another unique element of our model is the differential consideration of the influence of the suboptimal structures, often regarded as relevant in RNA-RNA functional interactions, in the availability of the target region to interact with a given asRNA. We showcase the utility of this model by predicting and experimentally validating highly "accessible" regions in 4 additional RNAs: a group II intron, the Spinach II, the 2-MS2 binding domain and the glgC 5'UTR. Additionally, we show the value of our approach by predicting sRNA-mRNA binding regions in two newly discovered, though uncharacterized, regulatory RNAs in *Zymomonas mobilis*.

3.2 RESULTS

3.2.1 Description of asRNA hybridization efficacy by a thermodynamic model that includes a regional measure of interaction availability

In the context of this work, hybridization efficacy is defined as the ability of a given oligonucleotide to establish base-pairing interactions as a cohesive unit with its corresponding target region within an RNA molecule. To quantitatively estimate asRNA hybridization efficacy, we assume that it is directly proportional to the ratio of the concentration of asRNA-target RNA in the bound state (B) over the concentration of the asRNA in the unbound state (U). By transition state theory, rate constants $\boldsymbol{k_B}$ and $\boldsymbol{k_U}$ describe the rate at which the asRNA binds or unbinds to the region on the target RNA, respectively. These rates can be calculated from changes in the Gibbs free energy (G) relative to the intermediate state (I) that denotes non-equilibrium (i.e. initial seeding interaction complex (Rodrigo, Landrain et al. 2012)).

$$\boldsymbol{k_B = \gamma_B e^{\left[-\frac{G(I)-G(U)}{RT}\right]}} \quad \textbf{(1)}$$

The pre-exponential factor γ_B scales the rate and is assumed to be independent of temperature. Similarly, the rate for the reverse, unbinding process can be quantitated as follows:

$$\boldsymbol{k_U = \gamma_U e^{\left[-\frac{G(I)-G(B)}{RT}\right]}} \quad \textbf{(2)}$$

The coupled first order differential equations incorporating these rate constants represent the asRNA-target RNA system, where $[U]$, $[T]$ and $[B]$ denote the concentration of the bound (asRNA-Target complex), target and unbound (asRNA) states, respectively.

$$\frac{d[B]}{dt} = \boldsymbol{k_B[U][T] - k_U[B]} \qquad \textbf{(3)}$$

$$\frac{d[U]}{dt} = \boldsymbol{k_U[B] - k_B[U][T]} \qquad \textbf{(4)}$$

Assuming that intermediates are unstable, that degradation effects are negligible (as experimentally validated for our system (Sowa, Vazquez-Anderson et al. 2015)), and that $[U]$ and $[B]$ are at equilibrium:

$$\frac{d[B]}{dt} = \frac{d[U]}{dt} = 0 \qquad (5)$$

$$\frac{[B]}{[U]} = \frac{k_B[T]}{k_U} = [T]\frac{\gamma_B}{\gamma_U} e^{\left[-\frac{\Delta G}{RT}\right]} = [T]\alpha e^{[-\beta \Delta G]} \qquad (6)$$

In Equation (6), α is the pre-exponential factor, β is a constant and $\Delta G = G(B) - G(U)$ is the energy difference between the asRNA-Target RNA bound and unbound states. Considering that the concentration of the target ($[T]$) is constant across experiments, we can incorporate it into the pre-exponential factor α. Rearranging equation (6) where ΔG is $\Delta G_{overall}$, we obtain

$$v = \log \frac{[B]}{[U]} = -\beta \Delta G_{overall} + \log \alpha \qquad (7)$$

Here v, termed hybridization efficacy, provides a measure of the asRNA-target RNA hybridization and can be estimated experimentally using the logarithm of the ratio of the fluorescence measurements representative of the asRNA-target interaction to the fluorescence measurements representative of background [(FL_{on}- FL_{off})/FL_{off}] obtained from the iRS[3] (see Methods for more details). Briefly, the iRS[3] reporter system is composed of an asRNA that targets a specific region within the target RNA and a cis-blocking element (CB) that sequesters a ribosomal binding site (RBS) and controls the expression of a downstream green fluorescent protein (GFP). Therefore, fluorescence is observed upon asRNA-target RNA hybridization (FL_{on}) as the CB-RBS interaction is disrupted and GFP is expressed due to interaction of the asRNA with the target RNA region (Figure 3.1). FL_{off} is the fluorescence measured in the absence of the target RNA.

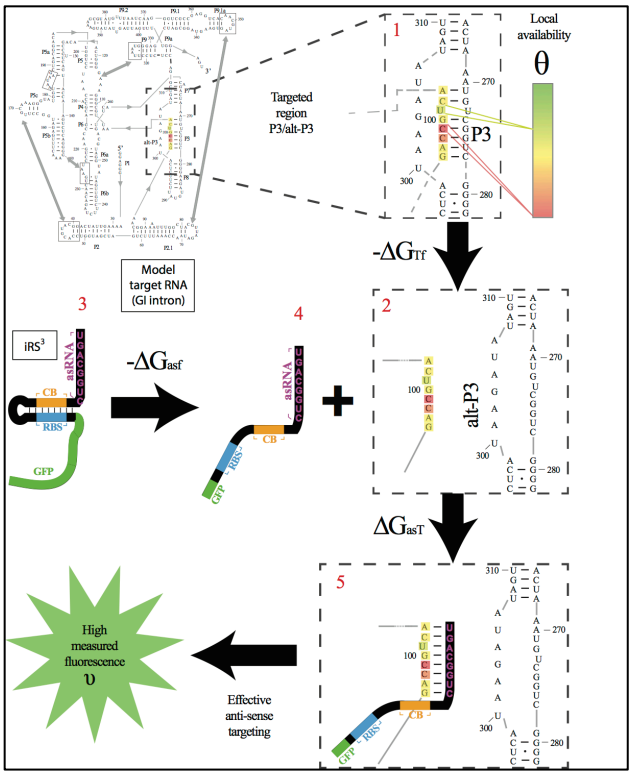


Figure 3.1. Proposed accessibility-based mechanism of anti-sense hybridization in living cells.

1) Example target region with color-coded local availability (estimated by base-pairing probabilities) is shown in canonical conformation, as would be expected in the native state. In asRNA targeting by the iRS^3, the targeted region must unbind from P3 to become single stranded as shown in 2) with a free energy change of $-\Delta G_{Tf}$. The iRS^3 consists of 4 main elements: a cis-blocking strand (CB, orange), a ribosome binding site (RBS, blue), the sequence encoding green fluorescence protein (GFP, green) and the probe (pink and black) of 8-27 nucleotides targeting a specific region shown in 2). The expected native state of the iRS^3 is shown in 3), and it must also unfold to bind the target region as shown in 4) with a free energy change of $-\Delta G_{asf}$. Finally, the two unfolded structures bind as in 5) with a free energy change of ΔG_{asT} to stabilize the unfolded iRS^3 and allow translation of GFP. Effective asRNA targeting results in a high fluorescent response.

A combination of equations (7) and (13) (see Methods section 3.4.5) yields the following linear model (upon taking the logarithm, and rearranging):

$$\boldsymbol{log}\left(\frac{\boldsymbol{FL_{on}-FL_{off}}}{\boldsymbol{FL_{off}}}\right)=\left(\frac{\boldsymbol{FL_{on}}}{\boldsymbol{FL_{off}}}-\mathbf{1}\right)\sim \boldsymbol{v}\equiv \boldsymbol{log}\left(\frac{[\boldsymbol{B}]}{[\boldsymbol{U}]}\right)\propto -\left(\Delta \boldsymbol{G_{asT}}-\Delta \boldsymbol{G_{Tf}}-\Delta \boldsymbol{G_{asf}}\right) \quad (\mathbf{8})$$

This model, comprised of the changes in free energy due to asRNA-target binding (ΔG_{asT}), target region unfolding (ΔG_{Tf}) and folding of the asRNA $\left(\Delta G_{asf}\right)$ depicted in Figure 3.1 captures the thermodynamic driving force of intermolecular base pairing and the penalties for breaking the structures of the asRNA and target regions. Hereafter, equation (8) represents the baseline thermodynamic model from which we depart for further optimization. It is worth noting that similar thermodynamic derivations have been previously used to describe accessibility-based antisense hybridization (Muckstein, Tafer et al. 2006, Busch, Richter et al. 2008). The novelty of this work lies in the treatment of target accessibility. We consider target accessibility a combination of two terms: (1) the energy penalty for the local disruption of the target region using only the minimum free energy structure and (2) the regional availability as estimated by the base-pairing probabilities of the ensemble of suboptimal structures. In part, the rationale behind the use of a regional availability factor is the hypothesis that suboptimal structures hallmark dynamic regions with a differential influence on hybridization efficacy.

Importantly, this availability factor is consequential with the equilibrium derivation represented by Equations 5 and 6. The pre-exponential factor α, also known as the frequency factor in the Arrhenius equation, is a constant that represents the frequency of collisions between reactant molecules. This parameter is often understood to be an inherent characteristic of interaction between molecules at a given temperature, and in many numerical settings, is taken to be a reasonable constant as a simplifying approximation (Voter 2007). In reality, however, the pre-factor is determined by the curvature-dominated structure of the potential energy landscape (Zwanzig 2001), and

hence varies between regions (stretches of nucleotides interacting with each other as cohesive units). To this end, when considering specific *regional* interactions between molecules, the frequency factor may vary. We hypothesize that this frequency factor α is a function of the structural availability of the hybridization region within the target RNA (Figure 3.1), in which a more available target region is more likely to be involved in a collision and vice versa (Figure 3.2). In this work, the "availability" of the target region is assessed at a regional level, termed *regional availability factor* ($\bar{\theta}$), in which a continuous target region of nucleotides is described by the summation of each nucleotide's local availability over the length of the target region:

$$\bar{\theta} = \sum_{i}^{j} \theta_k \qquad (9)$$

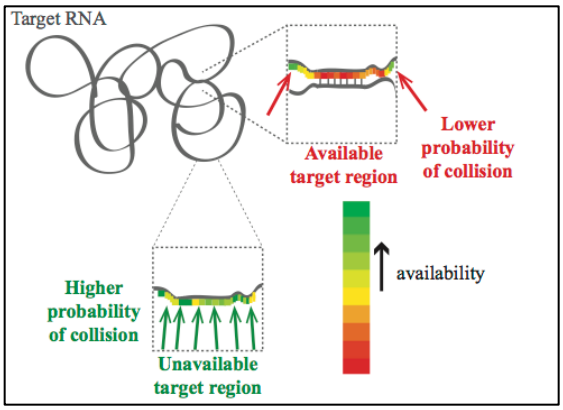


Figure 3.2. Structural target availability.

The influence of structural availability (as defined by the ensemble of suboptimal structures) on intermolecular interactions can be attributed to the frequency of collision. For instance, if a region is highly unavailable due to the presence of a strong secondary structure, there is a lower probability of physical interaction with another molecule, therefore rendering it a region to which it would be difficult to hybridize. On the contrary, a highly available region (that within relevant structural variations is single-stranded) would be more likely to successfully bind with another molecule.

In Equation 9, *i* and *j* represent the start and end of each region correspondingly and θ_k is the local availability of nucleotide k. In the formulation of the model, local availabilities within a target region are considered collectively in order to represent the region as a *single unit*. Therefore, our proposed biophysical model, considers the following four predictors:

$$\boldsymbol{v} \propto \Delta \boldsymbol{G_{asT}} + \Delta \boldsymbol{G_{Tf}} + \Delta \boldsymbol{G_{asf}} + \overline{\boldsymbol{\theta}} \qquad (\mathbf{10})$$

The single term for $\bar{\theta}$ could also be interpreted as an energy penalty due to availability, is slightly similar to the approach of DiChiacchio et al. in AccessFold (DiChiacchio, Sloma et al. 2016) and Tafer et al. in RNAplex (Tafer and Hofacker 2008), influencing the regional energy barrier (initiation energy) that the system is required to overcome in order to produce the bound complex. The local availabilities, θ_k, can be estimated by base pairing probabilities based on the ensemble of suboptimal structures of the target RNA. It should be noted that this term, unlike the energy of target unfolding which is based only on the minimum free energy structure, represents equilibrium structural fluxes (that often facilitate intermolecular interactions ((Lai, Proctor et al. 2013, Grohman, Gorelick et al. 2014))). These Boltzmann-distributed structural variations were obtained from the Nupack webserver (Zadeh, Steenberg et al. 2011) (see *Methods* section for details).

3.2.2 Model optimization using in vivo experimental profiling of asRNA hybridization efficacy

While there is novelty in considering ensemble-based base pairing probabilities as a regional availability factor, the most notable aspect of this study lies in the *in vivo* optimization of the above models using experimental hybridization data for a diversity of RNA targets. Conceivably, one of the greatest challenges in prediction of hybridization efficacy is the ability to account for asRNA-target interactions *in vivo*, where interactions

with other molecules are prevalent due to molecular crowding and complex patterns of ionic strength that vary across different organisms (Leamy, Assmann et al. 2016). To this end, our baseline thermodynamic (Eq. 8) and biophysical (Eq. 10) models were optimized by taking into account *in vivo* hybridization patterns collected directly within cells.

For this work, we have collected large sets of antisense hybridization data using a recently published fluorescence-based assay (iRS^3) for *in vivo* RNA profiling (Sowa, Vazquez-Anderson et al. 2015). Specifically, we interrogated 80 regions within three diverse target RNAs: the gI intron (393 nt, in which 35 regions were probed), the csrB regulator (369 nt, in which 27 regions were probed), and the glutamate tRNA (76 nt, in which 18 regions were probed). Figure 3.3 illustrates all the collected hybridization profiles for these three target molecules, where the heat maps depict differential levels of asRNA-target binding. A list of all 80 asRNAs designed for these molecules is included on Table B.2. These molecules make appropriate targets for this study given their complex structural features that challenge the ability to predict hybridization. For instance, in the gI intron, secondary structure domains that are essential for catalysis such as P4-P6 and P3-P9 (Figure 3.3A) (Beaudry and Joyce 1990, Jaeger, Michel et al. 1997) contain tertiary contacts (gray boxes in 3A) that are connected to each other via pseudoknots (covered by regions 8-10) (Ikawa, Yoshioka et al. 2001). Given that most current secondary structure prediction approaches fail to predict pseudoknots, predicting hybridization efficacy in these domains is extremely challenging. In addition, these interactions are capable of disrupting the folding pathway, e.g. from misfolded to the native state, generating low abundance intermediates in which certain regions are rendered single-stranded (Russell, Das et al. 2006, Mitchell, Jarmoskaite et al. 2013, Xue, Gracia et al. 2016). As discussed in a previous work (Sowa, Vazquez-Anderson et al.

2015), our experimental probing system bears the potential to sense transient states only present *in vivo*, which is consistent with the relatively high hybridization efficacy of regions 8-10. In the case of CsrB, six of the regions with the lowest hybridization efficacies (regions 5, 7, 10, 11, 20 and 22 in Figure 3.3B) contain the binding recognition motif (GGA) for its major target, the CsrA protein sequence (Romeo, Vakulskas et al. 2013) (Lapouge, Perozzo et al. 2013, Holmqvist, Wright et al. 2016); specifically the GGA motif in the stem loop of region 22, has been recently suggested as a strong binding site (Vakulskas, Leng et al. 2016). Our ability to see these patterns reflected in the level of hybridization potential of these regions indicates that our dataset captures the effect of *in vivo* interactions with other cellular factors. Lastly, our *in vivo* experimental data also captures expected high hybridization efficacy within the tRNA at the highly flexible anticodon arm (corresponding to region 8 in Figure 3.3C), consistent with molecular dynamic simulations and crystallographic B-factors for various tRNA models (Bahar and Jernigan 1998). Collectively, these observations validate the experimental data collected and used for model optimization.

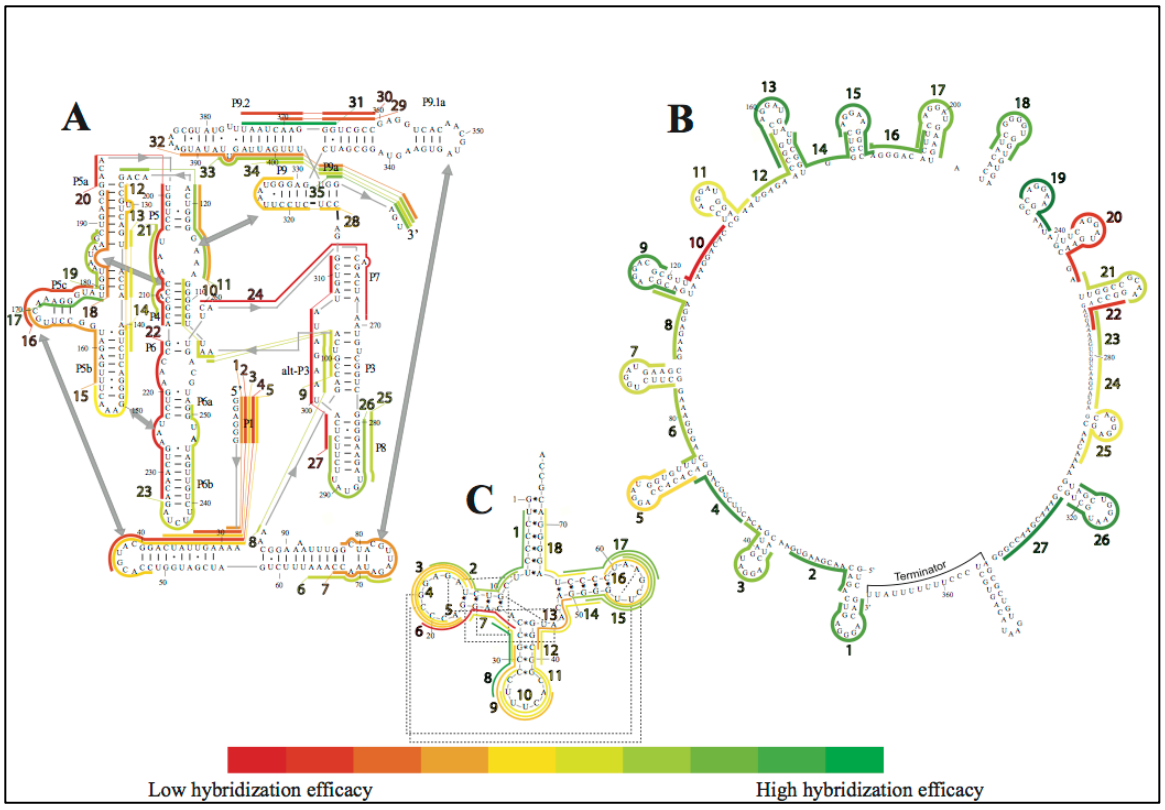


Figure 3.3. asRNA hybridization map as measured by *in vivo* oligonucleotide hybridization.

Heat maps of the asRNA hybridization efficacies for **(A)** The *Tetrahymema* group I intron (35 target regions). Numbers with a dash right next to a nucleotide indicate the standard indexing of the gI intron. Stems (domains) have been named by the convention in our previous work (42) using the letter "P" followed by a number for the gI intron. Tertiary contacts are indicated with a gray double-headed arrow. **(B)** The small RNA CsrB (27 target regions). **(C)** The glutamate tRNA (18 target regions), in which tertiary contacts are indicated with dashed lines. For all three heat maps, color-coded lines represent length and location of a region targeted by the iRS[3] asRNA and color represents hybridization efficacies that can be decoded using the bar scale at the bottom. The target regions/asRNAs were numbered in ascending order from 5' to 3' and labels were colored in accordance with relative hybridization efficacies.

Optimization of the baseline thermodynamic (Eq. 8) and biophysical (Eq. 10) models was performed from collected experimental data by: (1) setting an interval constraint on ΔG_{asf} (-19.3 kCal/mol < ΔG_{asf} < -17.8 kCal/mol) wherein this factor is negligible, (2) scaling all parameters to adjust for their relative importance (e.g. determination of parameter coefficients), and (3) incorporating the interplay between prediction parameters suggested by strong statistical interactions (see *Methods* section). All parameters resulting from this optimization are included in Table B.3. As shown in Figure 3.4, we observe that the regional availability factor ($\overline{\boldsymbol{\theta}}$) by itself and in relation with the energy of target unfolding (ΔG_{Tf}) is prominent in its influence as a predictor of hybridization efficacy. This observation underscores the importance of the differential relationship between the two target accessibility measures. Interestingly, this statistically-derived mathematical form marginally resembles the scaling of the stacking energies by base-pairing probabilities used by Sfold in siRNA design (Ding, Chan et al. 2004). Importantly, the optimization of the baseline thermodynamic (Eq. 8) and biophysical (Eq. 10) models led to the development of the inTher (***in vivo*** optimized **Ther**modynamic), equation (11) and ***in vivo*** optimized **Ther**modynamic **Acc**essibility-adjusted (**inTherAcc**), equation (12) models, respectively.

$$\boldsymbol{v} \propto \Delta \boldsymbol{G_{Tf}}(\Delta \boldsymbol{G_{asT}}) + \boldsymbol{\Delta G_{asT}} + \Delta \boldsymbol{G_{Tf}} \qquad (11)$$

$$\boldsymbol{v} \propto \overline{\boldsymbol{\theta}}(\Delta \boldsymbol{G_{Tf}}) + \Delta \boldsymbol{G_{Tf}}(\Delta \boldsymbol{G_{asT}}) + \boldsymbol{\Delta G_{asT}} + \Delta \boldsymbol{G_{Tf}} + \overline{\boldsymbol{\theta}} \qquad (12)$$

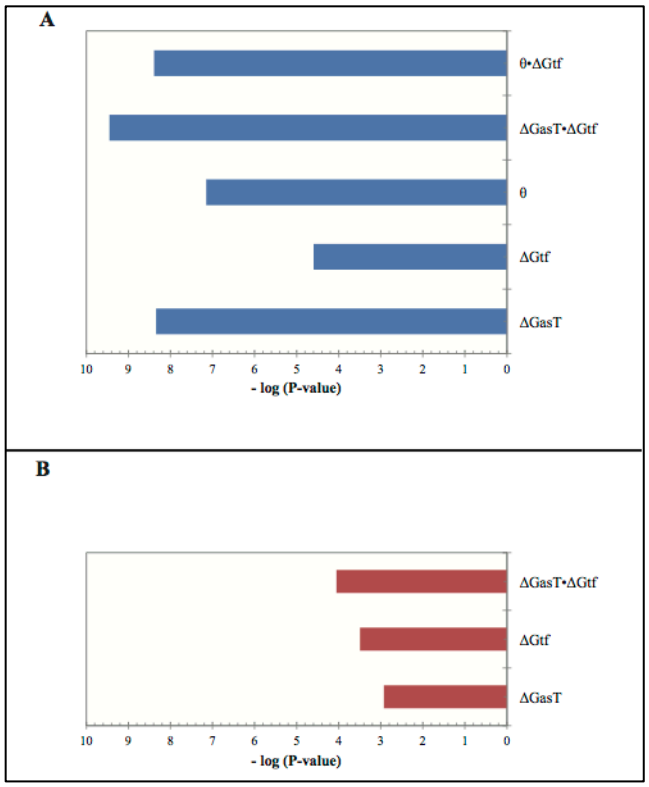


Figure 3.4. Relative significance of each term in the (A) inTherAcc and (B) inTher models.

Optimization of baseline thermodynamic (Eq. 8) and biophysical models (Eq. 10) with *in vivo* data produces significant models. The addition of the availability term (θ) and its statistically significant interaction with the unfolding energy of the target region (ΔG_{tf}) to the inTherAcc model increases the significance of common parameters seen in (B).

Importantly, as presented in Figure 3.5, regression analyses of the ability of these models to capture *in vivo* hybridization data shows that the proposed optimized versions of both, thermodynamic and biophysical models in (Equation (11) and Equation (12), Figure 3.5A and Figure 3.5B respectively) exhibits improved performance relative to the non-optimized models (Equations 8 and 10, Figure 3.5C and Figure 3.5D). Furthermore, the use of the regional availability factor in the inTherAcc model (Equation 12, Figure

3.5B) shows an additional enhancement relative to its counterpart inTher in its ability to capture *in vivo* asRNA hybridization data, making it the best model developed in this work. These findings set the grounds for a final test case in which inTherAcc predictions of highly "hybridizable" regions in 4 molecularly diverse RNAs were experimentally validated. Collectively, these results suggest that consideration of physical intracellular interactions (as captured by the collected data) is vital to improve the accuracy of hybridization behavior predictions.

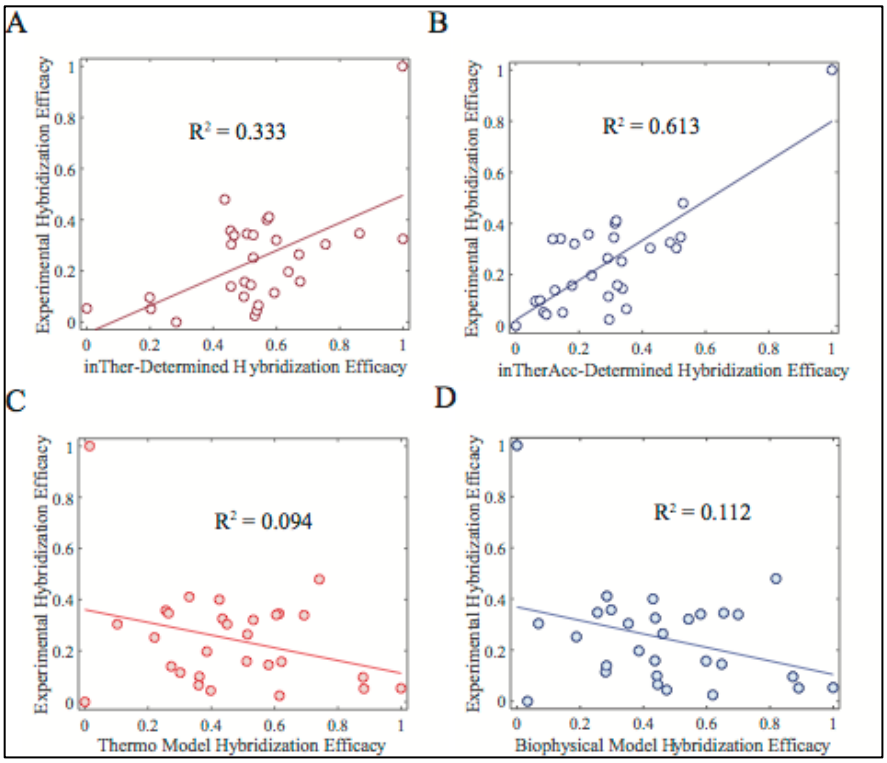


Figure 3.5. Improvement in performance for *in vivo* optimized models underscores the influence of intracellular factors.

Comparing linear correlations of (**A**) the *in vivo* optimized Thermodynamic model (Eq. 12) (**B**) the *in vivo* optimized Thermodynamic Accessibility-adjusted model (Eq. 13), (**C**) the un-optimized thermodynamics-only method (Eq. 8), and (**D**) the un-optimized biophysical model (Eq. 11) shows the ability of the inTher model family to capture the *in vivo* collected data. This improved performance can be attributed to the incorporation of statistical interactions between prediction parameters, which likely well represent the cellular environment.

3.2.3 The inTherAcc Model Proves Effective in Predicting High asRNA Hybridization Regions in Other RNA targets

Initial evaluation of the predictive capabilities of the inTherAcc model was performed by a 3-fold cross validation analysis using the same *in vivo* data set used for optimization. Our evaluation shows that the cross-validated R^2 is 0.37 and 0.09 for

inTherAcc and inTher models, respectively, confirming the increased predictive potential of the inTherAcc model. Given these results, we further tested the prediction capabilities of the inTherAcc model using 4 additional unique RNA targets: the 2-MS2 RNA tag (2-MS2), the model RNA LtrB group II intron (gII), the Spinach II RNA (SpnII) in a tRNA scaffold (Ponchon and Dardel 2007) and the glgC messenger RNA 5'UTR (glgC).

To interrogate highly "hybridizable" regions within these RNA molecules, 1300 target regions across the entirety of these four molecules were randomly compiled. The regions were randomly varied in length between 9-17 nucleotides (see Methods section). The hybridization efficacies of these regions were calculated using the inTherAcc model. Following predictions with the inTherAcc model, 49 regions were selected for experimental validation; 6 regions for the 2-MS2, 13 regions for Spinach, 13 regions for glgC and 17 regions for the larger gII were experimentally tested. In general, regions representing a wide range of predicted hybridization efficacy were selected, with a particular interest in those with highest ranked predicted efficacy (Figure 3.6A). The heat maps illustrated in Figure 3.6A depict relative levels of asRNA hybridization efficacy that were detected for each target molecule using the iRS^3 high throughput plasmid (iRS^3-GG) (see Figure B.1A) as described in the *Methods* section. It is worth noting that two of the top predicted regions (regions 2 and 17) for asRNA hybridization efficacy in the gII intron correspond to well-studied regions that contain one and two tertiary structure contacts, respectively (Cui, Matsuura et al. 2004). These contacts are known to be involved in long range interactions (generally weaker than secondary structure interactions)). In the case of the regions with the highest hybridization efficacy for 2-MS2, regions 5 and 6 both overlap with a 2-MS2 coat protein binding site (Shtatland, Gill et al. 2000) located in a loop. Likewise, it is noteworthy that region 2 within the GlgC 5'UTR, targeting its preferred CsrA interacting site (Baker, Morozov et al. 2002), appears

to be one of the regions with the lowest hybridization efficacies. On the other hand, regions 6 and 7 in the GlgC 5'UTR overlap with the relatively more single-stranded (Kertesz, Wan et al. 2010, Wan, Qu et al. 2014) SD and start codon regions, respectively, and show one of the highest hybridization efficacies. Lastly, in the Spinach molecule, region 4, covers the binding site for DFHBI (Strack, Disney et al. 2013, Warner, Chen et al. 2014), the target molecule of this aptamer. Overall, these observations indicated that our predictions of extreme hybridization potential captured important structural-functional features of these molecules.

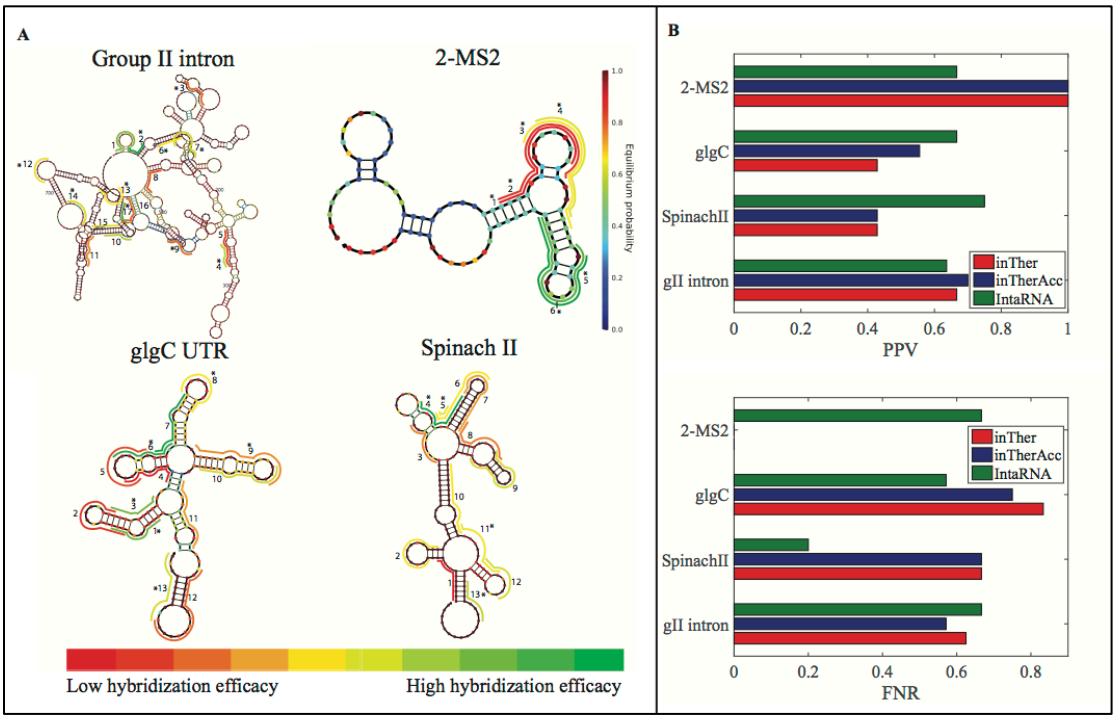


Figure 3.6. Experimental evaluation of hybridization efficacy in four RNAs shows inTherAcc model prediction accuracy comparable to that of benchmark IntaRNA.

(A) Relative hybridization efficacy of each tested region is indicated on the predicted secondary structure (Zadeh, Steenberg et al. 2011) of respective molecules via color-coded lines, in which green and red represent highest and lowest hybridization efficacy, respectively, per scale bar (bottom). Each nucleotide is colored based on equilibrium probability (bar on the left) according to Nupack (Zadeh, Steenberg et al. 2011) output. Regions which were correctly predicted by inTherAcc to be high or low are denoted by an asterisk. **(B)** Comparison of positive Predictive Value-PPV (top) for high hybridization efficacies and False Negative Ratio-FNR (bottom) for low hybridization efficacies, for inTher (red), inTherAcc (blue) and IntaRNA (green).

Importantly, when calculating the Positive Predictive Value (PPV) of regions with high hybridization efficacy and the False Negative Rate (FNR) of regions with low hybridization efficacy for all the data collected, inTherAcc (but not inTher) performed

overall comparably to IntaRNA predictions in terms of PPV and FNR. We chose to benchmark against IntaRNA since it is an accessibility-based approach, uses a seed interaction that resembles our *regional* interaction notion and has been tested for bacterial systems(Busch, Richter et al. 2008). However, inTherAcc displays improved prediction performance, relative to intaRNA, particularly for the gII intron (R^2 = 0.13 vs. 0.08) and 2-MS2 (R^2 = 0.949 vs. 0.014) as shown in Figure 3.7. No difference in performance was observed when considering the linear correlations for glgC and spII target RNAs (R^2<5% for all three models). In summary, these findings support the potential prospects of considering both, intaRNA and inTherAcc, complementary approaches in the prediction of hybridization efficacy (see Figure B.7 for a summary of all the prediction vs. experimental results).

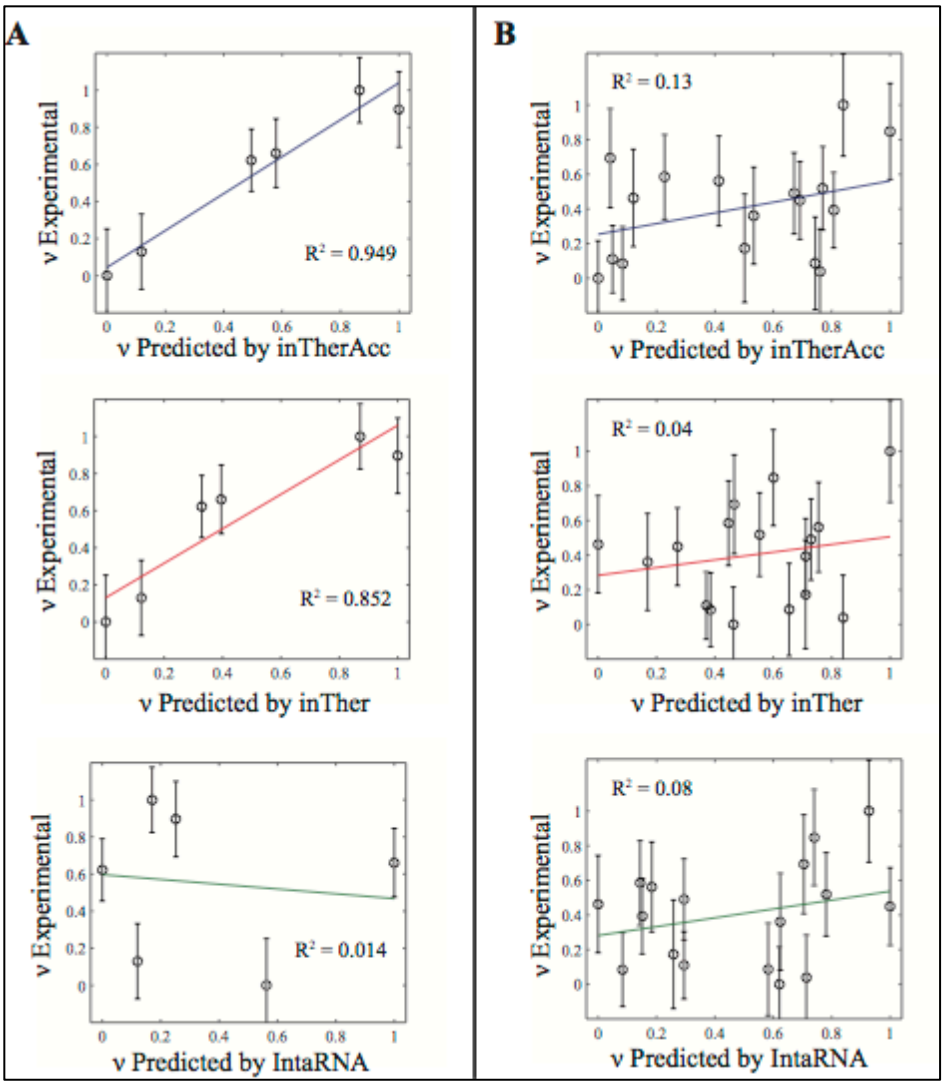


Figure 3.7. Regression analysis on experimental versus inTherAcc-(top), inTher-(center) and IntaRNA-(bottom) predicted hybridization efficacy for (A) 2-MS2 and (B) gII intron.

inTherAcc exhibits superior performance in predicting hybridization potential in **(A)** 2-MS2 and (B) group II intron compared to both inTher and IntaRNA models when considering linear regression fits. **(B)** Higher performance accuracy of hybridization efficacy in gII intron is achieved by inTherAcc due to its capability to predict extreme lows and highs. Error bars indicate standard error of the mean. Both predicted and experimental hybridization efficacies were linearly scaled from 0 to 1.

3.2.4 inTherAcc Aids in Prediction of Target mRNAs

As a final model validation, we evaluated the ability of inTherAcc to aid in prediction of target mRNAs of newly identified sRNAs in a different bacterium. We selected two sRNAs of relevance to ethanol tolerance, Zms4 (280 nt) and Zms6 (304 nt) of *Z. mobilis* (Cho, Lei et al. 2014). A RIP-seq experiment was performed by tagging each sRNA with 2-MS2 RNA. Following purification of the sRNAs and sequencing of the pulled-down (associating) RNAs , the most likely targets were identified as those that showed the greatest transcript enrichment compared to a control (2-MS2 with no sRNA attached). Because of the role of Zms4 and Zms6 under ethanol stress, we expect their mRNA targets to include stress-related genes. Indeed, as expected, many potential targets enriched by MS2 pulldown for both Zms4 and Zms6 were related to stress responses, including global stress response regulators, heat shock proteins, protein folding chaperones, and DNA repair proteins Because the inTherAcc model is well suited to help narrow the large pool of potential targets by predicting those with most favorable hybridization efficacies, potential regions of interest in both sRNAs were randomly compiled and ranked by hybridization efficacies using our inTherAcc model (Figure 3.8A), as described in the Methods section. As observed in Figure 3.8A, interesting "hot spots," defined as regions exhibiting predicted extreme (high or low) hybridization efficacies were identified and considered for further analysis. The rationale behind using regions with predicted high and low hybridization efficacies is based on the hypothesis that these regions are likely to be functional sites either highly available or unavailable based on active binding to *in vivo* factors. The reverse complement sequences of the five highest and five lowest predicted hybridization efficacies were selected for BLAST analysis to identify potential "top" likely interacting mRNA targets (for a total of 52-54 unique genes considered). Comparisons of these results with data obtained from RIP-seq

experiments supported the target prediction capability of the inTherAcc model. As shown in Figure 3.8B, inTherAcc predicted about 28 and 22 potential targets, respectively for Zms4 and Zms6, found in the set of RIP-seq-determined enriched transcripts. Importantly, about 8 and 7 potential targets respectively for Zms4 and Zms6 were found within the top 20% pulled-down targets (ranked by fold change enrichment relative to the 2-MS2-only control). In all cases for each region predicted to be an mRNA binding site, multiple potential targets were found suggesting the ability of these sRNAs to exert multiplex regulation (Table B.4). As expected, a considerable portion of enriched transcript associations of Zms4 and Zms6 correctly predicted by inTherAcc code for proteins involved in ethanol tolerance mechanisms, specifically those that facilitate (1) protein folding and transport, (2) redox metabolism, and (3) stress response (Ingram 1989, Cray, Stevenson et al. 2015), further validating our results. In addition, inTherAcc showed a comparable performance to benchmark IntaRNA (Table B.4 and Figure 3.8B). The limited number of matches in target prediction (Figure 3.8B) between both approaches underscores the potential complementarity between them. Collectively, these results show the potential of the model to aid in gene target prediction and, more specifically, to identify *potential functional regions* that act via base-pairing.

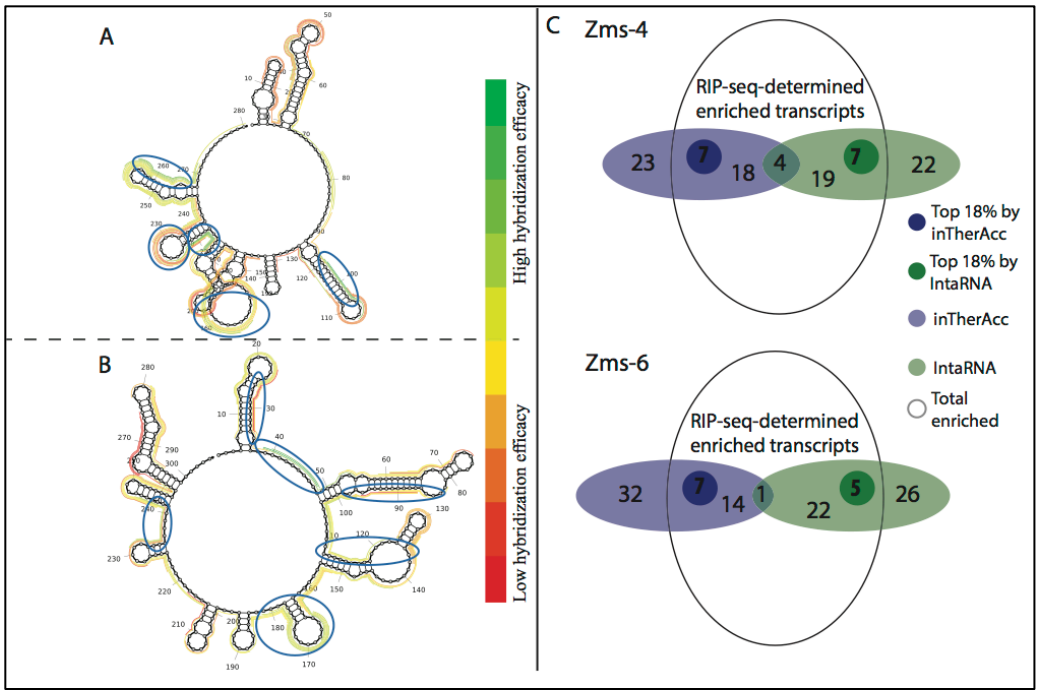


Figure 3.8. inTherAcc aids in prediction of mRNA targets for *Z. mobilis* (A) Zms4 and (B) Zms6.

Ten regions evenly distributed at the top (green) and bottom (red) hybridization efficacy scale were selected as potential mRNA-sRNA binding sites for further prediction of target mRNA candidates and comparison with RIP-seq data. The regions that matched with the 18% of top enriched candidates (log 2 of fold change sRNA/only MS2) are circled in blue. **(C) Overview of the prediction performance for IntaRNA (green) and inTherAcc (blue).** Venn diagram showing the total enriched candidates (fold log2 of sRNA/only MS2 > 0). A total of 52 and 54 candidates respectively for Zms-4 and Zms-6 were predicted using both approaches. Darker green and darker blue circles represent the top 18% enriched candidates that each approach predicted correctly.

3.4 Discussion and Conclusions

The inTherAcc model incorporates a series of thermodynamic terms to account for energetics of intramolecular folding, intermolecular binding, and the target region

availability using the Boltzmann distribution of possible structural configurations. The novelty of this work lies in that our approach integrates large scale *in vivo* data as well as the interplay between the components of target accessibility as understood by (1) an availability factor based on suboptimal structures and (2) thermodynamic consideration of RNA unfolding, identified during model optimization. Our results suggest that the family of inTher models that we have developed could assist current asRNA predictions to capture "hybridizability" *in vivo*. Our work also highlights the potential of using *in vivo* experimental datasets to increase prediction accuracy for effective selection of sites for asRNA targeting and provides a methodology to do so. The observed relationship between target RNA folding energy and regional target availability as estimated by a summation of local base-pairing probabilities was shown via statistical model optimization (Figure 3.4 and Table B.3) and suggests that scaling this free energy by its availability factor plays a significant role in determining efficacy of RNA hybridization *in vivo*. Other research groups have used similar scaling approaches with significant improvements in the performance of siRNA design and predictions (Ding, Chan et al. 2005, Shao, Wu et al. 2006, Shao, Chan et al. 2007). The main difference of our scaling scheme relative to these previous efforts is its regional nature. While previous works scaled the stacking energies of interacting nucleotides one by one according to nucleotide-specific base-pairing probabilities, this approach assumes that any given asRNA behaves as an indivisible unit. In addition, this *in vivo* optimization has brought about coefficients for our model that are meaningful in capturing intracellular behavior. For instance, as expected, we observed a strong influence of the intramolecular structure of the target region on the hybridization efficacy (Figure 3.4). Moreover, the estimated coefficients could be indicators of the presence of binding factors, the effect of molecular crowding, or even the presence of ionic species in the cellular milieu. For example,

divalent ion influence on ribozyme active site structural arrangement (Woodson 2005) was likely accounted for to an extent by optimizing inTher models with gI intron data. It is therefore not surprising that the optimized inTherAcc model was an improved predictor of gII intron hybridization efficacy. To the best of our knowledge, no approach in the past has attempted to consider the *in vivo* environment by optimizing a current biophysical model using large sets of *in vivo* data collected in bacteria and applying it to predict other RNA molecules, while simultaneously studying the influence of target accessibility.

Through *in vivo* optimization of model parameters, we achieved a highly reliable qualitative prediction of highly "hybridizable" regions in a wide array of RNA molecules. Overall, the inTherAcc model performs at levels above 63% and below 60% in PPV and FNR, correspondingly. It also is at least comparable to benchmark IntaRNA, bearing an advantage in specific cases likely due to the incorporation of *in vivo* factors during model optimization. Interestingly, some of the observed discrepancies between experimental and predicted hybridization efficacy in the glgC 5'UTR can be attributed to competitive binding between the asRNA and factors that naturally interact with this RNA that are not fully accounted for by the collected dataset. In many of these cases, we suspect that even our experimentally collected datasets fail to capture the full set of molecular interactions (e.g. with other intracellular factors) given that only limited environmental conditions were tested where the full range of these interactions does not occur. This is likely the case for regulatory RNA regions like the glgC UTR, in which different interactions are observed *in vivo* under nutritional stresses (not tested in this work). As a result, we hypothesize that further prediction accuracy can be achieved for these models by expanding the collected datasets to include a variety of environmental conditions (e.g. cellular stresses) to capture a broader range of interactions.

Remarkably, the inTherAcc approach provides the following general strategies in asRNA design: (1) the suggestion of a free energy interval within which the thermodynamic stability of the asRNA does not seem to influence hybridization efficacy, (2) the realization that both low and high inTherAcc-predicted hybridization efficacies could indicate functional sites that may be interesting targets for asRNAs and (3) evidence of the potential influence of suboptimal structures in hybridization efficacy that aids in identification of target dynamic regions. Overall, we envision that the inTherAcc approach will assist in the characterization of newly-identified regulatory RNAs and the design of synthetic elements that require RNA binding through complementarity by improving reliability of RNA targeting performance *in vivo,* particularly in bacteria.

3.4 METHODS

3.4.1 Plasmids and strains

As previously described in (Sowa, Vazquez-Anderson et al. 2015), the fluorescence-based iRS^3 system provides a measurement of asRNA-RNA hybridization by using various 8-27 nt sequences (asRNAs) that are complementary to a target RNA. In this system, a fluorescence shift is observed when an asRNA successfully binds the region of interest in the target RNA. A total of eighty asRNAs targeting unique regions in three target molecules (gI intron, CsrB and tRNA) were analyzed in this work for model optimization purposes. Forty-nine asRNAs targeting unique regions in four different target molecules (gII intron, SpinachII, glgC 5'UTR and 2-MS2) were also used to assess model prediction capabilities. To construct these experimental asRNA systems, a modified Golden Gate cloning-based plasmid was introduced for high-throughput cloning that included the following changes to the previously published "Wild Type Intron Probe I reporter" (Sowa, Vazquez-Anderson et al. 2015): a p-chlorophenylalanine negative

selection cassette (PheS) in place of the asRNA sequence (between EcoRI and the CB element flanked by two BsmbI restriction sites)(Kast and Hennecke 1991, Kast 1994). We termed this plasmid iRS^3 Golden Gate (IRS^3-GG) and it is illustrated in Figure B.1A. All target molecules, with the exception of the natively targeted tRNA, were separately introduced in the iRS^3-GG between the Xba1 and Sal1 restriction sites (see plasmid map in Figure B.1A). In the case of gII, SpII, glgC and 2-MS2, Gibson assembly (Gibson 2011) (using Gibson Assembly mix from NEB) was performed. CsrB was introduced via traditional restriction cloning. All primers used for cloning of target molecule into iRS^3-GG are listed in Table 3.1 of Supplementary Data.

All asRNA sequences within the plasmid (Table B.2), besides 11 asRNAs corresponding to regions within the gI intron that were previously synthesized and published (Sowa, Vazquez-Anderson et al. 2015), were either ordered from GenScript Inc., synthesized by a site-directed mutagenesis approach (QuikChange II Site-Directed Mutagenesis Kit, Agilent Technologies) by modifying a previously synthesized asRNA, synthesized via Gibson Assembly (Gibson 2011) or synthesized by using a high throughput Golden Gate approach as described in (Engler and Marillonnet 2014) on our iRS^3-GG plasmid. For the Golden Gate approach, complementary primers (ordered from IDT) containing each asRNA sequence with the proper flanking overhangs were annealed and cloned after digestion with BsmbI (Thermo Scientific) to replace the PheS cassette. To increase cloning throughput, two to five asRNAs were combined into a single reaction and later transformed into DH5α chemically-competent cells or NEB Turbo electro-competent cells and plated in Luria-Bertani (LB)/Agar media supplemented with p-cholorophenylalanine (p-Cl-Phe) to select for the clones harboring the appropriate asRNA. Once the asRNA sequences were confirmed by DNA sequencing, the newly synthesized plasmids were transformed into K-12 MG1655, our experimental strain, or,

in the case of CsrB, into a CsrB-deficient K-12 MG1655 strain. An overview of the specifics of the asRNA synthesis strategy is included in Figure B.1B.

For the evaluation of sRNA target prediction as aided by inTherAcc and IntaRNA, we utilized pBBR1MCS2-pgap vector for constitutive expression. Each sequence confirming the corresponding small RNA fragments between NheI and SalI which were synthesized by GenScript® and then cloned into pBBR1MCS2-pgap vector (Zou, Zhang et al. 2012), resulting in pBBR1MCS2-pgap-sRNA. For 2MS2BD-Zms4/Zms6/control constructs, gBlock® (NEB) of 2MS2BD-Zms4/Zm6/control was used for cloning into pBBR1MCS2-pgap vector, resulting in plasmids abbreviated 2MS2-Zms4/2MS2-Zms6/2MS2-control.

3.4.2 Selection of target RNAs

Rationale for target molecule selection was based on molecule complexity, size, and functional interactions. For instance, the gI intron is a relatively large (393 nucleotides), well-studied RNA model (Russell, Das et al. 2006, Wan, Suh et al. 2010) whose many structurally significant regions have been previously probed with the iRS[3] system (Sowa, Vazquez-Anderson et al. 2015). These studies have shown that this autocatalytic molecule may well-represent the complexity of structural features present in most RNAs targeted for regulation (e.g. UTRs of mRNAs (Ding, Tang et al. 2014))**.** On the contrary, the 76 nucleotide long glutamate-tRNA has a wide assortment of interactions with intracellular factors, including mRNAs, rRNAs, various modification enzymes and other proteins despite exhibiting tight tertiary structure comparable to that of the gI intron (Brion and Westhof 1997). The third molecule chosen for model optimization, CsrB, is a non-coding RNA whose multiple protein binding motifs contribute to the translational regulation of a large number of mRNAs (Babitzke and

Romeo 2007). Compared to the previously described molecules, CsrB (369 nucleotides) is less structurally sophisticated than the gI intron and the tRNA.

For assessing the prediction capabilities of our model, 4 alternative RNA molecules were used: the 2-MS2 coat protein binding domain (2-MS2), the model LtrB group II intron (gII), the Spinach II RNA (SpII) and the glgC messenger RNA 5'UTR (glgC). This set of RNAs cover a wide array of types, functions, structures and sizes. MS2 and Spinach II are commonly used to investigate RNA interactions, more specifically, to isolate RNAs to determine specific RNA interacting complexes (Faoro and Ataide 2014) and track RNA movement (Paige, Wu et al. 2011), respectively. Similarly, 5' UTRs often use their structure to regulate the translation of their associated mRNA. The gII intron was selected given the interest in targeting ribozymes for understanding the molecular mechanisms for catalytic activity, largely regulated by their complex folding (Frommer, Appel et al. 2015).

3.4.3 Fluorescence Measurements and Calculations of asRNA hybridization using the in vivo RNA Structural Sensing System (iRS^3)

In general, flow cytometry experiments were carried out as previously reported (Sowa, Vazquez-Anderson et al. 2015). All target molecules (except for the glutamate tRNA) were evaluated under overexpression conditions in which the hybridization efficacy is evaluated as the ratio between the fluorescence in the presence of the target RNA with baseline fluorescence (in the absence of the target RNA) subtracted out (FL_{on}-FL_{off}) to the baseline fluorescence (FL_{off}). For all hybridization calculations, FL_{off} was scaled by an adjustment factor of 0.65 to account for the excess abundance of the reporter probe relative to the target RNA, as approximated by recently obtained RNA-sequencing data (unpublished). In the case of the tRNA, the target was evaluated using native levels given its natural presence and abundance in E. coli cells using plasmid in Figure B.1C. In

this case, FL_{on} and FL_{off} represent the fluorescence in the presence of the asRNA (iRS[3]+specific oligonucleotide) and the fluorescence in the absence of the asRNA, respectively. FL_{off} fluorescence was measured right before induction (at time "0") and FL_{on} was collected 45 min after induction (See Figure B.2A for a correlation between uninduced and time "0"). Seeding cultures originated from independent overnights and uninduced and induced samples proceeded from the same initial seeding culture. Specifically, seeding was done in LB (40 mL+50 μg/mL of kanamycin) and split up into two 20 mL cultures at the time of induction (1-2 h of growth upon seeding) for the collection of model optimization data. When testing model predictions, seeding was done in LB (200 μL+50 μg/mL of kanamycin) and split up into two 100 μL cultures in 96-well plates at the time of induction.

3.4.4 In vivo DMS footprinting and calculation of regional availability ($\bar{\theta}$)

The DMS reactivity of the gI intron was obtained using a previously published protocol (Sowa, Vazquez-Anderson et al. 2015). In this work, we published the *in vivo* DMS reactivity profile for the full gI intron (Figure B.3). Previously, the reactivity for only select regions had been published (Sowa, Vazquez-Anderson et al. 2015). The nucleotide indexing for the gI intron follows the established consensus for this well-known molecule. These data were filtered and normalized using specialized software, the Capillary Automated Footprinting Analysis (CAFA) (Mitra, Shcherbakova et al. 2008). The reactivity values for the untreated sample were subtracted from the average reactivity value of two independent DMS treated samples. The DMS reactivity for Gs and Us was estimated by assuming the same reactivity as their pairing partners (if paired), and, when unpaired, an average reactivity value for "exposed" nucleotides was assigned. Special cases were those Gs and Us exposed in loops (G58, U59, G92, G112, G119, U120, G126,

U179, U185, G200, G201, U202, U225, G227, U247, G254, G279, U300, G303, U322, U323, G331, U340, G341, G357, G358, G368 and U372) that were assigned values more similar to their neighbors and other As and Cs present in loops. The regional target availability factor was then calculated using the average of the individual reactivity values of each nucleotide in the given target region over the length of the target region.

3.4.5 Derivation of the accessibility-based thermodynamic model

The quantity $\Delta G_{overall}$ is the overall free energy change related to the different mechanistic steps associated with asRNA binding to the target RNA region; the folding and binding processes considered are depicted in Figure 3.1. This quantity is represented as the combined contribution of the free energies of: (i) the Watson-Crick base-pairing of the asRNA to the target RNA region (ΔG_{asT}), (ii) the local unfolding of the target RNA region required for asRNA binding (ΔG_{Tf}), and (iii) the unfolding of the asRNA required for binding (ΔG_{asf}). The sum of these terms comprises the total energy of hybridization, $\Delta G_{overall}$:

$$\boldsymbol{\Delta G_{overall} = \sum_i \Delta G_i = \Delta G_{asT} - \Delta G_{Tf} - \Delta G_{asf}} \qquad \textbf{(13)}$$

In (13), subscripts asT, Tf and asf denote the asRNA-target RNA hybridization, the target RNA folding, and the asRNA folding respectively.

3.4.6 Calculation of free energy of hybridization (ΔG_{asT})

To calculate the Gibbs free energy of binding between the perfectly complementary stretch of nucleotides (asRNA) within the iRS[3]-asRNA system and the target region, the energy parameters for the nearest neighbor model published in (Xia, SantaLucia et al. 1998) were used. Only canonical base-pairing (Watson-Crick base-

pairs), penalties for self-complementarity within the asRNA, and AU ending were considered for the calculation of the stacking energies.

3.4.7 Calculation of free energy of the target region (ΔG_{Tf})

To calculate the Gibbs free energy of target region folding, the energy parameters for the nearest neighbor model previously published (Xia, SantaLucia et al. 1998) were used. The target region plus one extra nucleotide at each end was considered to account for stacking contributions of neighboring base-pairs. The folding of the target RNA was considered to be a local event due to the tight coupling of prokaryotic transcription and translation. Such assumptions of local folding have previously been used in a structural study of bacterial genes (Shao, Wu et al. 2006). To calculate the stacking energy contributions, the consensus secondary structure of the gI intron was considered (Tijerina, Mohr et al. 2007). For all the other target molecules, a secondary structure prediction from the RNAStructure webserver was used (Reuter and Mathews 2010). Since GU base-pairs are somewhat extensively found in the structure of our target RNAs, they were treated as nearest neighbor stacks, similar to Watson-Crick helices. In addition, the penalty for ending in a GU was the same as an AU ending. In our treatment of GU pairs we followed the parameters reported by Mathews et al.(Mathews, Sabina et al. 1999). No energy parameters for other structural motifs such as loops, bulges, etc. were taken into account.

3.4.8 Calculation of regional availability

To support high-throughput estimations of regional availability ($\bar{\theta}$), without involving experimental structural studies, local availability (θ_k) was estimated by base pairing probabilities determined by Boltzmann-distributed structural variations provided by the Nupack webserver (Zadeh, Steenberg et al. 2011). This structural accessibility

estimation was shown to capture of *in vivo* experimental DMS reactivity at the regional level, supporting the use of base pairing probabilities as a substitute for experimentally determined structures (Figure B.4).

3.4.9 Calculation of free energy of folding for the asRNA (ΔG_{asf})

The "allSub" subroutine of the RNAStructure webserver (Reuter and Mathews 2010) was used to predict the secondary structure of the asRNA+iRS[3] transcript (5'-6 nt + asRNA + 56 nt-3'). The Gibbs free energy of the minimum free energy structure was used to represent the asRNA folding energy (ΔG_{asf}). For the purpose of this analysis, the transcript considered 62 nucleotides in addition to the 8-27 nucleotides of the asRNA (for a total of 70-89 nucleotides). Additionally, six nucleotides upstream of the asRNA were included as part of the transcript to account for imprecision of transcriptional start sites. In this way, any potential interactions between the asRNA and the segment downstream from the RBS site were accounted for. The specific sequence is as follows: 5'GAA UUC -asRNA- UAC CAU UCA CCU CUU GGA UUU GGG UAU UAA AGA GGA GAA AGG UAC CAU GAG UAA AG 3'.

3.4.10 Model optimization via regression analysis using experimental hybridization data

Regression analysis was used to statistically evaluate the contributions of the proposed biophysical factors in the derived models. Briefly, a linear model relating the experimental response variable υ (defined as the logarithm of the ratio of FL_{on} to Fl_{off} measurements) to the previously described factors ($\bar{\theta}$, ΔG_{asT}, ΔG_{Tf}, ΔG_{asf}) was composed and the coefficients for the various factors were fit by ordinary least squares regression. Coefficient fitting and statistical analysis of parameter contributions to the overall model were performed using MatLab Math, Statistics and Optimization package

(specifically "fitlm" function). A total of 383 independent fluorescence measurements (representing asRNA hybridization efficiency) across all three optimization molecules were used for regression analysis. ΔG_{asf} was constrained to an interval between -19.3 kCal/mol and -17.8 kCal/mol, where its influence became statistically insignificant (p-value > 0.05), allowing the other more relevant factors ($\bar{\theta}$, ΔG_{asT}, ΔG_{Tf}) to be studied in isolation (see *Model optimization via regression analysis using experimental hybridization data*, Figure B.5). In addition, the predictors in equation (10) were normalized by the length of the asRNA, to decrease linear dependency on this design parameter. A 3-fold validation was performed to test for prediction power for both optimized models (groups were determined based on regions and their replicates). Each cross-validated R^2 was calculated as the adjusted coefficient of determination of the linear regression fit between the experimental data of each independent group and corresponding predicted values from a model derived from the remaining 2 independent groups.

For all regression analyses conducted in this work, factors and their potential interactions were considered statistically meaningful if their p-value (t-test) was lower than 0.005. Additionally, the quality of the regression was qualitatively evaluated by visual inspection, ensuring that the residuals showed a strong normal distribution (see Figure B.6)

3.4.11 Selection of target regions for evaluation of model prediction power

About 1300 of potential target regions within each molecule (gII intron, SpinachII, glgC 5'UTR, 2-MS2) were randomly generated. Starting from the first nucleotide of the molecule, regions of random length between 9 and 17 nucleotides were designed sequentially with one nucleotide overlap between each region. This process

was iterated 7 additional times with respect to integer-increasing nucleotide overlap between regions, ultimately producing 8 sets of target regions with 1-8 nucleotide overlaps. To ensure that every nucleotide of each molecule was included within each set of target regions, the last region within each set was not of random length. Instead, if the first nucleotide of a prospective region was within 9-17 nt of the last nucleotide, the final probe of the respective iteration was established as the region from the first nucleotide of the prospective region to the last nucleotide of the molecule. The full set of asRNAs targeting these regions was then filtered by their calculated folding free energies to select a subset of 366 asRNAs with ΔG_{asf} ranging from -19.3 to -17.8 kcal/mol.

Lastly, this filtered set of asRNA designs was used to predict hybridization efficacies via the optimized model. The total number and sequence of asRNAs for experimental validation for each RNA were chosen based on molecule length, biophysical model hybridization prediction, and targeting region. Seven asRNAs were chosen for the smallest target molecule (2MS2), thirteen for the "mid-sized" molecules (glgC 5'UTR and Spinach II) and eighteen for the largest (gII intron). Approximately 40% of asRNAs for each molecule were selected for their low predicted hybridization values, defined as a predicted hybridization efficacy equal to or less than the median of the asRNA pool within a molecule. The remaining asRNAs selected were within the predicted high hybridization efficacy pool, specifically, with predicted hybridization greater than the median. Precautions were taken to avoid selection of asRNAs targeting highly similar regions (greater than 5 shared nucleotides); however, exceptions were made when two asRNAs targeting similar regions showed interesting differences in terms of predicted hybridization efficacy (differences greater than the standard error of the pool).

3.4.12 Statistical Evaluation of Model Prediction Power

For each target region designed for experimental validation (above), hybridization efficacies as predicted by benchmark software IntaRNA were also estimated (28). First, the hybridization energy of each region was calculated using the pre-set seed, folding, and output parameters with inputs of target RNA sequence and asRNA sequence. The hybridization energy was then, normalized by the length of the asRNA oligonucleotide. The lowest (most negative) normalized energy values indicated a higher predicted hybridization potential. Predicted (by both *in vivo*-optimized models and IntaRNA) and experimental hybridization efficacies for each of the 4 molecules were then linearly scaled to fall between 0 and 1. To statistically evaluate the prediction potential of our models, experimental and predicted "high" hybridization efficacy was defined as any hybridization efficacy greater than one standard deviation above the hybridization efficacy mean of points below the median within experimental and predicted subsets, respectively. Any points below these thresholds were considered to have "low" hybridization efficacies within their categories. To evaluate the performance of our models, we also calculated the Positive Predictive Value (PPV) of regions with high hybridization potential and the False Negative Rate (FNR) of regions with low hybridization efficacy defined in this specific context as follows:

$$PPV = \frac{\#\ of\ high\ v's\ correctly\ predicted}{total\ \#\ of\ predicted\ high\ v's}$$

$$FNR = \frac{\#\ of\ low\ v's\ incorrectly\ predicted}{total\ \#\ of\ predicted\ low\ v's}$$

3.4. 13 Evaluation of Prediction of sRNA-mRNA Binding Regions

Approximately 150 potential binding regions within Zms4 and Zms6 (sRNAs recently discovered in *Z. mobilis*(Cho, Lei et al. 2014) but not fully characterized yet)

were randomly generated following the process described in "*Selection of target regions for evaluation of model prediction ability*". Hybridization efficacy of each region was predicted using the inTherAcc model. Ten total regions were selected for further target prediction analysis for each sRNA: 5 regions that exhibited the highest and 5 regions that exhibited the lowest predicted hybridization efficacy. During the selection process, regions with any overlap to a prior selected region were not considered in an attempt to select for unique regions. The reverse complement of the selected regions was inputted to nucleotide BLAST(Altschul, Gish et al. 1990) to identify potential target mRNAs of these two sRNAs in *Zymomonas mobilis subsp. mobilis* ZM4 (taxid:264203). For selected regions encompassing less than or equal to 10 or 12 nucleotides, 2 or 1 nucleotides of the neighboring sRNA sequence were added onto both ends, respectively, to increase sequence specificity of the hits obtained by BLAST. Five potential targeting arrangements were chosen for each region from BLAST results with the constraints of 1. Minimization of E-value, 2. Correct orientation of gene sequence, and 3. Location of sequence at most 400 nucleotides upstream of a TSS or 200 nucleotides downstream of a TTS. For each target region designed for experimental target validation (above), hybridization efficacies as predicted by benchmark software IntaRNA were also estimated (28). First, the hybridization energy of regions within each sRNA with target mRNAs was calculated using the pre-set seed, folding, and output parameters with inputs of sRNA sequence and *Z. mobilis* genome, target NCBI reference sequence NC_006526, within both -300 to +300 nucleotides around the start codon and stop codon, the maximum consideration window offered by the IntaRNA software. Results from both start and stop codon were consolidated within each sRNA and ranked according to energy values. An equal number of target genes to those of inTherAcc, harboring the

lowest energy of interaction with the sRNA were ultimately chosen as IntaRNA predictions for Zms4 and Zms6.

3.4.14 Strains and culture conditions for MS2 pull-downs

Zymomonas mobilis 8b strain was used in this study (Zhang, Eddy et al. 1995). *Z. mobilis* 8b strain was cultured in RM media (Glucose, 20.0 g/L; Yeast Extract, 10.0 g/L; KH_2PO_4, 2.0 g/L; pH 6.0) at 33 °C. *Escherichia coli* DH5α was used for plasmid construction and manipulation. Plasmids containing pBBR1MCS2-pgap-sRNA and 2MS2-Zms4/Zms6/control strains were cultured with 350 ug/ml of kanamycin for *Z. mobilis* 8b and with 50 ug/ml for *E. coli*. For the preparation of the samples for RNA sequencing, each overexpression, empty plasmid, and wildtype strain was initially grown in 5ml culture overnight. Then, cells were transferred into 500ml to adjust starting OD_{600nm} at 0.1. Cells were grown at 33 °C for 12 hrs. 50ml of cells were pelleted and stored at -80 °C for further processing.

3.4.15 RNA Preparation for Evaluation of Zms4 and Zms6 mRNA Targets

Total RNA of 2MS2-Zms4/2MS2-Zms6/2MS2-control strains was prepared according to a protocol previously published in (DiChiara, Contreras-Martinez et al. 2010) for all the growth conditions tested. Briefly, cells were grown anaerobically and collected at each time points for RNA Sequencing. All centrifugation was performed at 4°C. Cells were pelleted and resuspended in 1 ml TRIzol reagent (Invitrogen). Following pelleting, cells were transferred to screw cap tubes containing glass beads (Sigma) and incubated at 25°C for 5 min. Cells were lysed using a mini-beadbeater (BIOSPEC), with 100-s pulses three times. Cells were kept on ice for 10 min between each 100-s treatment. The beads and cellular debris were centrifuged at 4 °C for 2 min. The supernatant was transferred to a clean siliconized 2 ml tube. After addition of 300 μl of chloroform:

isoamyl alcohol mix (v/v 24:1), the samples were inverted for 15 s, and then incubated at 25 °C for 3 min. Then, tubes were centrifuged at 13,000 rpm for 10 min, and the aqueous top phase transferred to a clean siliconized 1.5 ml tube. Following this step, 270 μl of isopropanol and 270 μl of a mixture of 0.8 M sodium citrate and 1.2 M sodium chloride was added. The samples were mixed well, and then incubated on ice for 10 min. The RNA was pelleted by centrifugation at 13,000 rpm for 15 min. The pellet was washed with 1 ml 95% cold ethanol and centrifuged for 5 min. The pelleted RNA was allowed to air-dry for 5 min, and was resuspended in 50 μl RNase-free water (Ambion). RNAs were digested with DNase I (RNase-free, ThermoScientific) for 1hr at 37 °C to prevent genomic DNA contamination. By adding 0.5mM EDTA to the reaction mixture, samples were heat inactivated at 75 °C for 10mins. Then, RNAs were incubated with isopropanol and GlycoBlue™ (ThermoScientific) at -20 °C overnight. After centrifugation, pelleted RNAs were washed with 95% cold ethanol and centrifuged. RNAs were resuspended in 50 μl RNase-free water (Ambion) and stored at -80 °C for sequencing.

3.4.16 Purification of MS2-MBP fusion proteins

For use as an affinity tag, MS2 coat protein fused with maltose binding protein (MS2-MBP) (Said, Rieder et al. 2009) was expressed in *E. coli*. 100ml of cells were cultured and induced with 1mM IPTG at OD 0.5_{600nm} for 4 hrs. Cells were pelleted and resuspended in 10ml column buffer (20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, 10 mM β–mercaptoethanol pH7.4). 2mM PMSF (phenyl methylsulfonyl fluioride) was added to resuspended cells for preventing protein degradation. After the sonication on ice, DNase I was treated for 1 hr at 4 °C. Cell lysates were centrifuged at 15000 rpm and supernatants (MS2-MBP lysates) were collected. After vortexing and thoroughly suspending amylose magnetic beads (NEB), 200 µl of aliquot was washed with 1ml

column buffer twice. Entire MS2-MBP lysates were incubated with washed amylose magnetic beads for 2~3 hrs at 4 °C. Then, magnet was applied and supernatants were decanted. Beads were washed with 1ml wash buffer (column buffer + 0.1mM maltose) three times. 50 µl of elution buffer (column buffer + 10mM maltose) was added to beads for the elution of MS2-MBP and incubated for 15 minutes at 4°C. By applying magnet, eluted MS2-MBP fusion protein was collected. To increase the yield, elution step was repeated with 50 µL of elution buffer. Purified proteins were confirmed by SDS-PAGE gel and the concentration was measured using Bradford assay.

3.4.17 Affinity purification of MS2-MBP fusion proteins

2ug of purified MS2-MBP proteins were incubated with 100 μl of total RNAs (500ng/μl) extracted from the cells containing 2MS2BD-Zms4/Zms6/control for 1hr at 4 °C. Washed amylose beads were incubated with 2MS2BD-Zms4/Zms6/control+MS2-MBP complex for 2hrs at 4 °C. Supernatants were removed from the beads by applying the magnet. Beads were washed three times with wash buffer and incubated with 50 μl of elution buffer for 15 mins. Elution step was repeated so that total 100 μl of elutions were collected. For the precipitation of RNA, equal volume of isopropanol and 10 μl of GlycoBlue™ was added to elution sample and then, incubated overnight at -20 °C. RNAs were pelleted at 15,000 rpm for 15 mins at 4 °C and washed with 1 ml ethanol. Air-dried RNA pellet was resuspended in 50 μl RNase-free water. RNAs for sequencing were stored at -80 °C.

3.4.18 Transcriptomics data analysis

Prepared RNA was quantified and qualified using Bioanalyzer before sequencing. NEBNext® Multiplex RNA Library Prep Set for Illumina® (New England Biolabs Inc.) was used for generating RNA libraries. Sequencing was performed using Illumina®

NextSeq technology with PE 2*150 run (Genomic Sequencing and Analysis Facility at the University of Texas at Austin). All sequenced libraries were mapped to the *Z. mobilis* 8b complete genome (pending publish) using bwa (0.7.12-r1039) (Li and Durbin 2009).We used three replicate for each sample. Generated sam files were further analyzed using Cuffdiff (v2.2.1 (4237)) (Trapnell, Roberts et al. 2012) to generate normalized count matrix. Analysis followed the procedures and steps described in the package documentation and unless stated otherwise default parameters were used.

Chapter Four

High throughput in vivo sensing of accessible interfaces in a large ensemble of small RNAs reveals Hfq as a universal structural relaxer

**Article in preparation*

4.1 INTRODUCTION

Recent efforts have rendered useful high throughput approaches to understand structure and intermolecular interactions involving RNA in living cells. However, often times profiles of RNA structures only partially explain patterns of regional accessibility, where accessibility is understood as the ability of a given stretch of nucleotides to establish intermolecular interactions that could be important to RNA function. We have developed a novel high-throughput method based on synthetic biology and machine learning approaches to characterize functional RNA structures in living cells. Specifically, we have engineered a system in which accessibility is correlated to transcriptional elongation, as evaluated by RNA-seq. We demonstrate the use of this method in understanding binding interfaces in a variety of RNAs by simultaneously interrogating over 1000 regions in 72 identified bacterial sRNAs *in vivo*. Among the72 characterized sRNAs, only a few have been previously extensively characterized regarding their target mRNAs and their dependency on Hfq. This work reveals patterns of functional structure related to high and low accessibility that likely hallmark regulatory activity. Specifically, our results suggest that interacting regions display either extremely high or low accessibility, preferentially evolve within the most 5' two thirds of the sRNA molecule and harbor a sequence motif reminiscent of the ubiquitous RNA recognition motif YUNR (Franch, Petersen et al. 1999). We also evaluate the contribution of the Hfq

chaperone to natural patterns of accessibility for this sRNA collection and show that it serves as a universal structural relaxer of regulatory RNAs in *E.coli*.

4.2 RESULTS

4.2.1 Harnessing transcriptional regulation for high throughput characterization of RNA accessible interfaces

To allow large-scale identification of accessible RNA interfaces, we constructed a system that couples *in vivo* hybridization to transcriptional elongation control. While assessment of hybridization potential has shown informative in mapping accessible RNA interfaces(Sowa, Vazquez-Anderson et al. 2015)(Vazquez-Anderson J, Mihailovic M, in Review, 2016), we hypothesized that transcriptional elongation control would support high throughput studies by allowing coupling to next generation sequencing. As shown in Figure 1, this system consists of the following main 5 elements: (1) a variable probe region (Probe X), an oligonucleotide (9-26 nt) complementary to a region within a target RNA (taRNA), (2) the RSE, a ribosomal binding site sequestration element blocking (3) the RBS, a strong ribosomal binding site followed by (4) the elongation switch (ES) located directly upstream of (5) the RNA elongation reporter (RER). The ES consists of a small 24-amino acid peptide known as tnaC followed by the rho-dependent transcription terminator rho utilization site (rut). These two components function together to regulate transcriptional elongation by a mechanism known as nascent polypeptide-mediated ribosome stalling(Wilson, Arenz et al. 2016), previously applied to convert translational to transcriptional control(Liu, Qi et al. 2012). The plasmid design of this system is shown in Figure C.1A. Overall, this scheme is based on the premise that differential translation of tnaC, governed by exposure of the RBS, influences transcriptional elongation and this ultimately correlates to the accessibility of the target region. Accessibility is herein

defined as the ability of a given stretch of nucleotides to establish intermolecular interactions. The exposure of the RBS is determined by the stability of the hairpin (RSE paired with RBS), which is controlled by the interaction of the probe region with the corresponding antisense target region, as previously reported(Sowa, Vazquez-Anderson et al. 2015). In this case, if accessible, the target region strongly binds to the probe, destabilizing the hairpin and allowing for the translation of tnaC. Importantly, translation of tnaC leads to ribosome stalling (Figure 4.1 top), thereby blocking the rut site and preventing the Rho factor from binding; in this case, full transcriptional elongation occurs and a full-length transcript is synthesized. In contrast, if inaccessible, the target region does not base pair with the probe region, tnaC is not translated, and the rut site is available for the rho factor to bind and arrest transcription prematurely; this leads to a short partial-length transcript), as shown in Figure 4.1 (bottom). This design offers a measure of RNA accessibility per region suitable to be characterized by varying transcript lengths using RNA-seq. Remarkably, in this way INTERFACE allows for the simultaneous characterization of local accessibility profiles within large assortments of RNAs in the transcriptome. “Accessosome”: a thorough landscape of accessible interfaces throughout any assortment of RNAs in the transcriptome.

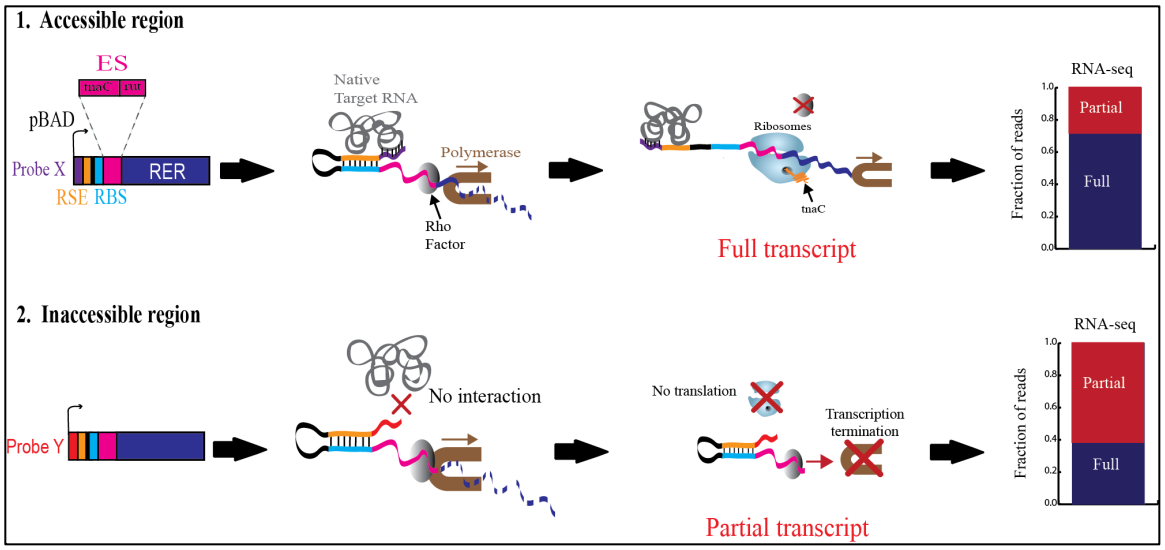


Figure 4.1. Modular engineering of a synthetic transcriptional control for high throughput characterization of RNA accessible interfaces

1.Accessible region: the system considers 5 main components that function in concert to generate a full-length transcript measurable via RNA-seq. In this case, the probe binds strongly, releasing RBS for the translation of tnaC that, in turn, stalls the ribosomes. The stalled ribosomes blocked the Rho factor from binding to the *rut* site and allow transcription elongation. 2. Inaccessible region: in this case tnaC is not translated due to lack of binding between the probe and the target RNA causing the *rut* site to become available. The Rho factor binds the *rut* site preventing transcriptional elongation and ultimately generating a partial-length transcript.

As shown in Figure 4.2, implementation of INTERFACE to RNA characterization consists of 3 main steps: (1) cell collection, (2) RNA extraction, (3) DNA library generation and (4) RNA-seq. First, we transformed a library of variants of the INTERFACE plasmid into a relevant *E. coli* strain; each INTERFACE variant represented an oligonucleotide probe that targets a specific target RNA region (Figure 4.2). We then collected total RNA from the sample and generated a DNA library as outlined in the Methods section (Figure 4.2). Importantly, no fragmentation was

performed to preserve the ability to correctly assign a 3' end to the corresponding 5' end of each transcript and reliably identify the extent of transcriptional elongation for each transcript (based on the target region to which the distinct probe hybridized). Following DNA purification, samples were sequenced using a standard paired-end Illumina-platform protocol. As part of this protocol, the RNA-sequencing data was analyzed for its overall quality and the adaptor sequences were trimmed followed by reads alignment to the differentially elongated sequence (ES in Figure 4.1) with varying probe sequences (Methods). Finally, data was independently filtered for reads containing each probe sequence of interest and corresponding R2 (3' end) reads were paired to R1 (5'end) reads to determine probe-specific transcriptional elongation. In this way, transcriptional elongation was correlated to the accessibility of each target RNA region.

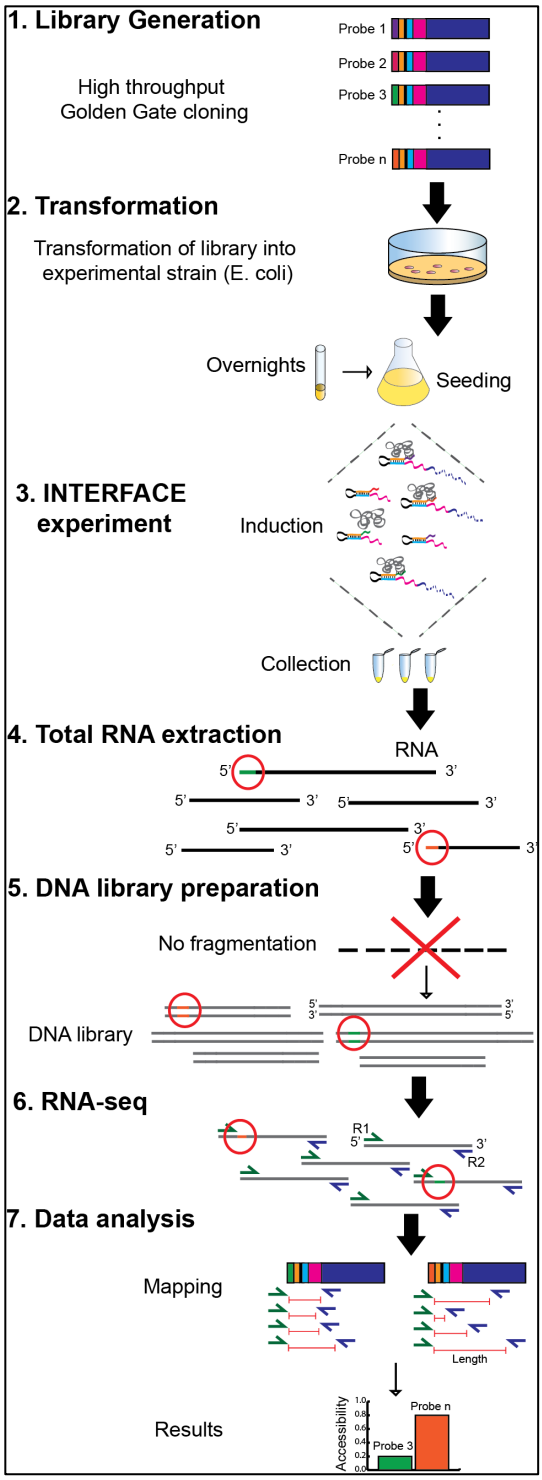


Figure 4.2. INTERFACE experimental workflow.

4.2.2 Validating molecular features governing the ability of INTERFACE to capture regional accessibility

Two basic mechanistic premises enable the INTERFACE approach: (1) that RBS exposure governs transcriptional elongation and (2) that elongation patterns are dependent on the presence of the elongation switch and are reliable indicators of regional accessibility. First, to confirm that RBS exposure governs transcriptional elongation, we designed two control experiments in the presence of a random probe with low genome complementarity: a construct comprised of a permanently sequestered RBS (RSE effectively sequesters RBS due to complementarity) and a construct comprised of a permanently opened RBS (RSE is mutated to free the RBS). Consistent with expectations, the sequestered RBS and the open RBS controls exhibited partial and full transcript lengths, respectively (Figure 3.3A). From Figure 3.3A we observed major peaks present at different loci of the INTERFACE transcript: the first group of peaks (~75 nt) signals the presence of the ES (specifically *tnaC*), the second group of peaks (~140 nt) indicates the presence of the *rut* terminator and finally the last group of peaks (~200 nt) shows up at the RER and ultimately denotes the extent of transcriptional elongation. Second, to test the dependency of differential elongation on the elongation switch, we assayed the accessibility of well-characterized regions(Sowa, Vazquez-Anderson et al. 2015) along the model *Tetrahymena* group I (gI) intron target RNA. Our initial choice of this non-native RNA as a control (expressed in trans, with plasmid shown in Figure C.1B) minimizes potential interference from native RNAs. Figure 4.3B shows representative sample regions within this model RNA, one highly accessible (nucleotides 400-409) (Figure 4.3B top) and one highly inaccessible (nucleotides 361-380) (Figure 4.3B bottom), as previously reported(Sowa, Vazquez-Anderson et al. 2015). Importantly, we observed the expected shift towards longer transcripts for the most

accessible region, only in the presence of the elongation switch (Figure 4.3B). Moreover, this bias towards larger transcript sizes was absent from transcriptomic data when probing the highly inaccessible region. It is worth noting that these transcript size biases were only observed in the presence of a target RNA, indicating our ability to gather specific information about a target RNA of interest *in vivo*.

To evaluate the expected high throughput potential of this approach, we fully characterized the gI intron in a single experiment by evaluating a library containing INTERFACE plasmids for 30 independent probes (in a combination of 9-mers and 16-mers for 100% coverage) (Figure 4.3C, Figure C.2). These results are consistent with evidence from previous studies(Zarrinkar and Williamson 1994, Doherty and Doudna 1997, Sowa, Vazquez-Anderson et al. 2015)(Vazquez-Anderson J, Mihailovic M, in Review, 2016). In particular, as shown in Figure 4.3C, we demonstrate the ability of the INTERFACE approach to similarly identify highly transient and dynamic regions in the unique way of *in vivo* hybridization methods (distinct from single-nucleotide probing methods as they bear the ability to capture low-abundance dynamic regions), as previously reported(Sowa, Vazquez-Anderson et al. 2015).

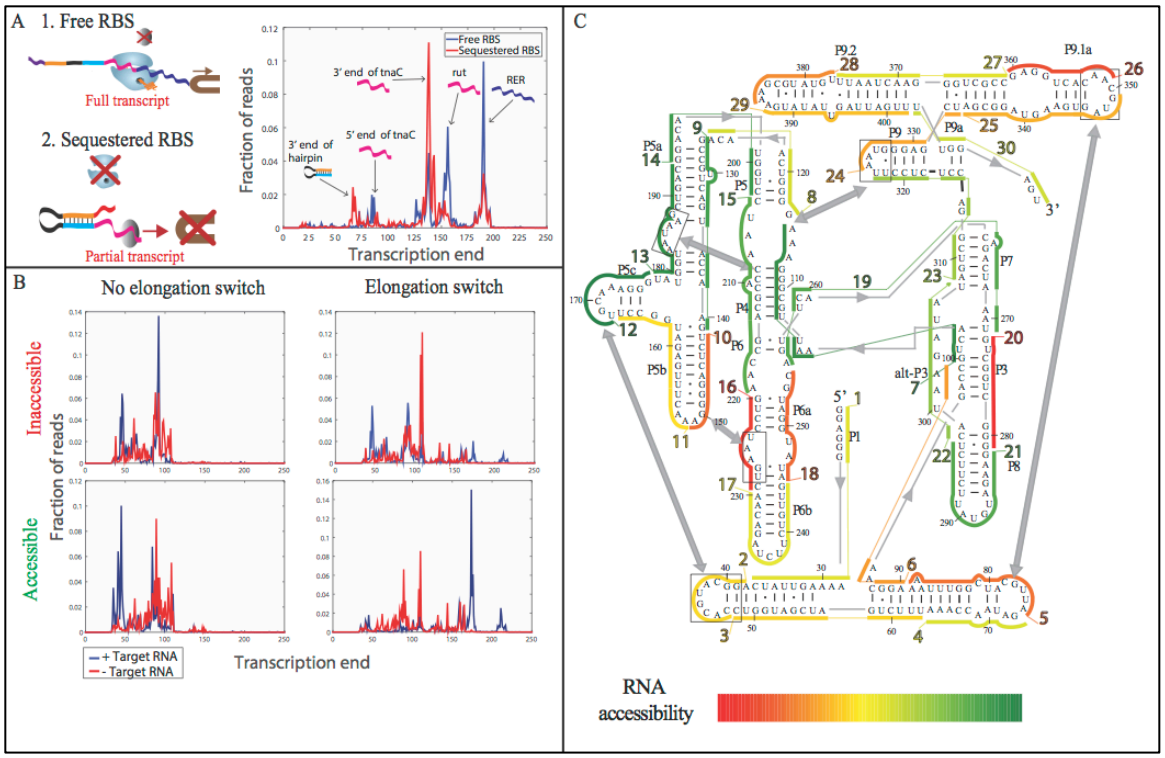


Figure 4.3. INTERFACE allows for high-throughput characterization of functional structure.

A. RBS exposure controls transcriptional elongation. A sequestered RBS control shows differentially less transcriptional elongation than the free RBS control as per the weighted average of the length per read calculated from RNA-seq results. B. The inclusion of the Elongation Switch (ES) enables differential transcriptional elongation that correlates with accessibility. C. INTERFACE is capable of fully characterizing an RNA molecule in a single experiment. The model RNA gI intron was characterized showing similar results to previous studies.

4.2.3 Large-scale characterization of accessible interfaces in native regulatory RNAs aided by machine learning reveals potential functional regions

To understand universal molecular features of sRNAs features that contribute accessible local surfaces for *in vivo* interactions, we designed a single INTERFACE experiment to collectively characterize accessibility profiles in a large group of trans-

encoded sRNAs. To simultaneously minimize the number of experimentally probed target regions while optimizing the information collected regarding dynamic sRNA regulatory interfaces, a machine-learning algorithm was coupled to a biophysical model to select optimal experimental target regions (Figure 4.4A). We considered optimal experimental target regions to be those more likely to show dynamic behavior in terms of their likelihood to form intermolecular interactions, indicated by the most and least accessible regions (as previously shown, (Sowa, Vazquez-Anderson et al. 2015) (Vazquez-Anderson J, Mihailovic M, in Review, 2016)). Approximately 70,000 potential targeting regions along 72 sRNAs (Table C.1) were initially generated with a length constraint of 9-16 nt. The predicted accessibility of each region was evaluated using an adapted version of a previously developed biophysical model that predicts regional hybridization potential (Vazquez-Anderson J, Mihailovic M, in Review, 2016). Predicted regional hybridization potentials were then provided to a machine-learning algorithm, known as sparse knowledge gradient (KG)(Li, Liu et al. 2015, Li, Liu et al. 2016), to provide experimental suggestions based on value-of-information analysis by combining Bayesian optimization problem with a regularized regression approach, Lasso(Tibshirani 1996) (Methods). In this way, we collected an initial list of suggested regions for probing within each target sRNA, fulfilling the constraints of: (1) minimized target region overlap, (2) target region length specification (9 – 16 nt), and (3) full coverage of each sRNA molecule. As shown by simulations that compared the selection of dynamic interfaces obtained by this machine learning algorithm relative to random design (exploration) or to the adapted biophysical model (exploitation) only (Figure 4.4B), the inclusion of this computational approach in our experiment reduced our probing efforts by enriching the number of interesting (dynamic) regions that were assayed. Ultimately, we selected the top KG-ranked (~971) regions from these predictions for a full coverage

of the target 72 sRNAs. These 72 sRNAs were selected using carefully curated databases(Li, Huang et al. 2013, Wang, Liu et al. 2015) for which experimental evidence of their existence has been reported (Table C.1). The INTERFACE accessibility maps are shown in Figure C.3. Among the 72 characterized sRNAs, no more than a couple dozen have been extensively characterized regarding their target mRNAs and their dependency on Hfq.

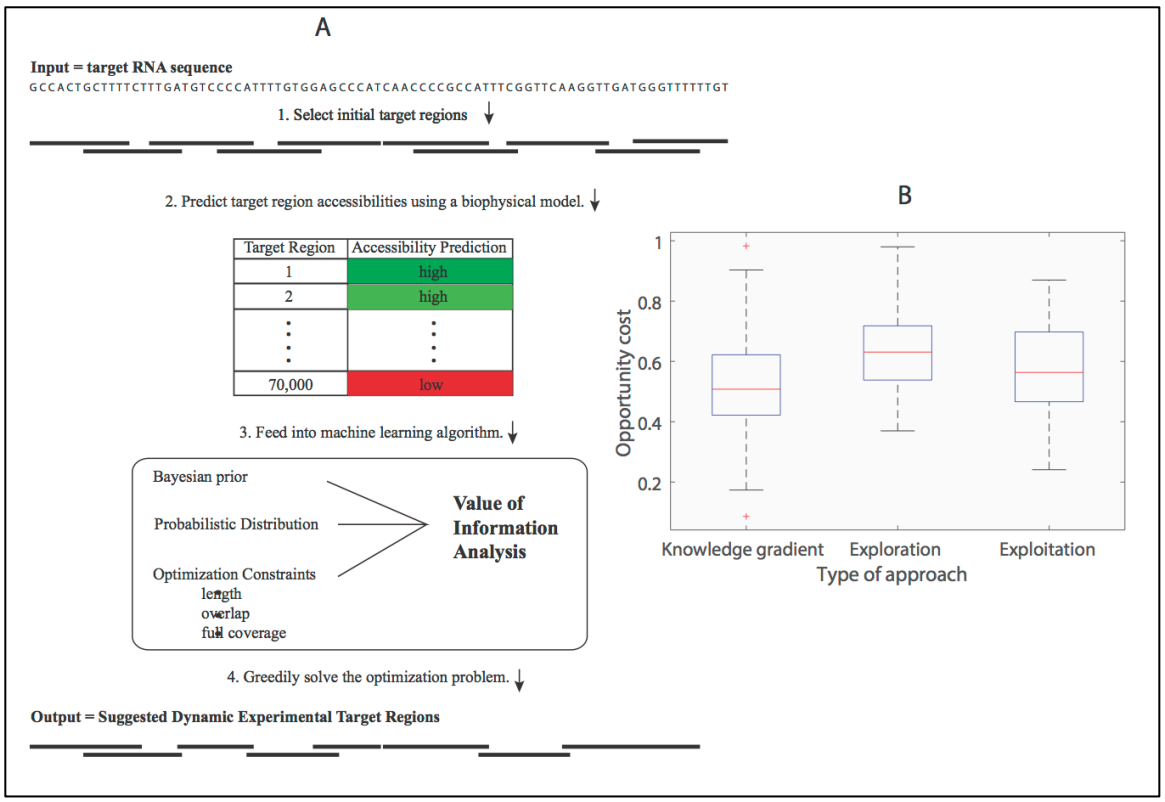


Figure 4.4. Characterizing the acessosome for a large ensemble of sRNAs.

A. Workflow for machine-learning algorithm to select for accessible interfaces. B. Our machine-learning algorithm based on knowledge value outperforms the exploration and exploitation strategies (p-value<0.001, 2-tailed t-test). Using exploitation (only the biophysical model) requires less experimental effort than the exploration approach (p-value<0.001, 2-tailed t-test).

To assess how the collected accessibility data related to the sRNA functional interfaces, we analyzed the acquired profiles for 21 sRNAs (with a total of 66 local regions) that have been well characterized in terms of their experimentally confirmed binding sites. Using the sRNATarBase(Wang, Liu et al. 2015), we collected exhaustive information on experimentally confirmed mRNA binding sites this set of of 21 sRNAs (Table C.1). Importantly, we found up to twice as many regions harboring sRNA-mRNA binding sites within the low (<0.3) and high (>0.7) INTERFACE-assessed intervals, respectively, relative to the mid range ($0.3 \leq mid \leq 0.7$). Additionally, we observed a clear preference of these binding sites to fall within the first two thirds of the sRNA molecule (5' to 3' end) (Figure 4.5A).

To further differentiate sRNA-mRNA binding sites, we performed a sequence motif search using the GAM2 tool from the MEME suite (Frith, Saunders et al. 2008) and discovered a highly prevalent sequence motif (ADUCA) shown in Figure 4.5B (see Methods for details). These results suggest that interacting regions tend towards extremes in regards to accessibility, preferentially evolve within the most 5' two thirds of the sRNA molecule and harbor a sequence motif reminiscent of the ubiquitous RNA recognition motif YUNR (Franch, Petersen et al. 1999). Altogether these three observations could be a strong indicator of the presence of sRNA-mRNA binding sites.

As shown in Figure 4.5A, INTERFACE-determined accessibility profiles capture well experimentally confirmed binding sites as either highly or lowly accessible. This reflects the possibility that highly active sites are either *actively* bound by a cellular

factor/target (and therefore lowly accessible) or are configured as regions that are highly available for binding (displaying high accessibility). Indeed, low accessibility, characteristic of *active* regulatory control, was observed in known binding sites of sRNAs known to be active under the experimental conditions of this study. For instance, the exponential growth conditions of this study likely require replacement of select outer membrane proteins (OMP) to maintain structural integrity(Arunasri, Adil et al. 2014). Consistent with this physiological expectation, we captured low accessibility (0.14) of the region within the rseX sRNA that regulates *ompA* (region 5 of rseX, Figure 4.5C) indicating the inability of the INTERFACE to access the binding region due to active ompA upregulation. Similarly, the lrp *repression* region within the MicF sRNA (region 2 of MicF, Figure 4.5C) appears to be active under these experimental conditions based on observed low accessibility (0.2), which is expected of the double negative Lrp-MicF feedback loop in a *nutrient-rich* environment(Holmqvist, Unoson et al. 2012). In contrast, active regulation by the DsrA and Spot 42 sRNAs of stress-responsive mRNAs(Lease, Smith et al. 2004) is not anticipated under the experimental conditions used in this study. This is consistent with the observed INTERFACE-determined high accessibility in DsrA and Spot 42 regions involved in regulation of *rpoS* (region 3 of DsrA, Figure 4.5C), *hns* (region 7 of DsrA, Figure 4.5C), and sugar-responsive mRNAs(Beisel and Storz 2011), such as *gltA* (region 5 of Spot 42, Figure 4.5C) and *sucC* (region 9 of Spot-42, Figure 4.5C), respectively. The sensitivity of INTERFACE to in vivo cellular conditions unveils the prospects of using this method to map dynamic functional accessibility changes of full regulatory networks.

After establishing the value of identifying lowly and highly accessible regions as a way to establish their functionality within the molecules for establishing regulatory interactions, we collected INTERFACE profiles for the remaining ~51 uncharacterized sRNAs. Importantly, we use the patterns learned from our profiling of well-characterized sRNAs to propose potential mRNA binding sites for these more obscure sRNAs. To this end, we first classify all regions based on the same criteria used in Figure 4.5A. Next, using GAM2scan form the MEME suite (Frith, Saunders et al. 2008) to search for highly similar sequences to the motif in Figure 4.5B (score>=5 according to GAM2scan results) in the regions contained within each grouping. We scored each bin in Figure 4.5D using the normalized (0 to 1) prevalence of these motifs in the pool of regions within each bin. Figure 4.5D shows a strong match in the pattern observed in Figure 4.5A suggesting high potential for INTERFACE accessibility to be used as a predictor of functional regions.

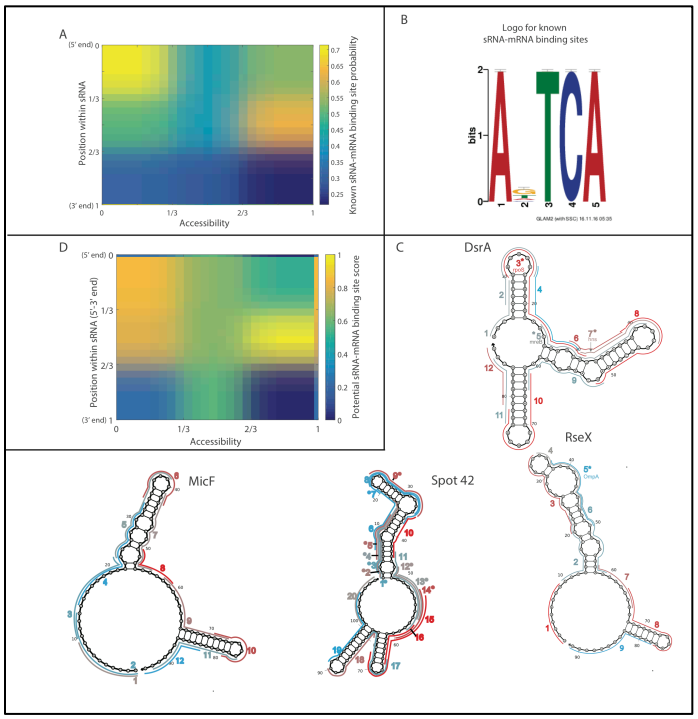


Figure 4.5. Large-scale characterization of accessible interfaces in native regulatory RNAs reveals functional regions

A. An analysis of sRNA-mRNA known binding sites in 18 sRNAs (Table C.1) shows an enrichment of binding sites in the 2/3 most 5' end and in the extremes of accessibility (1/3 most inaccessible and 1/3 most accessible). B. A sequence logo obtained using the webserver MEME (Frith, Saunders et al. 2008). Specifically the tool GAM2 was employed to find gapped sequence motifs, a significant motif was found in the regions harboring known sRNA-mRNA binding sites. In a parallel analysis no significant motifs were found in the regions not known to harbor binding sites. C. Accessibility maps for dsrA, Spot-42, rseX, and MicF are representative of the conclusions drawn from previous data. Specifically binding sites are shown to preferentially appear in regions either highly or lowly accessible. D. Analysis of the rest of sRNAs shows a strong correlation with the distribution of known binding sites in terms of position, accessibility and the sequence motif found previously, supporting the prospects of using INTERFACE accessibility as a predictor of functional regions.

4.2.4 Hfq facilitates sRNA-mRNA interaction by releasing target sRNA structure and increasing accessibility

To understand the global impact of Hfq chaperoning within this large set of sRNAs, we also mapped the INTERFACE accessibility of the 971 regions within the 72 sRNAs to both the wild type BW25113 strain (as described above) and an hfq-knockout strain (JW4130-1(Baba, Ara et al. 2006). Upon collecting information available for all 72 sRNAs studied Table C.1), we found that approximately 34 sRNAs have a confirmed dependency on Hfq in *E. coli*; of these, exact binding sites have been experimentally confirmed for less than half and approximately 16 have been proposed to be Hfq-independent. Interestingly, although more than 20% of sRNAs in our experimental set are believed to be Hfq-independent, the distribution of INTERFACE accessibility changes upon the presence of Hfq (parent – ΔHfq strain), where we observe a clear skewed towards positive values (Figure 4.6A). Importantly, this marked accessibility change in the absence of Hfq supports a previously-proposed hypothesis in which Hfq serves as a structural-releasing chaperone for its target sRNAs(Ishikawa, Otaka et al. 2012). Importantly, upon comparison of the number of regions affected across strains between Hfq-dependent and Hfq-independent sRNAs, we observed a stronger Hfq-dependent skew in INTERFACE determined accessibility for regions within known Hfq-dependent sRNAs relative to regions within Hfq-independent sRNAs (representative examples shown in Figure 4.6B). The effect of Hfq on these regions was assessed by tallying the number of regions where a skew in accessibility was observed, lower p-values < 0.05 and higher magnitude of accessibility change).

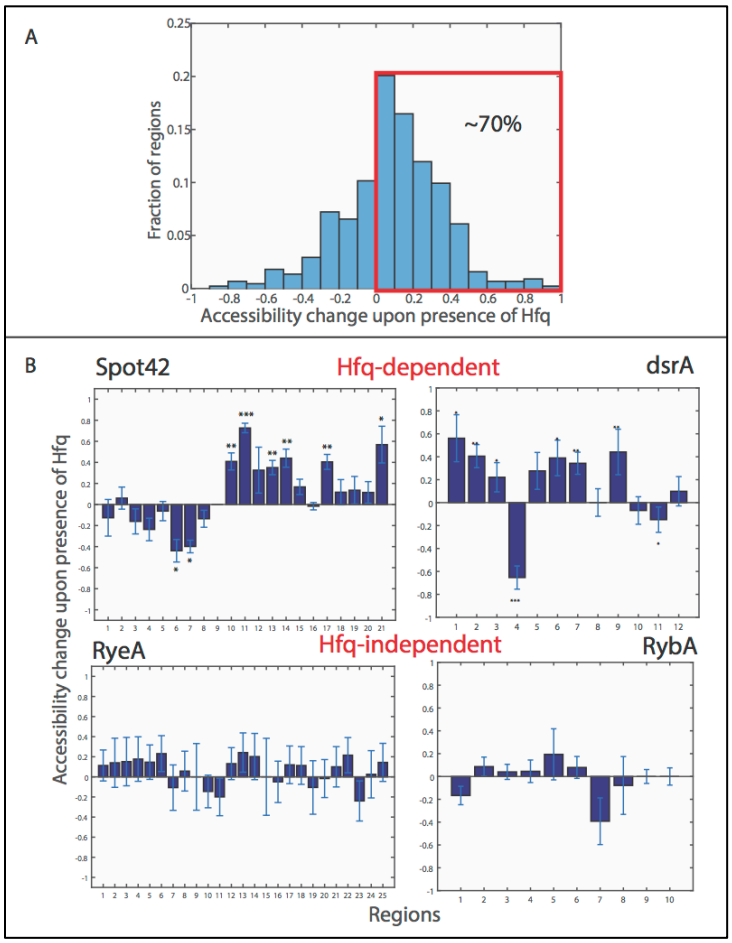


Figure 4.6. INTERFACE analysis reveals structural changes upon protein binding in small RNAs

A. INTERFACE reveals role of Hfq as a structural relaxer of sRNAs indicated by the skew in the positive difference in accessibility upon presence of Hfq relative to the absence of Hfq (parent – ΔHfq). Approximately 70% of the all regions analyzed (~970) fell within the positive interval of the difference in accessibilities across strains (parent – ΔHfq). B. INTERFACE is sensitive to different levels of Hfq dependency. Bar graphs represent the changes in accessibility upon the presence of Hfq, stars indicate statistical significance (* p-value< 0.05, ** p-value<0.01 and *** p-value<0.001 by the two-tailed t-test). Spot42 a strong Hfq-dependent sRNA shows global effects in its accessibility patterns versus RyeA, an Hfq-independent sRNA, with no significant changes in its structure.

Founded on our ability to capture Hfq-dependency, we propose an INTERFACE-based level of Hfq-dependency in accordance to the following two characteristics per sRNA for all uncharacterized sRNAs: (1) the fraction of regions showing significant accessibility changes between wt and ΔHfq strains and (2) the maximum absolute regional accessibility difference between wt and ΔHfq strains (Table 4.1) (see Methods for details). Interestingly, the INTERFACE-based Hfq dependency classification is mostly consistent with dependencies characterized by quantitative metrics including competition assays, co-immunoprecipitation, and electrophoretic mobility shift assays (for ~ 20 sRNAs for which relative Hfq-binding strength information is available) (Table 4.1). These observations altogether support strong prospects for the application of INTERFACE to the characterization of Hfq-like regulators and their effects on RNA binding partners.

Table 4.1. INTERFACE reveals differential dependency of Hfq for all sRNAs analyzed and correlates strongly with pull down data obtained from the literature.

sRNA	Hfq dependen	INTERFACE-determined Hfq Interaction Strength	Literature-determined Hfq Interaction Streng	Hfq dependency Reference	Hfq-interaction Strength Reference	comments
sroB (ChiX)	Yes	weak	strong	Moon and Gottesman (2011)	Moon and Gottesman (2011)	strong in comparison to Spot42, ArcZ, CyaR, GcvB, MgrR, DsrA
RybB	Yes	mid	strong	Zhang, et al. (2003)	Wassarman, et al. (2001)	
RydC	Yes	weak	strong	Zhang, et al. (2003)	Olejniczak (2011)	
RprA	Yes	weak	strong	Tree, et al. (2014)	Olejniczak (2011)	
RyeB (sdsr)	Yes	strong	strong	Zhang, et al. (2003)	Wassarman, et al. (2001)	
SraD (MicA)	Yes	strong	strong	Tree, et al. (2014)	Olejniczak (2011)	
SraH (ArcZ)	Yes	strong	strong	Zhang, et al. (2003)	Soper, et al. (2010)	Higher affinity than rprA which is considered to interact with Hfq strongly
GlmZ	Yes	weak	strong	Tree, et al. (2014)	Göpel, et al. (2013)	Binds to Hfq with high affinity
Spot_42	Yes	strong	strong	Kim, et al. (2015)	Olejniczak (2011)	
MicF	Yes	weak	strong	Zhang, et al. (2003)	Olejniczak (2011)	
GcvB	Yes	strong	mid	Tree, et al. (2014)	Moon and Gottesman (2011)	Lesser affinity for Hfq than ChiX
dsrA	Yes	strong	mid	Tree, et al. (2014)	Olejniczak (2011); Soper, et al. (2010); Moon an	Mid in competition assay; weak in comparison to rprA; weak in comparison to ChiX
ryhB	Yes	strong	mid	Tree, et al. (2014)	Olejniczak (2011)	
MgrR	Yes	strong	mid	Kim, et al. (2015)	Moon and Gottesman (2011)	Lesser affinity for Hfq than ChiX
CyaR	Yes	strong	mid	Tree, et al. (2014)	Moon and Gottesman (2011)	Lesser affinity for Hfq than ChiX
GlmY	Yes	mid	weak	Göpel, et al. (2015)	Göpel, et al. (2013)	Much weaker affinity than GlmZ
istR-1, istR-2	Yes	strong	weak	Olejniczak (2011)	Olejniczak (2011)	
OxyS	Yes	weak	weak	Tree, et al. (2014)	Olejniczak (2011); Henderson, et al. (2013)	Weak in competition assay; weak interaction compared to rprA-Hfq
dicF	Yes	weak	weak	Zhang, et al. (2003)	Olejniczak (2011)	
RyeA (SraC)	Yes	N/A	weak	Pandey, et al. (2014)	Wassarman, et al. (2001)	
tpke11	Yes	strong	N/A	Zhang, et al. (2003)		
SgrS	Yes	mid	N/A	Ishikawa, et al. (2012)		
McaS	Yes	weak	N/A	Jorgensen, et al. (2013)		
FnrS	Yes	strong	N/A	Tree, et al. (2014)		
MicC	Yes	strong	N/A	Tree, et al. (2014)		
RyeF	Yes	mid	N/A	Zhang, et al. (2003)		
rseX	Yes	strong	N/A	Kim, et al. (2015)		
OmrA	Yes	weak	N/A	Tree, et al. (2014)		
OmrB	Yes	weak	N/A	Tree, et al. (2014)		
GadY	Yes	mid	N/A	Kim, et al. (2015)		
ryjB	Yes	weak	N/A	Zhang, et al. (2003)		
MicL	Yes	mid	N/A	Guo, et al. (2014)		
IpeX	Yes	weak	N/A	Catillo-Keller, et al. (2006)		
ryjA	Yes	weak	no detectable binding	Wassarman, et al. (2001)	Wassarman, et al. (2001)	
RygC	No	mid	N/A	Pandey, et al. (2014)		
ffs (4.5S)	No	strong	N/A	Zhang, et al. (2003); Pandey, et al. (2014)		
RybA (mntS)	No	weak	no detectable binding	Gerstle, et al. (2012)	Wassarman, et al. (2001)	
psrD	No	strong	N/A	Pandey, et al. (2014)		
rdlA	No	weak	N/A	Pandey, et al. (2014)		
rdlB	No	weak	N/A	Bak, et al. (2015)		
rdlC	No	weak	N/A	Bak, et al. (2015)		
rydB	No	weak	N/A	Pandey, et al. (2014)		
isrB	No	mid	N/A	Pandey, et al. (2014)		
isrC	No	mid	N/A	Pandey, et al. (2014)		
ryfA	No	weak	N/A	Pandey, et al. (2014)		
ryfB	No	weak	N/A	Bak, et al. (2015)		
RygD (sibD)	No	mid	N/A	Pandey, et al. (2014)		
symR	No	strong	N/A	Kawano, et al. (2007)		
ryfD	No	mid	N/A	Pandey, et al. (2014)		
rdlD	No	mid	N/A	Pandey, et al. (2014)		
sroC	N/A	weak	N/A	Papenfort and Vanderpool (2015)		
arrS	N/A	strong	N/A			
SraL	N/A	strong	N/A			
SroA	N/A	strong	N/A			
tp2	N/A	mid	N/A			
tff	N/A	mid	N/A			
nc2	N/A	strong	N/A			
sokB	N/A	mid	N/A			
sroD	N/A	N/A	N/A			
tpke70	N/A	mid	N/A			
sroE	N/A	strong	N/A			
ryfC (ohsC)	N/A	strong	N/A			
InvR	N/A	weak	N/A			
sroG	N/A	mid	N/A			
psrN	N/A	weak	N/A			
sroH	N/A	mid	N/A			
nc5	N/A	weak	N/A			
SibB	N/A	weak	N/A			
SibE	N/A	mid	N/A			
SibA	N/A	mid	N/A			

4.3 DISCUSSION

A synthetic transcriptional control was coupled to a reporter system to allow for the evaluation of accessibility of 971 sRNA interfaces via RNA-sequencing, including 66 previously identified mRNA binding regions. Our results suggest that INTERFACE can aid in determination of functional regions in RNA, as well as sense structural changes that support intermolecular interactions e.g. Hfq-dependency. The ability of INTERFACE to capture dynamic behavior of interacting interfaces within sRNAs on this larger scale supports a hypothesis in which regions that are actively being occupied due to interaction with target mRNAs are rendered inaccessible to the INTERFACE probe, while interaction interfaces which are not active appear highly accessible to the INTERFACE probe (Figure 4.5A). This phenomenon, which could be explained by competitive binding between natural targets of the sRNA and the reporter probe, upholds the use of the INTERFACE to sense active or inactive RNA-RNA interactions and aid in determination of functional regions.

As for the role of Hfq in sRNA chaperoning, our results strongly suggest a universal role of Hfq as a structural relaxer. Moreover, since no regional effects where observed (no reduced accessibility in Hfq binding sites), INTERFACE results highly supports a hypothesis for the Hfq chaperoning in which Hfq only binds utilizing only 2-3 nt(Dimastrogiovanni, Frohlich et al. 2014) and unfolds the sRNA to facilitate binding to its target mRNA. Paradoxically, Hfq-independent sRNAs still showed significant changes, although weakened, supporting at least an indirect effect by Hfq on these regulatory RNAs. These striking findings, to the best of our knowledge, are the first to confirm in a global perspective of the functional RNA structure the roles of Hfq.

It is clear that, by exploiting the ability of this system to sense RNA-RNA and RNA-protein interactions, further understanding of regulatory RNA networks can be

attained. The accessosome provided by INTERFACE sits at the core of the RNA function-structure relationship (Figure 4.7). Potential applications of information gained by such research include synthetic antisense RNA design, characterization of stress-responsive sRNAs, and even metabolic engineering. Identification of regions that exhibit functional characteristics identified in this study (low/high accessibility), which have not previously been linked with regulatory function, could enable such discovery upon further targeting with synthetic antisense RNA (and monitoring gene expression changes). Using INTERFACE under varying environmental conditions could grant identification of stress response of various regulatory RNAs. Combining knowledge of functional/regulatory regions, protein dependency, and activity within a multitude of RNAs concomitantly supports system modeling, informing subsequent genomic engineering for fine, multiplex tuning which would be invaluable to metabolic engineering. The breadth of applications contingent upon information obtained by RNA hybridization accessibility patterns constitutes the need for high-throughput methods such as INTERFACE.

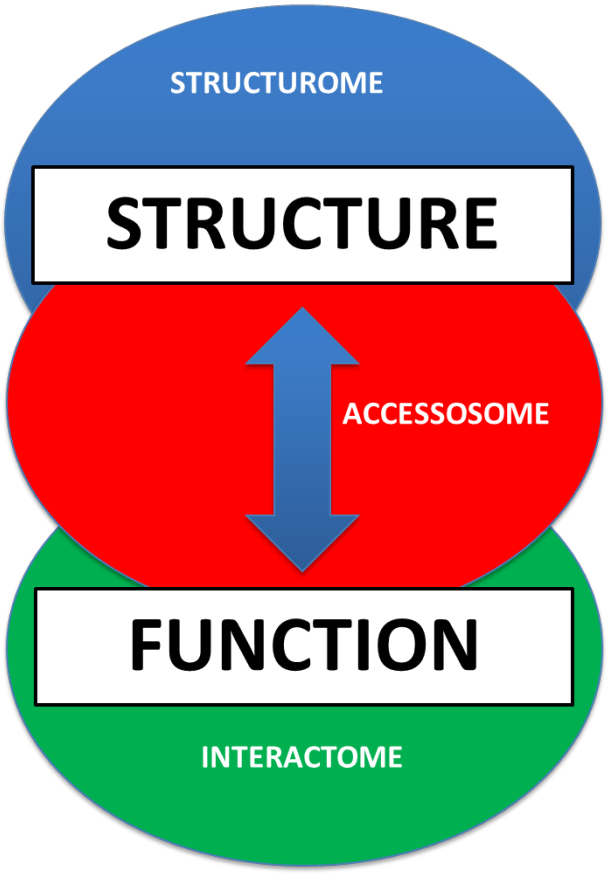


Figure 4.7. The accessosome sits at the "core" of the structure-function relationship

4.4. METHODS

4.4.1 Plasmids and strains

Three different E. coli strains were used in this work: K-12 MG1655 for the experiments performed via overexpression of the gI intron to establish the technique, BW25113 (Keio collection parent strain(Baba, Ara et al. 2006)) and the Hfq-deficient strain (*ΔHfq* JW4130-1 from the Keio Collection(Baba, Ara et al. 2006)) for experiments performed to characterize accessible interfaces in native RNAs. The Hfq-deficient strain was "cured" following a FLP recombination protocol(Baba, Ara et al. 2006) and the knockout was confirmed via genomic PCR. Two main plasmids were constructed

departing from the original "Wild Type Intron Probe I reporter-(WTIPIR)" plasmid(Sowa, Vazquez-Anderson et al. 2015) for this work: INTERFACE (for native target RNAs) and Overexpression INTERFACE (for overexpressed target RNAs) (Figure C.1A and 1B respectively). These constructs mainly differ from WTIPIR in that we introduced (1) a p-chlorophenylalanine negative selection (PheS) cassette in place of the asRNA sequence (between EcoRI and the CB element flanked by two BsmbI restriction sites) for cloning optimization purposes (Kast and Hennecke 1991, Kast 1994), (2) an adaptor containing a *tnaC* sequence and a rho-dependent terminator *rut* (rho utilization site) and (3) a truncated version of GFP (reporter) where the start and end codons have been preserved to allow for transcript size characterizations.

4.4.1.2 Synthesis of constructs

All probes were introduced in the different plasmids by using a Golden Gate (GG) cloning scheme (Vazquez-Anderson J, Mihailovic M; in review). Thirty new probes targeting the gI intron were cloned into the HT-GG iRS[3] plasmid using GG cloning. Using the same approach, five previously designed probes (original-4, 6, 7, 9 and 10) (Sowa, Vazquez-Anderson et al. 2015) were introduced into a version of the Overexpression INTERFACE plasmid lacking the elongation switch for the control experiments in Figure 4.3B. For the synthesis of the 1016 probe-library (sRtar) targeting 72 experimentally confirmed small RNAs in E. coli and the glutamate tRNA (used as a control), we utilized a high throughput version of Golden Gate cloning in which we combined up to 10 probes per synthesis reaction into the INTERFACE. We synthesized the sequestered and free RBS controls by introducing a "no interaction" probe (a randomized 15-mer tested for minimum complementarity to genome in E. coli) into the INTERFACE. For the free RBS control, a randomized RSE was introduced via Gibson

assembly(Gibson 2011) to prevent formation of the stem-loop that serves to block the RBS. All constructs were initially synthesized using *E. coli* electro-competent cells (Turbo, NEB) and subsequently transformed into the appropriate experimental strain. Each LB-agar plate containing the sRtar library (~100 plates) was thoroughly scraped using LB. The resulting combined "goop" was recovered for 1.5 h at 37 °C and stored at -80 °C with 30% glycerol. Sequencing of each individual clone was used to confirm each construct except for the sRtar library, for which diversity was confirmed at different steps of the library generation by sequencing randomly-selected subsets of colonies (>60). Clones for the sRtar library were further individually verified upon RNA-seq analysis, allowing confirmation of >95% and >98% of clones for parent and DHfq strains, respectively. The inability to confirm a small percentage of clones was likely due to limited sequencing depth.

4.4.2 INTERFACE experiments

INTERFACE experiments were performed following the previously reported protocol(Sowa, Vazquez-Anderson et al. 2015). For the experiments used to establish the system with an overexpressed target RNA (Overexpression INTERFACE), we made individual overnights (biologically independent samples) for each construct, equal parts of each resulting culture were combined and 400 µL of this mixture were seeded into 40 mL of LB. The cell cultures were run by triplicates and at four different induction conditions (N=12): (1) no anhydrotetracycline (aTc), no arabinose (ara); (2) 20 µL of aTc (final concentration: 100 ng/µL), no ara; (3) no aTc, 800 µL of 20% ara (final concentration: 0.8%) and (4) 20 µL aTc, 800 µL 20% ara.

For the experiment intended to characterize native target RNAs (INTERFACE), 100 mL of LB were seeded with 600 µL of the sRtar library directly from gradually-

thawed freezer stocks. 400 µL of each of the following control constructs were distributed to the seeded library solution: (1) free RBS (overnight), (2) sequestered RBS (overnight), and (3) the libraries of probes targeting the following molecules (stored individually): CsrB, glu-tRNA, DsrA and RyhB (freezer stocks). Samples were grown in triplicates, under two induction conditions (no ara and 2 mL of 20% ara), and in two separate strains (parent and DHfq) for a total of 12. Kanamycin was added to all cultures to obtain a final concentration of 50 µg/mL. In both experiments samples were induced 1-1.5 h post-seeding, recovered 5 hours post-induction, and immediately processed for total RNA extraction.

4.4.2.3 Total RNA extraction

Following collection of samples 5-hours post-induction, total RNA was extracted from a sample of 1-5 mL of culture as per the protocol in (Hee Cho et al., 2014). Next, total RNA samples were treated with RNAse-free DNase I (PI-90083 Thermo Fisher Scientific Inc.). After DNase I treatment, 10 µL of GlycoBlue (AM9516 Life Technologies) were added to an equal volume solution of isopropanol (brand) and RNA samples (55 µL). Ethanol precipitation was then performed as described in (Cho, Lei et al. 2014). Finally, the quality of RNA was evaluated by using a bioanalyzer (Agilent) at the Genomic Sequencing and Analysis Facility (GSAF at UT Austin) to confirm no significant degradation had occurred.

4.4.4 Computation selection of accessible interfaces

4.4.4.1 Estimation of binding potential using a biophysical model

We used an initial, un-optimized version of a model reported in (Vazquez-Anderson J, Mihailovic, M; in Review) to explain hybridization efficacy, ν, obtained via regression analysis, as follows:

$$\upsilon = \bar{\theta}\left(\Delta G_{tf} - \Delta G_{asT}\right) + \Delta G_{asF}$$

In this model, the ΔG terms represent the free energies which must be considered for the interaction of the target region with the reporter probe, in which subscripts "tf," "asT," and "asF" represent target unfolding, binding between the reporter probe and target, and reporter probe unfolding, respectively. The model also includes a pseudo frequency factor ($\bar{\theta}$) to account for the global ensemble of structures within the target region. This term is evaluated at a *regional* level and thus calculated as the summation of each nucleotide's local accessibility over the length of the target region, as estimated by base-pairing probabilities from the using the AllSub subroutine in the RNA-structure webserver (Wuchty, Fontana et al. 1999, Duan, Mathews et al. 2006, Reuter and Mathews 2010).

4.4.4.2 Machine learning algorithm

To optimally select for accessible regions in 72 experimentally-confirmed bacterial small RNAs in E. coli, we adapted a machine-learning algorithm called sparse knowledge gradient (SpKG)(Li, Liu et al. 2015, Li, Liu et al. 2016) to a weighted set cover problem. The SpKG algorithm is developed to solve the sequential ranking and selection problem, in which, at each time period, one or several experimental suggestions are provided based on value-of-information analysis by taking into account the new observations. It can be used to adaptively select the targeted regions within a large molecule to identify which regions are more amenable to establish interactions with other molecules (Li, Reyes et al. 2015). However, in this setting in which many target regions need to be suggested before the experiments, we adapt the SpKG algorithm to a weighted set cover problem to maximize the value of information of all suggested probes that can provide a full coverage of each molecule and have minimum overlap.

In the following, we provide the mathematical formulation of the problem. For any RNA molecule with length L, suppose there are n potential targeting regions with the length specification. Let $[i_1, j_1], [i_2, j_2], \dots, [i_n, j_n]$ be intervals on $[1, L]$ that denote these n potential target regions. Let $x_1, x_2, \dots, x_n \in \{0,1\}$ be binary variables that denote either the kth target region is selected or not. We use $v_k, k = 1, \dots, n$ to represent the knowledge gradient value of the kth target region. These values can be computed via the SpKG algorithm described in(Li, Liu et al. 2016). Our optimization problem can be written as

$$max \sum_{k=1}^{n} v_k x_k - \lambda \sum_{k=1}^{n} x_k$$

$$s.t. \bigcup_{\{k:x_k=1\}} [i_k, j_k] = [1, L]$$

$$x_k \in \{0,1\} \; for \; all \; k = 1, \dots, n.$$

Here λ is a tunable parameter that penalizes the number of target regions selected to minimize overlapping. This optimization problem is essentially a weighted set cover problem. It is one of the Karp's 21 NP complete problems and cannot be solved in polynomial time. In order to solve it efficiently, we use a greedy algorithm, which was first analyzed in(Johnson 1974), to approximately solve it.

Algorithm:

1. $\mathrm{C} \leftarrow \emptyset, \mathrm{I} \leftarrow \emptyset$. (Here C is the set of nucleotides covered so far; I is the set of index for the selected targeting regions.)
2. While $\mathrm{C} \neq [1, \mathrm{L}]$ do
   for all $\mathrm{k} \notin \mathrm{I}$, let $\alpha_k = \frac{\lambda - v_k}{|[i_k, j_k] - C|}$
   choose $k^* = argmin \; \alpha_k$
   update $\mathrm{C} \leftarrow \mathrm{C} \cup [i_k, j_k], \mathrm{I} \leftarrow \mathrm{I} \cup \{k^*\}$.

3. Output.

4.4.4.3 Computational simulations to test for algorithm performance

To quantify the performance of this algorithm in synthetic simulations, we compare it with a 2 comparatively naïve algorithms—exploration and exploitation. The exploration algorithm generates the probes with length 12 uniformly for each RNA molecule. In the exploitation algorithm, target regions are selected as those with highest predicted accessibility using the aforementioned biophysical model. Additionally, identical coverage and overlap constraints to the SpKG algorithm were imposed. In these controlled synthetic simulations, we sample the true accessibility coefficients from a stochastic process. Taking into account the noise of each experiment, we then normally sample the observations from the three sets of target regions generated using the SpKG, the exploration (uniform), and the exploitation algorithms. We consider a metric called opportunity cost, which provides an estimation of "how far" we reach in terms of identifying the most accessible target region in a given number of sequential experiments.

4.4.5 Synthesis of DNA libraries for next generation sequencing

Following RNA extraction, and in preparation for RNA-seq, we directly proceeded to use the NEBNext Multiplex Small RNA Sample Prep set for Illumina (NEB E7330) to prepare the DNA libraries without fragmentation. RNA Fragmentation was not performed to guarantee that each 5' read and 3' read from RNA-seq could be reliably assumed as the true starts and ends respectively of the corresponding transcripts. The protocol provided for preparation of DNA libraries by the supplier (NEB) was followed with a few adaptions. Briefly, a 1:2 dilution of the 3' SR adaptor for Illumina was ligated overnight (18 h at 16°C) using between 0.5 μg to 1 μg of non-fragmented RNA as the starting material. Next, the SR RT Primer for Illumina was annealed to the 3'adaptor

ligated RNA samples and then the 5' SR adaptor for Illumina was ligated (1 h at 25°C). Subsequently, a reverse transcription reaction was performed (60 min at 50°C) to obtain cDNA, which was immediately enriched via a standard PCR amplification as recommended by the supplier with a modified extension time of 1 min per cycle instead of 15 s for a total of 15 cycles. The resulting DNA was purified using the AMPure Bead XP system and a magnetic rack in at least two wash cycles with freshly-prepared 80% ethanol.

4.4.6 Illumina sequencing of DNA libraries

Once we obtained the DNA libraries above, DNA samples were submitted to the GSAF (UT Austin) for sequencing. First, the samples were analyzed for their size distribution using a bioanalyzer (Agilent). To enrich for the transcripts of interest (INTERFACE plasmid transcripts), in the case of the sRNA experiment, and enhance mapping depth of every single probe, the GSAF facilities performed a Pippin purification preferentially selecting for transcript sizes between 120 (exact length of tnaC sequence) and 310 nt (observed maximum size). Finally, DNA libraries were prepared for RNA-seq using standard Illumina kits and were run using a NextSeq equipment in a 75X2 paired-end scheme.

4.4.7 Computational processing pipeline of sequencing results

The computational pipeline used to process the RNA-seq results includes the following steps: (1) performing a quality check on base sequencing quality using fastqc, a program offering analysis on attainment of passing quality scores (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), (2) using Cutadapt (http://cutadapt.readthedocs.io/en/stable/guide.html) to trim adaptor sequences despite low adaptor contamination (<0.5%), (3) submitting the sequences for the INTERFACE

plasmid (pBAD/plTetO-probe-RSE-RBS-ES-RER) as the "reference genome", (4) mapping the RNA-seq reads using BWA MEM for paired-end sequences, (4) converting the resulting sam-type file to a bam-type and subsequently to a more manageable bedtools file, (5) using awk to develop a script to filter for the R1 reads that contained at least 6 nucleotides of the probe sequence, and (6) obtaining R2 reads corresponding to the R1 reads from the previous step using their unique identifier and the Linux command "join." In the process of filtering, we discarded concatenated reads (<2% of total filtered reads) due to over-ligation of 5' adaptor from the library preparation step.

4.4.8 Calculation of relative accessibility

Finally, using awk, we developed a code to generate a file that contain a summary of the number of reads per probe ending at different positions within the INTERFACE sequence provided in the alignment. The transcript length with respect to each target region was calculated as the number of nucleotides between the observed transcription start site (TSS) for each promoter (consistent with the TSS reported in the literature: pBAD(Smith and Schleif 1978) and plTetO (Lutz and Bujard 1997)) and the transcription end site, both obtained from the RNA-seq results processed following the procedure described above. To calculate relative accessibility, MatLab was used to calculate the weighted averages of the read length per probe. Finally, a baseline (85 nt) was estimated from an thorough analysis of every single probe characterized in this study and subtracted from each transcript length. In the case of the experiments using the gI intron, the relative accessibility per probe was estimated as the ratio of the weighted average of the transcript length in the presence of the target RNA (adding both inducers: aTc and ara) to the weighted average of the transcript length in the absence of the target RNA (no induction). In contrast, for the sRNA experiment, the relative accessibility was calculated utilizing

only the weighted average of the transcript length in the presence of the INTERFACE transcript because the target RNA is natively present, hence, we are unable to manipulate its expression. Next, this weighted average of the transcript length was linearly normalized to fall between 0 and 1 for each molecule to reduce transcript abundance effects.

4.4.9 Estimating Hfq-dependency class from accessibility changes between parent and Hfq-deficient strains

To estimate the level of Hfq-dependency of sRNAs using INTERFACE, the differences in relative accessibility between parent and Hfq-deficient strains were calculated for every region targeted in this study. The average of two noteworthy characteristics was calculated per sRNA: (1) the fraction of regions which showed significant differences (p-value < 0.05) between strains as well as (2) the absolute value of the maximum difference. Hfq-dependency was estimated as "strong" for sRNAs which exhibited above average behavior in both selected fields, "mid" for sRNAs which exhibited above average behavior in one of two selected fields, and "weak" for sRNAs lacking above-average behavior in both categories.

4.4.10 Sequence motif discovery and search

Using "The MEME Suite", specifically the motif discovery tool GAM2 we analyzed the sequences for all regions harboring a previously reported sRNA-mRNA binding site. GAM2 was independently applied utilizing all preset parameters (except for the number or iterations, n, that was set to 12,800) to both, the pool of regions harboring known sRNA-mRNA binding sites and the pool of regions not harboring any known sRNA-mRNA binding sites. By comparison, motifs were considered significant when

found in at lest 90% of the sequences fed into the algorithm and the motif was not listed in the databases available in GAM2scan of The MEME suite.

To search for a known motif (in this case the motif discovered in the pool of regions harboring known sRNA-mRNA binding sites) we used GAM2scan from the The MEME suite. The algorithm was set to find at least a number of motifs equal to the number of sequences fed but only those motifs with as score greater or equal than 5 were considered.

Chapter Five

Conclusions and perspectives

In this dissertation, I recount a set of novel molecular tools that provides specific understanding, assists in the selection, and enables high throughput characterization of RNA accessible interfaces. These tools associate seamlessly to potentiate each other's capabilities as demonstrated in the works described in Chapter 3 and 4. This toolkit offers a comprehensive picture of the ability of RNA to interact with other molecules and thus empowers its exploitation as a regulatory entity. Collectively, this research has, with a cumulus of evidence, supported the centrality of RNA structural accessibility in RNA structure-function research. Importantly, my work will be of value to enable the scientific community to continue shedding light onto the roles of RNA in various contexts.

In the work described in Chapter 2, an *in vivo* oligonucleotide hybridization system was engineered and applied to the characterization of *in vivo* RNA structural accessibility. To achieve this goal, we exploited post-transcriptional regulation in a scheme that involved fluorescence as a measurable outcome. This study represents the proof of concept of what we termed the iRS^3. Our results positioned this system as a tool that could be applied to study RNA structures in vivo in a variety of contexts. A highlight of this work is the contrast between this approach and chemical probing, specifically the capacity to capture dynamic regions that hallmark potential regulatory regions. This key ability led me to devise an application of potentially high impact and broad scope: characterization of functional regions in regulatory RNAs.

In the third chapter, I present the development of a biophysical approach for the prediction of hybridization efficacy (i.e. structural accessibility) in RNAs. In summary, I showed improved prediction capabilities of a thermodynamic model upon optimization

with large sets of experimental data collected using the iRS[3], relative to un-optimized approaches. Demonstration of this performance includes comparable prediction capabilities to benchmark IntaRNA and enhanced linear fits for complex large RNAs such as the group II intron. A total of 130 regions within 7 different RNAs were characterized and an important pattern was observed: the iRS[3] possesses the capacity to identify functional regions i.e. binding sites for other molecules. This realization provided further evidence to the observed unique ability of the iRS[3] to identify dynamic regions as seen in the previous chapter. This computational approach will be instrumental in manipulating and engineering synthetic RNA schemes for gene expression control. Importantly, the inTherAcc model would be used in selecting accessible regions in a high throughput characterization of regulatory RNAs.

Finally in the last chapter, I introduced an innovative system called INTERFACE, for the high throughput characterization of RNA functional structure. INTERFACE exploits the *in vivo* oligonucleotide hybridization scheme described in chapter 2 coupled to transcriptional elongation control. Through RNA-seq we showed that the full accessibility landscape of an RNA molecule of any size can be readily characterized in a single experiment. More importantly I showcased the power of INTERFACE utilizing a version of the biophysical model in chapter 3 coupled to a machine-learning algorithm to select for accessible regions in the small RNA regulatory network. The regions selected, approximately 1000, were characterized using INTERFACE. The results obtained strongly suggest that highly and, to a lesser extent, lowly accessible regions are likely to be involved in interactions with other molecules. These findings underscore the prospects of using INTERFACE in the transcriptome-wide identification of potentially functional RNA regions. Another striking realization was the first strong evidence for the global role of Hfq as a structural relaxer of small RNAs. We foresee that INTERFACE will be used

to study dynamic behavior of regulatory RNAs and for the multiplex characterization of the molecular implications of environmental cues that trigger regulatory responses. We anticipate this information will be instrumental in the mapping of complex molecular networks, particularly those governed by molecular regulation.

The pioneering work detailed in this dissertation sets the grounds for broad applications that span from the high throughput identification of RNA processing factors to the comprehensive characterization of complex networks that control bacterial virulence in pathogenic bacteria. In fact, I currently actively collaborate with Kevin Vasquez (group member) in a project that involves the former theme while I support Mia Mihailovic in pursuing the goal depicted in the latter. In addition, there are ongoing research efforts using the iRS^3 to understand structural features of important small RNAs such as DsrA and CsrB. In the case of CsrB, Abigail Leistra (group member) has collected promising evidence of the possibility of modular engineering of this global regulator. Altogether, these works will bring about maturity to this early technology.

An important perspective that remains a challenge in the near future is the transfer of these tools to other bacterial organisms and even to higher order organisms such as yeast and mammalian cells. At this point, I do not foresee extreme difficulties in transferring the iRS^3 to other organisms because it is based on universal principles such as Watson-Crick base-pairing and translational regulation. In contrast, INTERFACE could face obstacles when transferred to higher-order organisms given that it is based on the fact that, in bacteria, translation and transcription are coupled. One potential strategy to overcome this issue is to exploit other types of transcriptional elongation control by introducing controllable transcriptional terminators. In the hypothetical scenario of successful transfer of these technologies to pathogenic bacteria and eukaryotes, better understanding of regulatory networks, such as those in charge of regulating virulence or

even tumor growth in the case of mammalian cells, could be attained. At this point in time, considering that *E. coli* is regarded as a "biotechnological factory", the prospects of exerting multiplex fine-tuning of regulatory networks is within reach. In summary, I believe that altogether these technologies offer the possibility to develop important applications in the metabolic engineering and human health fields.

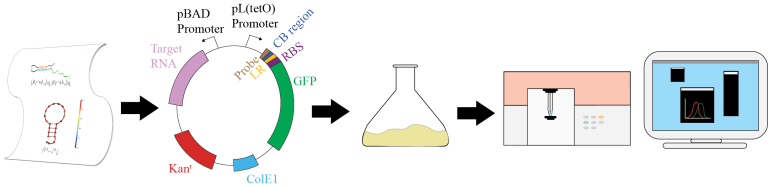
In the RNA folding field, there are important applications that could be realized using this toolkit, such as monitoring of conformational changes during folding and formation of ribonucleoproteins. Specifically, by using oligonucleotide hybridization (*in vitro* or *in vivo*, for one or multiple molecules) at different time-scales and in the presence of various folding factors, a comprehensive snapshot of the conformational changes of the functional structural landscape could be obtained. The sole vision of possessing the ability to monitor folding pathways from a functional structure angle could excite the entire RNA-folding community into studying fundamental RNA folding dynamics using our approaches in conjunction with other methods.

I expect that my PhD research becomes a cornerstone in work performed at the Contreras group aimed at discovering, understanding, characterizing and predicting the roles of regulatory RNAs in bacteria. I trust that the toolkit of molecular tools hereby presented and developed in collaboration with several of my lab mates, will be instrumental to the broader scientific community in furthering the command over RNA molecular functions.

Appendices

APPENDIX A: SUPPLEMENTARY DATA FOR CHAPTER TWO

Supplementary figures for Chapter two



1. Design Probes 2. Clone Plasmids 3. Design Experiment 4. Measure Fluorescence and Analyze Data

1. Design Probes
   - A. Thermodynamic properties
     - a. GC content (recommended between 40-60%)
     - b. Melting temperature (probe to target RNA: 38-50°C)
     - c. Similar Gibbs free energies (all probes should bind favorably to target RNA)
     - d. Structural predictions on thermodynamic/folding properties of probe and probe-GFP reporter (iRS3) construct (Nupack (32)). Probes should have only small interactions with the GFP stem.
   - B. Decide on probe length
     - a. Longer probes allow for greater coverage of the target RNA.
     - b. Shorter probes allow for great resolution of the exposed and protected areas.

2. Clone Plasmids
   - A. Target RNA can be cloned using traditional restriction enzyme cloning (XbaI/SalI digest, Supplementary Sequence File).
   - B. Probes can be cloned into iRS3 using traditional cloning (forming a Probe-hair pin-GFP insert). Alternatively they can be cloned using Gibson Assembly.

3. Design Experiment
   - A. Experiment to look at structural effects of RNA mutants.
     - a. Can be run with only a few probes to look at key areas of interest.
     - b. Run with biological quadruplicates of both induced and uninduced samples.
   - B. Examine relative accessibility across target RNA.
     - a. Choose probes that target a variety of areas of interest.
     - b. Measurements become a standard of relative accessibility along the molecule.

4. Measure Fluorescence and Analyze Data
   - A. Data gathered in flow cytometer should be analyzed for outliers, standard deviation, and average fluorescence of each induced and uninduced construct.
   - B. Fluorescence data can be analyzed using fluorescence values (uninduced and induced), shift (difference in fluorescence, induced minus uninduced) or ratio (induced to uninduced).

Figure A.1. General Methodology for iRS^3 experimental design.

The flow chart presents the general steps for designing an iRS^3 experiment. The outline provides more detailed considerations for designing an experiment.

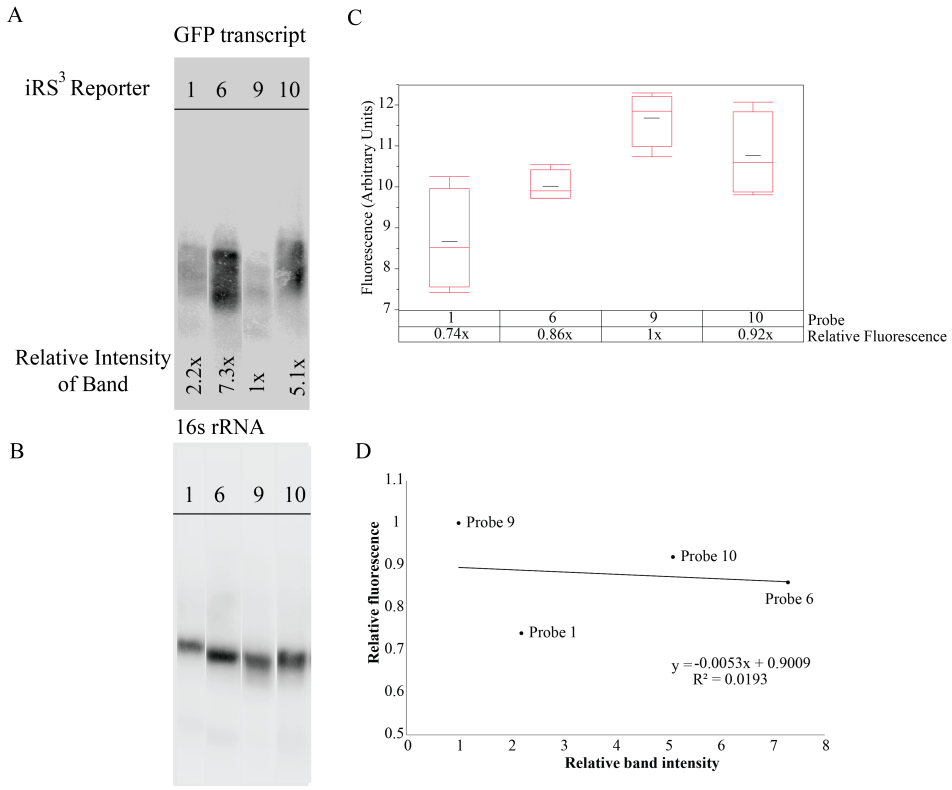


Figure A.2. Northern Analysis of iRS[3] reporters show that reporter transcript levels do not correlate with fluorescence output.

Total RNA was harvested from cells expressing Probe 1, 6, 9, and 10 reporter transcripts 5 hours after induction of iRS[3] transcript only, ran down an agarose-formaldehyde gel and blotted with specific oligonucleotide probes. (A) Northern blot using primer J (**Table A.2**) which is specific to the middle of GFP. (B) Northern blot using primer I (**Table A.2**) which is specific to 16s rRNA as an endogenous loading control. All probe reporters were run down the same gel, but the lanes were reordered for simplicity. (C) Fluorescence of strains containing Probe reporters 1, 6, 9, and 10 at 5 hours after induction of iRS[3] transcription only. The elements of the box plots are as defined in the Figure 2.3 caption. (D) The values of relative fluorescence were plotted against values of relative band intensity from northern blots to test for any significant correlation between the fluorescence produced by each construct and its transcript abundance. The slope and R^2 value tend to zero as a good indicator of the lack of correlation.

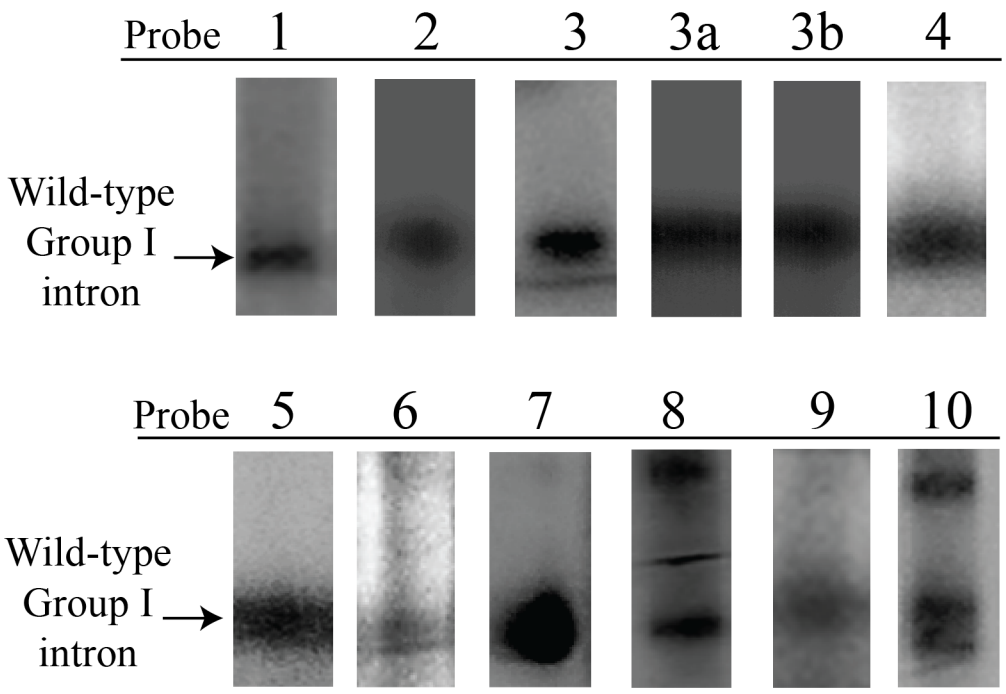


Figure A.3. All probes bind in vitro to the denatured gI intron in the context of total RNA extract.

Total RNA extract containing overexpressed gI intron was heated to 95°C (Methods: *In vitro* binding assays). After heating, the transcripts were removed from 95°C bath and immediately hybridized to P^{32} labeled probe. The probes were allowed to hybridize with the gI intron for 30 minutes at 37°C. The resulting hybridized mixture was then loaded into a 6% native PAGE gel. After running, the gel was dried and exposed to a phosphor screen for at least 4 hours. The black arrow points to the wild type gI intron band. Each gel was run along with the PhiX174 DNA/HinfI ladder to confirm size, *in vitro* transcribed gI intron (three concentrations between 50-500 ng), and a no-probe control (controls not shown).

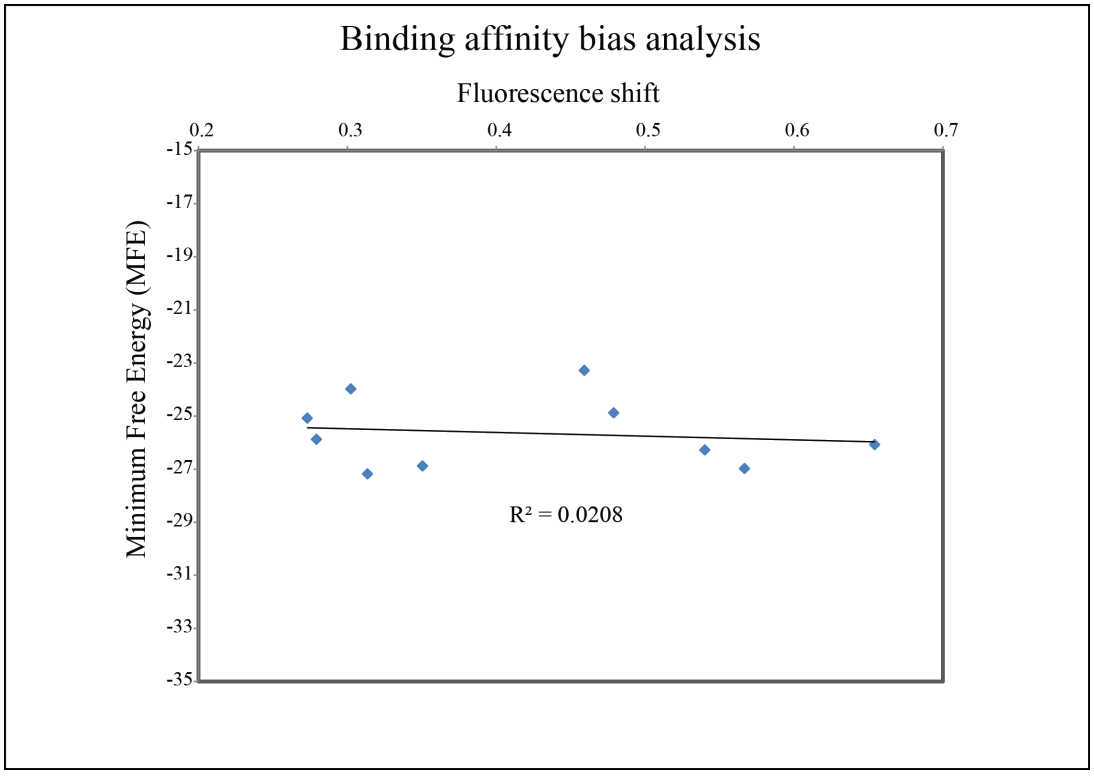


Figure A.4. Test for binding affinity bias.

The correlation between iRS[3] binding affinity to the gI intron and generated fluorescence was assessed thermodynamically and plotted. The plot contains data for various probe lengths ranging from 15-18 nucleotides. Using NUPACK software, we determined the minimum free energy (MFE) for the bound complex formed by the free-standing probes and targeted region (plotted on the y-axis). On the x-axis, we plotted the fluorescence shifts (difference between the averaged normalized medians of fluorescence of the non-induced and induced samples).

APPENDIX B: SUPPLEMENTARY DATA FOR CHAPTER THREE

Supplementary figures Chapter Three

Supplementary Figure 1

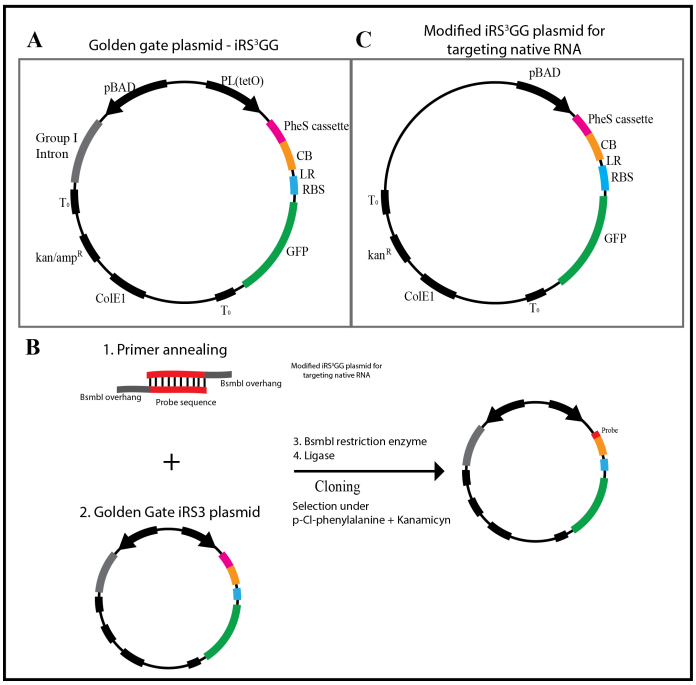


Figure B.1. Experimental plasmids and Golden Gate cloning procedure for synthesizing asRNAs of interest.

(A) iRS[3] experimental plasmid diagram, for use in experimentally targeting over-expressed molecules. **(B)** Modified iRS[3] experimental plasmid for use when evaluating hybridization efficacy of asRNAs targeting native molecules. **(C)** Synthesis of asRNAs (probe) by the Golden Gate cloning procedure. 1. Primers with BsmbI overhangs were ordered (IDT) and annealed by heating up to 95 C and maintaining the temperature at 52 C for 10 min. 2. The annealed primers along with the iRS3GG plasmid, the BsmbI restriction enzyme (Thermo Scientific) and T4 DNA ligase (NEB Labs) are incubated at 37 C for 45 min. 3. Two uL of the reaction are transformed into E. coli electro-competent cells and plated onto negative selection LB-agar media with Kanamicyn and p-Cl-phenylalanine to select for the plasmids without the PheS cassette. The diagram of the iRS3GG plasmid is presented in the bottom of the figure. See Methods for details.

Supplementary Figure 2

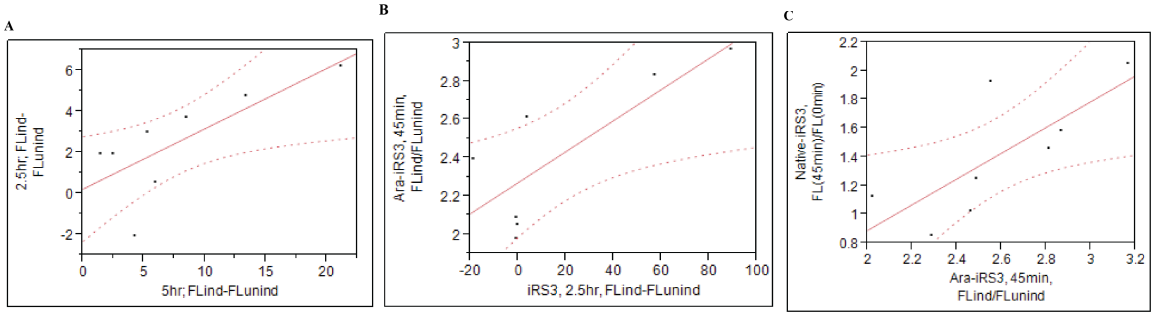


Figure B.2. Engineering the iRS^3 for characterizing native transcripts.

A. Evidence supporting the use of shorter time intervals in the collection of fluorescence data. Previous work in (43) used shifts in fluorescence (as opposed to ratios) collected 5 hours after induction. The linear fit shown supports shortening the time after induction to 2.5 h (p-value < 0.04). These data were collected by overexpressing the Arg-tRNA^{CCU}. **B.** Use of the pBAD promoter to control the iRS^3 expression is supported by its strong correlation to the original system. In order to target native transcripts, the iRS^3 must be under the tighter control of the pBAD promoter (since target cannot be turned on and off). The significant correlation (p-value < 0.04) supports the use of this engineered system, supports the use of ratios instead of differences, and allows for shortening the time of collection to 45 min after induction. These data were collected by overexpressing the Glu-tRNA^{UUC}, the tRNA ultimately used in this work. **C.** The Native-iRS^3 senses native transcripts. In this case, no target (Glu-tRNA^{UUC}) was overexpressed (y-axis) and signals were compared to the trends observed when the target was overexpressed (x-axis). In both cases the iRS^3 is expressed from the pBAD promoter-controlled region in the plasmid depicted in Figure B.1C. The significant correlation (p-value < 0.04) supports the use of the plasmid in Figure B.1C for characterizing native tRNAs such as the Glu-tRNA^{UUC}. Confidence curves shown in plots represent 95% confidence level.

Supplementary Figure 3

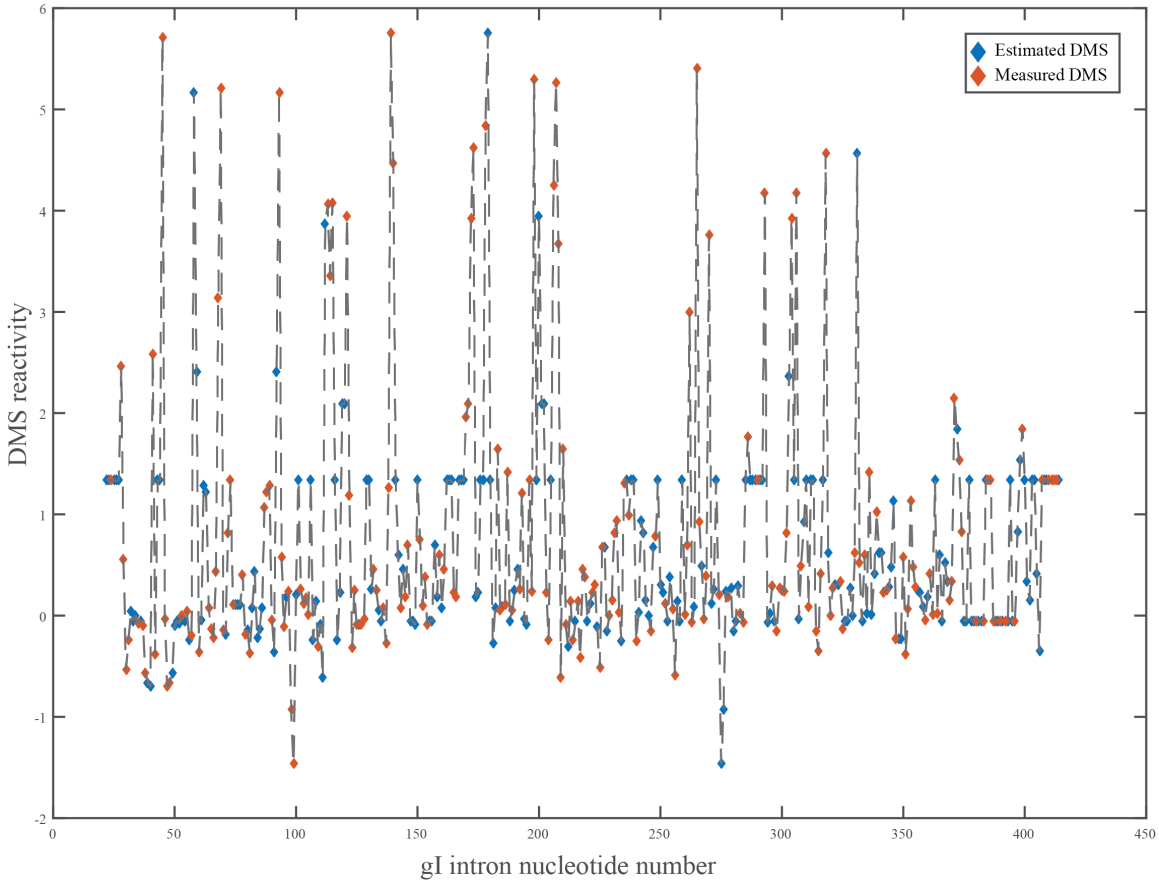


Figure B.3. DMS reactivity data for group I intron.

DMS reactivity data per nucleotide position for the *Tetrahymena* gI intron. DMS reactivity values for As and Cs were calculated by subtracting the no DMS control from the average of reactivity of two independent DMS treated samples (data partially published in (42), see Methods section for more details). Values of DMS reactivity for Gs and Us were estimated by their pairing counterparts when paired and, when unpaired, assigned an average of all accessibilities considered exposed across the molecule based on a calculated threshold (See Methods Section for details).

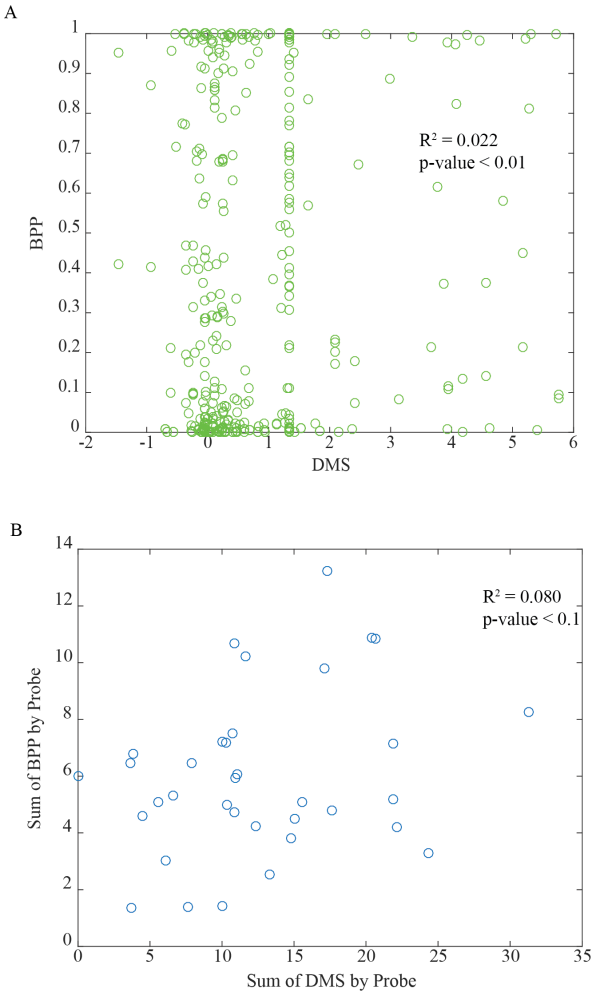


Figure B.4. Local versus regional base pairing probability-DMS correlations show no observable correlation for the group I intron at the local level

(A), but an obvious one at the regional level **(B)**. The notion of structural predictions being representative of experimental structure at the regional level, but not the local level, further supports the notion of regional characteristics as important influencers of target region behavior.

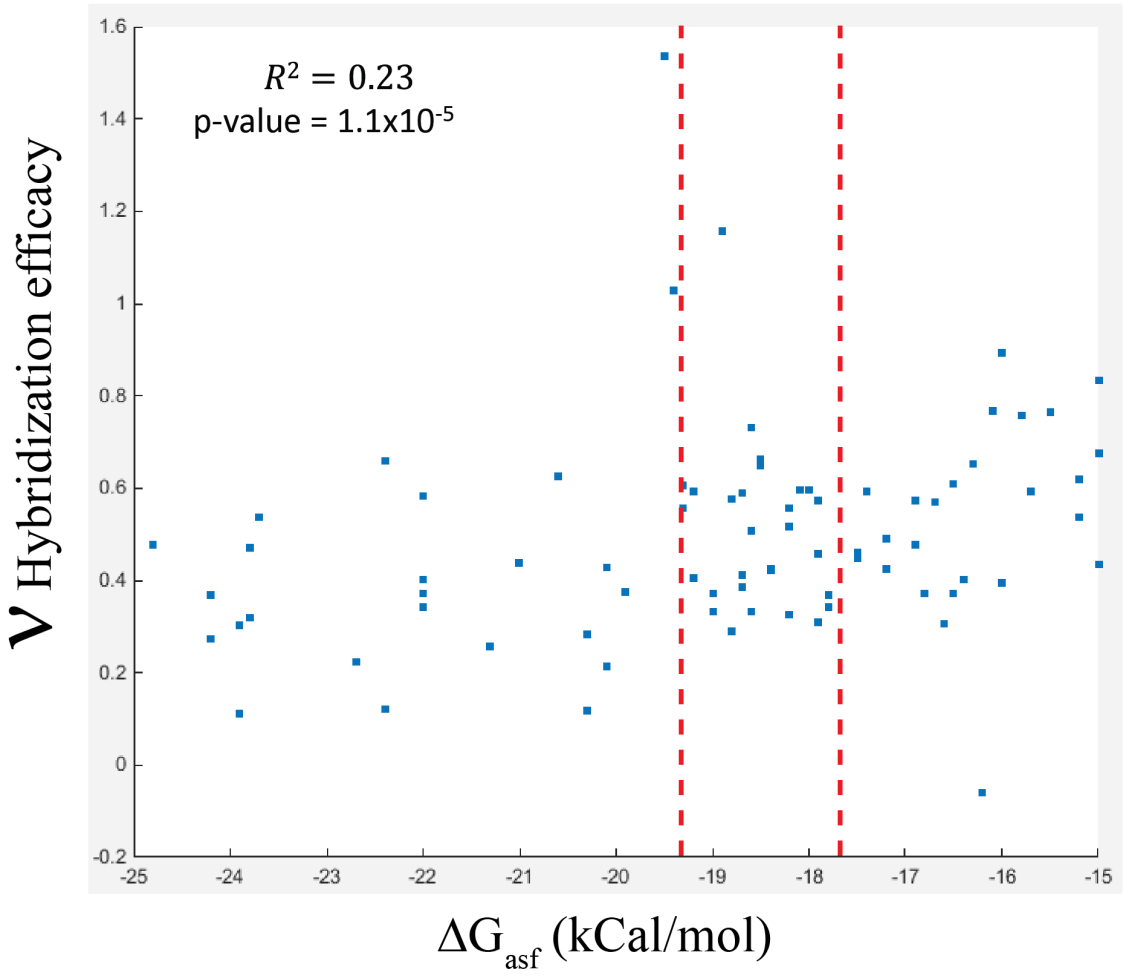


Figure B.5. Undesirable correlation between folding energy of asRNA and hybridization efficacy

Linear regression shows an undesirable correlation between energetics of probe unfolding and hybridization efficacy, due to tool idiosyncrasies that artificially inflate the importance of asRNA. This discovery led to the constraint on this predictor, resulting in an interval (red) considering only data in which the influence of asRNA structure on the measured response becomes insignificant levels

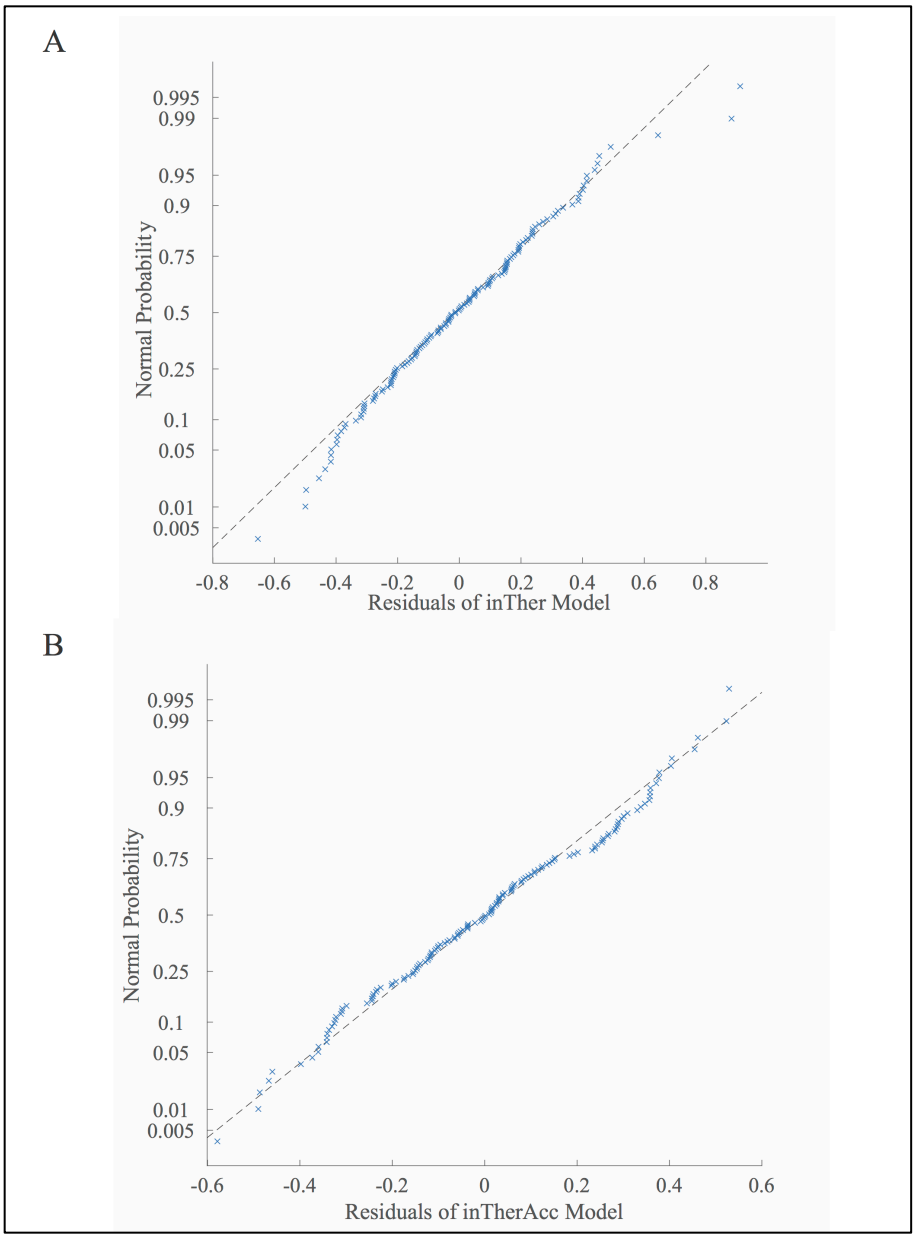


Figure B.6. Linear regression residuals for inTher (A) and inTherAcc (B).

The normal probability plot of residuals for linear regressions of the **(A)** inTher model (Eq. 10) and **(B)** inTherAcc model (Eq. 11) were calculated. Both residuals show characteristics of a normal population supporting the validity of the models derived. See Methods for details.

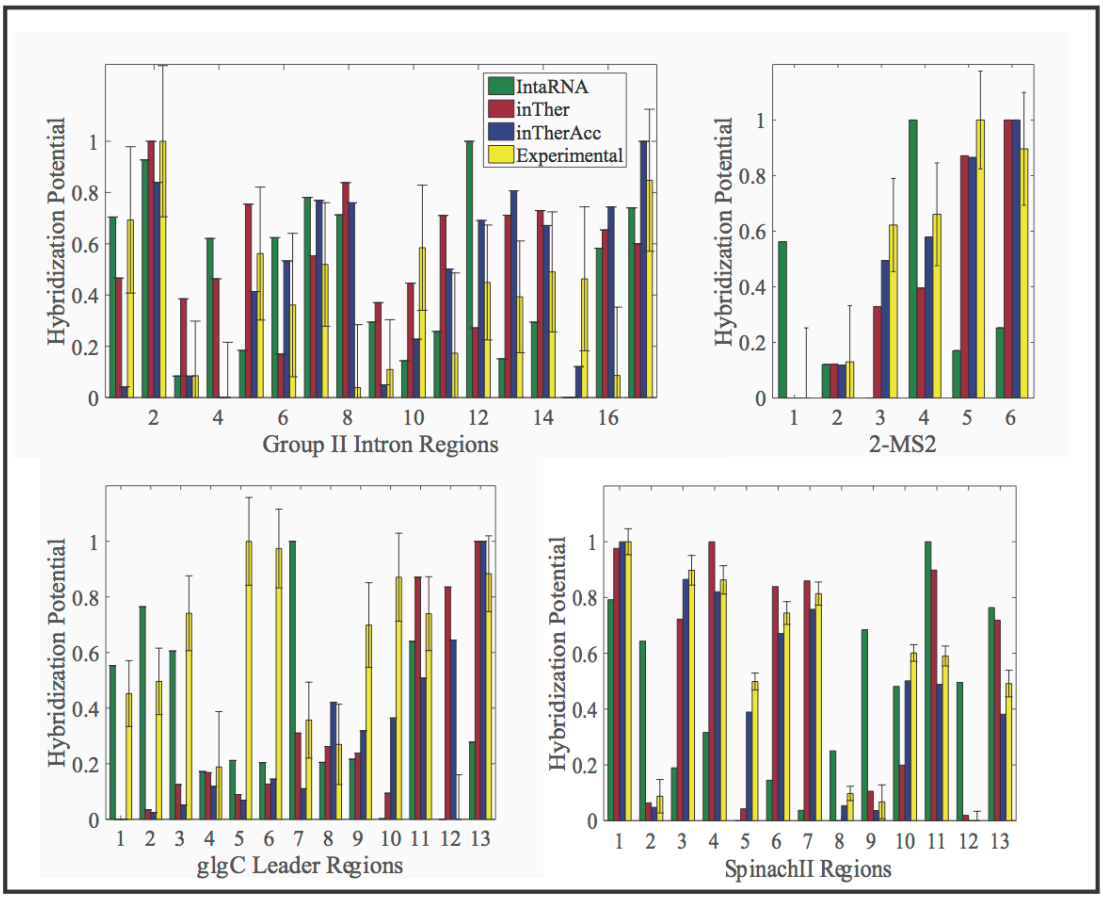


Figure B.7. Detailed results on prediction performance benchmark study

Predictions of IntaRNA, inTher, and inTherAcc models are compared to respective experimental scaled hybridization efficacies in group II intron, 2-MS2, glgC leader, and SpinachII. All predicted and experimental values were linearly scaled to fall between 0 and 1. Error bars indicate standard error of the mean.

Supplementary tables Chapter Three

Table B.1. List of asRNAs used in this study

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
gI intron-1	8	22	29	UUCCCUCC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-2	16	22	37	GAUAACUUUUCCCUCC	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-3	12	22	33	ACUUUUCCCUCC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-4	22	22	43	AUGCCUGAUAACUUUUCCCUCC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-5	27	22	48	GGUGCAUGCCUGAUAACUUUUCCCUCC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-6	10	64	73	AAACCAAUAG	gI intron	Golden Gate	AATTCAAACCAATAGT	TGGTACTATTGGTTTG	
gI intron-7	16	67	82	CCGAUGCAAUCUAUUG	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-8	10	95	104	UUGACGGUCU	gI intron	Site Directed	NA	NA	Published in (20)

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
						Mutagenesis			
gI intron-9	15	98	112	CCCGCAAUUUGACGG	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-10	14	113	126	CUGUUGACCCCUUU	gI intron	GenScript Inc.	NA	NA	
gI intron-11	10	113	122	AAAGGGGUCA	gI intron	Golden Gate	AATTCAAAGGGGTCAT	TGGTATGACCCCTTTG	
gI intron-12	14	127	140	UUGGUACUGAACGG	gI intron	GenScript Inc.	NA	NA	
gI intron-13	10	134	143	GUACCAAGUC	gI intron	Golden Gate	AATTCGTACCAAGTCT	TGGTAGACTTGGTACG	
gI intron-14	15	141	155	AGUUUCCCCUGAGAC	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-15	15	156	170	GCAAGGCCAUCUCAA	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-16	10	169	178	UACCCUUUGC	gI intron	Golden Gate	AATTCTACCCTTTGCT	TGGTAGCAAAGGGTAG	
gI intron-17	9	171	179	AUACCCUUU	gI intron	GenScript Inc.	NA	NA	
gI intron-18	15	179	194	CGUCAGCUUAUUACC	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-19	10	180	189	GCUUAUUACC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-20	20	195	214	UGCGUGGUUAGGACCAUGUC	gI intron	GenScript Inc.	NA	NA	
gI intron-	10	204	213	GCGUGGUUAG	gI	Golden	AATTCGCGTGGTT	TGGTACTAACCAC	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
21					intron	Gate	AGT	GCG	
gI intron-22	19	215	233	UGUUGACUUAGGACUUGGC	gI intron	GenScript Inc.	NA	NA	
gI intron-23	18	234	251	CCAUAUCAACAGAAGAUC	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-24	10	260	269	UUAGUCUGUG	gI intron	Golden Gate	AATTCTTAGTCTGTGT	TGGTACACAGACTAAG	
gI intron-25	10	279	287	CAUCUUCCCC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-26	17	280	296	GAAGAAUACAUCUUCCC	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-27	18	295	312	CGACUAUAUCUUAUGAGA	gI intron	GenScript Inc.	NA	NA	
gI intron-28	15	315	329	CCCAUUAAGGAGAGG	gI intron	GenScript Inc.	NA	NA	
gI intron-29	10	361	370	UUCCCAGCGG	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-30	15	361	375	AUUAGUUCCCAGCGG	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-31	10	366	375	AUUAGUUCCC	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-32	27	388	414	CGAGUACUCCAAAACUAAUCAAUA	gI intron	Site Directed	NA	NA	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
				UAC		Mutagenesis			
gI intron-33	22	393	414	CGAGUACUCCAAAACUAAUCAA	gI intron	Site Directed Mutagenesis	NA	NA	
gI intron-34	16	399	414	CGAGUACUCCAAAACU	gI intron	GenScript Inc.	NA	NA	Published in (20)
gI intron-35	10	405	414	CGAGUACUCC	gI intron	Site Directed Mutagenesis	NA	NA	
2MS2-1	14	36	49	UGGAGUCGAC	MS2	Golden Gate	AATTCTGGAGTCGACCTGCT	TGGTAGCAGGTCGACTCCAG	
2MS2-2	16	36	51	UCUGGAGUCGACCUGC	MS2	Golden Gate	AATTCTCTGGAGTCGACCTGCT	TGGTAGCAGGTCGACTCCAGAG	
2MS2-3	13	44	56	UGUUUUCUGGAGU	MS2	Golden Gate	AATTCTGTTTTCTGGAGTT	TGGTAACTCCAGAAAACAG	
2MS2-4	11	45	55	GGGAAUACUGCAGACA	MS2	Golden Gate	AATTCGGGAATACTGCAGACAT	TGGTATGTCTGCAGTATTCCCG	
2MS2-5	17	58	74	AGACAUGGGUGAUCCUC	MS2	Golden Gate	AATTCAGACATGGGTGATCCTCT	TGGTAGAGGATCACCCATGTCTG	
2MS2-6	16	59	74	AGACAUGGGUGAUCCU	MS2	Golden Gate	AATTCAGACATGGGTGATCCTT	TGGTAAGGATCACCCATGTCTG	
glgC-1	10	9	18	GUGCAGGUCC	glgC	Golden Gate	AATTCGTGCAGGTCCT	TGGTAGGACCTGCACG	
glgC-2	14	15	28	CACAAUCCGUGUGC	glgC	Golden Gate	AATTCCACAATCCGTGTGCT	TGGTAGCACACGGATTGTGG	
glgC-3	11	27	37	UGGAACACACA	glgC	Golden	AATTCTGGAACA	TGGTATGTGTGTT	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
						Gate	CACAT	CCAG	
glgC-4	11	40	50	UUUUUAUCAUC	glgC	Golden Gate	AATTCTTTTTATCATCT	TGGTAGATGATAAAAAG	
glgC-5	11	48	58	UAACUCCUUUU	glgC	Golden Gate	AATTCTAACTCCTTTTT	TGGTAAAAAGGAGTTAG	
glgC-6	11	53	63	AUGACUAACUC	glgC	Golden Gate	AATTCATGACTAACTCT	TGGTAGAGTTAGTCATG	
glgC-7	14	59	72	AAACUAACCAUGAC	glgC	Golden Gate	AATTCAAACTAACCATGACT	TGGTAGTCATGGTTAGTTTG	
glgC-8	10	72	81	UUCUUCUCUA	glgC	Golden Gate	AATTCTTCTTCTCTAT	TGGTATAGAGAAGAAG	
glgC-9	17	88	104	GGCGCGCCAACAUUAAG	glgC	Golden Gate	AATTCGGCGCGCCAACATTAAGT	TGGTACTTAATGTTGGCGCGCCG	
glgC-10	14	102	115	CAAUGGCAGCUGGC	glgC	Golden Gate	AATTCCAATGGCAGCTGGCT	TGGTAGCCAGCTGCCATTGG	
glgC-11	13	119	131	UCAGGGCAACAGA	glgC	Golden Gate	AATTCTCAGGGCAACAGAT	TGGTATCTGTTGCCCTGAG	
glgC-12	17	134	150	CCACGUCCUCCCGCCAG	glgC	Golden Gate	AATTCCCACGTCCTCCCGCCAGT	TGGTACTGGCGGGAGGACGTGGG	
glgC-13	12	150	161	UCAGGCGGGUAC	glgC	Golden Gate	AATTCTCAGGCGGGTACT	TGGTAGTACCCGCCTGAG	
SpinachII-1	11	1	11	GCUAUCCGGGC	Spinach II	Golden Gate	AATTCGCTATCCGGGCT	TGGTAGCCCGGATAGCG	
SpinachII-2	12	13	24	CUCUACCGACUG	Spinach II	Golden Gate	AATTCCTCTACCGACTGT	TGGTACAGTCGGTAGAGG	
SpinachII-3	9	43	51	UUUCAUUCA	Spinach II	Golden Gate	AATTCTTTCATTCAT	TGGTATGAATGAAAG	
SpinachII-4	16	61	76	AGCCUACUGGACCCGU	Spinach II	Golden Gate	AATTCAGCCTACTGGACCCGTT	TGGTAACGGGTCCAGTAGGCTG	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
SpinachII-5	9	65	73	CUACUGGAC	SpinachII	Golden Gate	AATTCCTACTGGACT	TGGTAGTCCAGTAGG	
SpinachII-6	17	65	81	GAAGCAGCCUACUGGAC	SpinachII	Golden Gate	AATTCGAAGCAGCCTACTGGACT	TGGTAGTCCAGTAGGCTGCTTCG	
SpinachII-7	17	76	92	AGUAGGCUGCCGAAGCA	SpinachII	Golden Gate	AATTCAGTAGGCTGCCGAAGCAT	TGGTATGCTTCGGCAGCCTACTG	
SpinachII-8	15	89	103	UCUACUCAACAAGUA	SpinachII	Golden Gate	AATTCTCTACTCAACAAGTAT	TGGTATACTTGTTGAGTAGAG	
SpinachII-9	16	101	116	ACGGAGCUCACACUCU	SpinachII	Golden Gate	AATTCACGGAGCTCACACTCTT	TGGTAAGAGTGTGAGCTCCGTG	
SpinachII-10	15	122	136	GCGGAUAGAUGUAAC	SpinachII	Golden Gate	AATTCGCGGATAGATGTAACT	TGGTAGTTACATCTATCCGCG	
SpinachII-11	14	134	147	ACCCUGGACCCGCG	SpinachII	Golden Gate	AATTCACCCTGGACCCGCGT	TGGTACGCGGGTCCAGGGTG	
SpinachII-12	10	142	151	UUGAACCCUG	SpinachII	Golden Gate	AATTCTTGAACCCTGT	TGGTACAGGGTTCAAG	
SpinachII-13	9	161	169	UGGCGCCCG	SpinachII	Golden Gate	AATTCTGGCGCCCGT	TGGTACGGGCGCCAG	
gII intron-1	15	2	16	CGUUAUGGAUGUGUU	gII intron	Golden Gate	AATTCCGTTATGGATGTGTTT	TGGTAAACACATCCATAACGG	
gII intron-2	11	16	26	UCUGGGCGCAC	gII intron	Golden Gate	AATTCTCTGGGCGCACT	TGGTAGTGCGCCCAGAG	
gII intron-3	12	89	100	CUUUUCGGUUAG	gII intron	Golden Gate	AATTCCTTTTCGGTTAGT	TGGTACTAACCGAAAAGG	
gII intron-4	14	307	320	UGUGGUGAUAACAG	gII intron	Golden Gate	AATTCTGTGGTGATAACAGT	TGGTACTGTTATCACCACAG	
gII intron-5	9	307	315	UGAUAACAG	gII intron	Golden Gate	AATTCTGATAACAGT	TGGTACTGTTATCAG	
gII	10	389	398	GUGUUAAGUC	gII	Golden	AATTCGTGTTAAG	TGGTAGACTTAAC	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
intron-6					intron	Gate	TCT	ACG	
gII intron-7	14	430	443	UUCCUCCUUUCUAU	gII intron	Golden Gate	AATTCTTCCTCCTTTCTATT	TGGTAATAGAAAGGAGGAAG	
gII intron-8	14	478	491	GUACUCCGUACCCU	gII intron	Golden Gate	AATTCGTACTCCGTACCCTT	TGGTAAGGGTACGGAGTACG	
gII intron-9	9	596	604	CCAUUGUUG	gII intron	Golden Gate	AATTCCCATTGTTGT	TGGTACAACAATGGG	
gII intron-10	10	620	629	UUUUUACUGA	gII intron	Golden Gate	AATTCTTTTTACTGAT	TGGTATCAGTAAAAAG	
gII intron-11	12	634	645	UAUAUUUUCUUG	gII intron	Golden Gate	AATTCTATATTTTCTTGT	TGGTACAAGAAAATATAG	
gII intron-12	9	707	715	CCCAACGCG	gII intron	Golden Gate	AATTCCCCAACGCGT	TGGTACGCGTTGGGG	
gII intron-13	11	742	752	ACAAGAGUUUU	gII intron	Golden Gate	AATTCACAAGAGTTTTT	TGGTAAAAACTCTTGTG	
gII intron-14	12	781	792	GUGUUUAUGAAU	gII intron	Golden Gate	AATTCGTGTTTATGAATT	TGGTAATTCATAAACACG	
gII intron-15	11	795	805	UAAAAAUUCAC	gII intron	Golden Gate	AATTCTAAAAATTCACT	TGGTAGTGAATTTTTAG	
gII intron-16	15	807	821	UGUUAUUGUUCGUUC	gII intron	Golden Gate	AATTCTGTTATTGTTCGTTCT	TGGTAGAACGAACAATAACAG	
gII intron-17	11	824	834	GAGUAUACGGC	gII intron	Golden Gate	AATTCGAGTATACGGCT	TGGTAGCCGTATACTCG	
csrB-1	10	4	13	ACUCCCUGUC	csrB	Gibson Assembly	GAATTCACTCCCTGTCTACCATTCACCTCTTGGATTTG	TGGTAGACAGGGAGTGAATTCGGTCAGTGCGT	
csrB-2	13	15	27	CACUUCGUUGUCU	csrB	Gibson Assembly	GAATTCCACTTCGTTGTCTTACCATTCACCTC	GAATTCCACTTCGTTGTCTTACCATTCACCTCTTGGATTT	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
							TTGGATTT		
csrB-3	14	32	45	UGUCAUCAUCCUGA	csrB	Gibson Assembly	CGAATTCTGTCATCATCCTGATACCATTCACCTCTTGGATTT	TCAGGATGATGACAGAATTCGGTCAGTGCGTC	
csrB-4	12	46	57	GUCCUGCAGAAG	csrB	Gibson Assembly	CGAATTCGTCCTGCAGAAGTACCATTCACCTCTTGGATTT	CTTCTGCAGGACGAATTCGGTCAGTGCGTC	
csrB-5	14	57	70	ACCAUCCUGGUGUG	csrB	Gibson Assembly	GAATTCACCATCCTGGTGTGTACCATTCACCTCTTGGATTT	GGTACACACCAGGATGGTGAATTCGGTCAGTGCGTC	
csrB-6	11	73	83	CUUUCCCUGAA	csrB	Gibson Assembly	GAATTCCTTTCCCTGAATACCATTCACCTCTTGGATTTG	TGGTATTCAGGGAAAGGAATTCGGTCAGTGCGT	
csrB-7	12	85	96	UUCAUCCAGAAG	csrB	Gibson Assembly	GAATTCTTCATCCAGAAGTACCATTCACCTCTTGGATTTG	TGGTACTTCTGGATGAAGAATTCGGTCAGTGCGT	
csrB-8	13	99	111	CGUCAUCCUCUUC	csrB	Gibson Assembly	CGAATTCCGTCATCCTCTTCTACCATTCACCTCTTGGATTT	GAAGAGGATGACGGAATTCGGTCAGTGCGTC	
csrB-9	11	109	119	GCGUCCUGCGU	csrB	Gibson Assembly	CGAATTCGCGTCCTGCGTTACCATTCACCTCTTGGATTT	ACGCAGGACGCGAATTCGGTCAGTGCGTC	
csrB-10	11	122	132	GGUGUCCUUUA	csrB	Gibson Assembly	CGAATTCGGTGTCCTTTATACCATTCACCTCTTGGATTT	TAAAGGACACCGAATTCGGTCAGTGCGTC	
csrB-11	12	135	146	UUCUCCAUCCUG	csrB	Gibson	GAATTCTTCTCCATCC	GGTACAGGATGGAGAAGAATTCG	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
						Assembly	TGTACCATTCACCTCTTGGATTT	GTCAGTGCGTC	
csrB-12	11	147	157	ACCGGUUCUCA	csrB	Gibson Assembly	CGAATTCACCGGTTCTCATACCATTCACCTCTTGGATTT	TGAGAACCGGTGAATTCGGTCAGTGCGTC	
csrB-13	11	154	164	CAUCCUGACCG	csrB	Gibson Assembly	CGAATTCCATCCTGACCGTACCATTCACCTCTTGGATTT	CGGTCAGGATGGAATTCGGTCAGTGCGTC	
csrB-14	11	166	176	GACCCACCGAA	csrB	Gibson Assembly	AATTCGACCCACCGAATACCATTCACCTCTTGGATTTG	GGTATTCGGTGGGTCGAATTCGGTCAGTGCGTC	
csrB-15	11	176	186	UGGCCUUCCUG	csrB	Gibson Assembly	CGAATTCTGGCCTTCCTGTACCATTCACCTCTTGGATTT	CAGGAAGGCCAGAATTCGGTCAGTGCGTC	
csrB-16	12	184	195	AAGUGUCCCUGG	csrB	Gibson Assembly	GAATTCAAGTGTCCCTGGTACCATTCACCTCTTGGATTT	GGTACCAGGGACACTTGAATTCGGTCAGTGCGTC	
csrB-17	13	193	205	CUUCAUCCUGAAG	csrB	Gibson Assembly	GAATTCCTTCATCCTGAAGTACCATTCACCTCTTGGATTT	GGTACTTCAGGATGAAGGAATTCGGTCAGTGCGTC	
csrB-18	11	212	222	ACCACCCCGAU	csrB	Gibson Assembly	CGAATTCACCACCCCGATTACCATTCACCTCTTGGATTT	ATCGGGGTGGTGAATTCGGTCAGTGCGTC	
csrB-19	11	229	239	AUUGCUUCCUG	csrB	Gibson Assembly	CGAATTCATTGCTTCCTGTACCATTCACCTCTTGGATTT	CAGGAAGCAATGAATTCGGTCAGTGCGTC	
csrB-20	12	244	255	UCGUUCAUCCUG	csrB	Gibson	CGAATTCTCGTTCATC	CAGGATGAACGAGAATTCGGTCA	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
						Assembly	CTGTACCATTCACCTCTTGGATTT	GTGCGTC	
csrB-21	12	255	266	CUUGCGGCCAAU	csrB	Gibson Assembly	GAATTCCTTGCGGCCAATTACCATTCACCTCTTGGATTT	GGTAATTGGCCGCAAGGAATTCGGTCAGTGCGTC	
csrB-22	10	267	275	UUCCUCUGGC	csrB	Gibson Assembly	GAATTCTTCCTCTGGCTACCATTCACCTCTTGGATTT	GGTAGCCAGAGGAAGAATTCGGTCAGTGCGTC	
csrB-23	11	270	280	AACUUUUCCUC	csrB	Gibson Assembly	CGAATTCAACTTTTCCTCTACCATTCACCTCTTGGATTT	GAGGAAAAGTTGAATTCGGTCAGTGCGTC	
csrB-24	11	281	291	UCAUCCUUGAC	csrB	Gibson Assembly	GAATTCTCATCCTTGACTACCATTCACCTCTTGGATTT	GGTAGTCAAGGATGAGAATTCGGTCAGTGCGTC	
csrB-25	12	294	305	UUGUUGCUCCCUG	csrB	Gibson Assembly	GAATTCTTGTTGCTCCTGTACCATTCACCTCTTGGATTT	GGTACAGGAGCAACAAGAATTCGGTCAGTGCGTC	
csrB-26	12	310	321	AGCAUUCCAGCU	csrB	Gibson Assembly	CGAATTCAGCATTCCAGCTTACCATTCACCTCTTGGATTT	AGCTGGAATGCTGAATTCGGTCAGTGCGTC	
csrB-27	12	324	335	CCGGUUCGUUUC	csrB	Gibson Assembly	CGAATTCCCGGTTCGTTTCTACCATTCACCTCTTGGATTT	GAAACGAACCGGGAATTCGGTCAGTGCGTC	
tRNA-1	9	2	10	CGAAGGGGA	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCCGAAGGGGATACCATTCACCTCTTGGATTTGGG	GAATGGTATCCCCTTCGGAATTCTATGGAGAAACAGTAGAGAGTTGC	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
tRNA-2	9	8	16	UCUAGACGA	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCTCTAGACGATACCATTCACCTCTTGGATTTGGG	GAATGGTATCGTCTAGAGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-3	9	13	21	GGGCCUCUA	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCGGGCCTCTATACCATTCACCTCTTGGATTTGGG	GAATGGTATAGAGGCCCGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-4	10	13	22	UGGGCCUCUA	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCTGGGCCTCTATACCATTCACCTCTTGGATTTGGG	GAATGGTATAGAGGCCCAGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-5	9	14	22	UGGGCCUCU	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCTGGGCCTCTTACCATTCACCTCTTGGATTTGGG	GAATGGTAAGAGGCCCAGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-6	9	20	28	GUGUCCUGG	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCGTGTCCTGGTACCATTCACCTCTTGGATTTGGG	GAATGGTACCAGGACACGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-7	9	23	31	GCGGUGUCC	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCGCGGTGTCCTACCATTCACCTCTTGGATTTGGG	GAATGGTAGGACACCGCGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-8	9	26	34	AGGGCGGUG	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCAGGGCGGTGTACCATTCACCTCTTGGATTTGGG	GAATGGTACACCGCCCTGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-9	9	29	37	UGAAAGGGC	tRNA-	Gibson	ATAGAATTCTGA	GAATGGTAGCCCTTTCAGAA	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
					Glu^{UUC}	Assembly	AAGGGCTACCATTCACCTCTTGGATTTGGG	TTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-10	9	31	39	CGUGAAAGG	tRNA-Glu^{UUC}	GenScript Inc./restriction cloning	NA	NA	
tRNA-11	9	33	41	GCCGUGAAA	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCGCCGTGAAATACCATTCACCTCTTGGATTTGGG	GAATGGTATTTCACGGCGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-12	13	39	51	CCCUGUUACCGCC	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCCCCTGTTACCGCCTACCATTCACCTCTTGGATTTGGG	GAATGGTAGGCGGTAACAGGGGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-13	9	41	49	CUGUUACCG	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCCTGTTACCGTACCATTCACCTCTTGGATTTGGG	GAATGGTACGGTAACAGGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-14	9	49	57	UCGAACCCC	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCTCGAACCCCTACCATTCACCTCTTGGATTTGGG	GAATGGTAGGGGTTCGAGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-15	9	50	58	UUCGAACCC	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCTTCGAACCCTACCATTCACCTCTTGGATTTGGG	GAATGGTAGGGTTCGAAGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-16	9	53	61	GAUUCGAAC	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCGATTCGAACTACCATTCACCTCTTGGATTT	GAATGGTAGTTCGAATCGAATTCTATGGAGAAACAGTAGAGAGTTGC	

asRNA	Length	Target start	Target end	Sequence	Target RNA	Synthesis method	Golden Gate primers		Remarks
							Forward	Reverse	
							GGG		
tRNA-17	9	57	65	AGGGGAUUC	tRNA-Glu^{UUC}	Gibson Assembly	ATAGAATTCAGGGGATTCTACCATTCACCTCTTGGATTTGGG	GAATGGTAGAATCCCCTGAATTCTATGGAGAAACAGTAGAGAGTTGC	
tRNA-18	9	63	71	TCCCCTAGG	tRNA-Glu^{UUC}	GenScript Inc./restriction cloning	NA	NA	

Table B.2. List of target molecules and cloning strategy used in this study.

Target Molecule	Target Molecule Sequence	Cloning Method	Primers		Vector Backbone Primers	
			Forward	Reverse	Forward	Reverse
glgC	TCTGGCAGGGACCTGCACACGGATTGTGTGTGTTCCAGAGATGATAAAAAAGGAGTTAGTCATGGTTAGTTTAGAGAAGAACGATCACTTAATGTTGGCGCGCCAGCTGCCATTGAAATCTGTTGCCCTGATACTGGCGGGAGGACGTGGTACCCGCCTGA	Gibson Assembly	ACTGCCGCCAGGCATCTAGATCAGGCGGGTACCACGTC	ACTGTTTCTCCATAGTCGACTCTGGCAGGGACCTGCAC	GTCGACTATGGAGAAACAGTAGAG	TCTAGATGCCTGGCGGCA
MS2	TAATTGCCTAGAAAACATGAGGATCACCCATGTCTGCAGGTCGACTCCAGAAAACATGAGGATCACCCATGTCTGCAGTATTCCCGGGTTCATT	Gibson Assembly	ACTGCCGCCAGGCATCTAGAAATGAACCCGGGAATACTG	ACTGTTTCTCCATAGTCGACTAATTGCCTAGAAAACATGAGG	"	"
SpinachII	GCCCGGATAGCTCAGTCGGTAGAGCAGCGGCCGAGATGTAACTGAATGAAATGGTGAAGGACGGGTCCAGTAGGCTGCTTCGGCAGCCTACTTGTTGAGTAGAGTGTGAGCTCCGTAACTAGTTACATCTATCCGCGGGTCCAGGGTTCAAGTCCCTGTTCGGGCGCCATCTTTCTTTTT	Gibson Assembly	ACTGCCGCCAGGCATCTAGAAAAAGAAAGATGGCGCCC	ACTGTTTCTCCATAGTCGACGCCCGGATAGCTCAGTCG	"	"

Target Molecule	Target Molecule Sequence	Cloning Method	Primers		Vector Backbone Primers	
			Forward	Reverse	Forward	Reverse
gII intron	GAACACATCCATAACGTGCGCCCAGATAG GGTGTTAAGTCAAGTAGTTTAAGGTACTA CTCTGTAAGATAACACAGAAAACAGCCAA CCTAACCGAAAAGCGAAAGCTGATACGGG AACAGAGCACGGTTGGAAAGCGATGAGTT ACCTAAAGACAATCGGGTACGACTGAGTC GCAATGTTAATCAGATATAAGGTATAAGT TGTGTTTACTGAACGCAAGTTTCTAATTTC GGTTATGTGTCGATAGAGGAAAGTGTCTG AAACCTCTAGTACAAAGAAAGGTAAGTTA TGGTTGTGGACTTATCTGTTATCACCACAT TTGTACAATCTGTAGGAGAACCTATGGGA ACGAAACGAAAGCGATGCCGAGAATCTG AATTTACCAAGACTTAACACTAACTGGGG ATACCCTAAACAAGAATGCCTAATAGAAA GGAGGAAAAAGGCTATAGCACTAGAGCTT GAAAATCTTGCAAGGGTACGGAGTACTCG TAGTAGTCTGAGAAGGGTAACGCCCTTTA CATGGCAAAGGGGTACAGTTATTGTGTAC TAAAATTAAAAATTGATTAGGGAGGAAAA CCTCAAAATGAAACCAACAATGGCAATTT TAGAAAGAATCAGTAAAAATTCACAAGAA AATATAGACGAAGTTTTTACAAGACTTTA TCGTTATCTTTTACGTCCAGATATTTATTA CGTGGCGACGCGTTGGGAAATGGCAATGA TAGCGAAACAACGTAAAACTCTTGTTGTA TGCTTTCATTGTCATCGTCACGTGATTCAT AAACACAAGTGAATTTTTACGAACGAACA ATAACAGAGCCGTATACTCCGAGAGGGGT ACGTACGGTTCCCGAAGAGGGTGGTGCAA ACCAGTCACAGTAATGTGAACAAGGCGGT ACCTCCCTACTTCACCATATCA	Gibson Assembly	ACTGCCGCCAG GCATCTAGATG ATATGGTGAAG TAGGGAG	ACTGTTTCTCCAT AGTCGACGAACA CATCCATAACGT G	"	"

Target Molecule	Target Molecule Sequence	Cloning Method	Primers		Vector Backbone Primers	
			Forward	Reverse	Forward	Reverse
csrB	GTCGACAGGGAGTCAGACAACGAAGTGA ACATCAGGATGATGACACTTCTGCAGGAC ACACCAGGATGGTGTTTCAGGGAAAGGCT TCTGGATGAAGCGAAGAGGATGACGCAG GACGCGTTAAAGGACACCTCCAGGATGGA GAATGAGAACCGGTCAGGATGATTCGGTG GGTCAGGAAGGCCAGGGACACTTCAGGAT GAAGTATCACATCGGGGTGGTGTGAGCAG GAAGCAATAGTTCAGGATGAACGATTGGC CGCAAGGCCAGAGGAAAAGTTGTCAAGG ATGAGCAGGGAGCAACAAAAGTAGCTGG AATGCTGCGAAACGAACCGGGAGCGCTGT GAATACAGTGCTCCCTTTTTTTATT	Restriction Enzyme Cloning (SalI and XbaI)	GAGCCATATGA CCGTCGACAGG GAGTCAGAC	TCCGCTTCTAGAA ATAAAAAAAGGG AGCACTGT	NA	NA
tRNA-Glu^{UUC}	GTCCCCTTCGTCTAGAGGCCCAGGACACC GCCCTTTCACGGCGGTAACAGGGGTTCGA ATCCCCTAGGGGACGCCA	N/A, studied native transcripts only				

Table B.3. Estimated coefficients and statistical measures of goodness of fit for the regressions of inTher and inTherAcc models, which were optimized using in vivo data.

***inTher* Model**

	Coeff. Value	P-value
Intercept	1.83	2.02E-04
ΔG_{asT}	0.96	1.13E-03
ΔG_{tF}	1.92	2.99E-04
ΔG_{asT} : ΔG_{tF}	1.27	8.85E-05

***inTherAcc* Model**

	Coeff. Value	P-value
Intercept	2.101	1.02E-05
ΔG_{asT}	1.838	4.56E-09
ΔG_{tF}	2.013	2.53E-05
θ	2.395	7.06E-08
ΔG_{asT} : ΔG_{tF}	2.090	3.55E-10
θ : ΔG_{tF}	2.832	4.06E-09

Table B.4. Summary of performance results for the prediction of mRNA targets for Z. mobilis.

Zms-4											
intaRNA					inTherAcc						
Gene	Energy	IntaRNA Ranking	log 2 fold change (sRNA/MS2 only control)	Experimental ranking	Gene	log 2 fold change (sRNA/MS2 only control)	Region	inTherAcc ranking	Extreme	Region BLAST ranking	Exper-imental ranking
ZMO0205	-17.49	25	0.923225612	3	**ZMO0176**	0.599478765	21	1	high	3	44
ZMO1041	-16.37	40	0.828108232	7	**ZMO1961**	0.57091544	55	4	low	5	52
ZMO0089	-18.92	19	0.617040784	37	**ZMO1335**	0.546674558	21	1	high	2	63
ZMO0885	-18	20	0.601829473	42	**ZMO0570**	0.47759167	71	2	high	5	92
ZMO0395	-20.3	7	0.579522929	50	**ZMO1697**	0.466774026	21	1	high	1	98
ZMO1335	-187.7	1	0.546674558	63	**ZMO0662**	0.408734183	23	5	low	4	138
ZMO1837	-18.97	18	0.454343608	111	**ZMO0731**	0.406662923	43	4	high	4	140
ZMO1622	-19.7	9	0.388983563	153	**ZMO1993**	0.369586541	30	1	low	1	170
ZMO0471	-19.64	13	0.365823892	175	**ZMO1063**	0.342934029	30	1	low	5	194
ZMO1083	-16.16	43	0.35677769	181	**ZMO2014**	0.311430483	26	3	low	4	222
ZMO1042	-16.37	39	0.301431468	245	**ZMO0208**	0.296899392	48	3	high	3	254
ZMO0888	-20.2	8	0.28376573	271	**ZMO0716**	0.293823265	43	4	high	3	259
ZMO0392	-20.9	5	0.268298177	293	**ZMO1623**	0.265157228	26	3	low	2	299
ZMO1623	-19.67	11	0.265157228	299	**ZMO0060**	0.216261044	70	5	high	2	376
ZMO0050	-16.05	48	0.217360146	373	**ZMO0773**	0.20480575	21	1	high	5	389
ZMO1084	-16.16	44	0.214714721	378	**ZMO0047**	0.189251415	55	4	low	1	421
ZMO1822	-17.202	29	0.186033458	426	**ZMO1060**	0.187421478	26	3	low	3	423
ZMO0116	-16.66	38	0.182849428	433	**ZMO1001**	0.179657399	48	3	high	5	441

Zms-4											
intaRNA					inTherAcc						
Gene	Energy	IntaRNA Ranking	log 2 fold change (sRNA/MS2 only control)	Experimental ranking	Gene	log 2 fold change (sRNA/MS2 only control)	Region	inTherAcc ranking	Extreme	Region BLAST ranking	Exper-imental ranking
ZMO0227	-16.83	33	0.168988676	463	**ZMO0182**	0.176555199	71	2	high	3	450
ZMO1918	-16.88	32	0.160752163	475	**ZMO0778**	0.174391482	43	4	high	5	453
ZMO0394	-19.18	16	0.146717471	492	**ZMO0480**	0.143717518	11	2	low	4	498
ZMO0335	-16.04	49	0.117119149	543	**ZMO0959**	0.124752042	11	2	low	2	528
ZMO1867	-16.02	51	0.098170755	578	**ZMO1064**	0.081710433	30	1	low	5	620
ZMO1404	-16.12	47	0.097264669	581	**ZMO0293**	0.081145441	71	2	high	4	624
ZMO1868	-16.02	52	0.088809648	597	**ZMO1231**	0.078508193	55	4	low	2	631
ZMO0293	-16.16	42	0.081145441	624	**ZMO0219**	0.076093482	23	5	low	2	633
ZMO1231	-16.68	37	0.078508193	631	**ZMO0500**	0.043918869	48	3	high	4	701
ZMO0242	-19.37	14	0.070699766	647	**ZMO1992**	0.040909372	30	1	low	1	707
ZMO0393	-20.9	6	0.068510918	656	**ZMO0628**	0.016717232	21	1	high	4	767
ZMO0964	-16.27	41	0.01894543	762							

Top 18% of mRNAs experimentally enriched (sRNA purification vs MS2 only control

Match Matches between both computational approaches (inTherAcc, IntaRNA)

Zms-6											
intaRNA					inTherAcc						
Gene	Energy	IntaRNA ranking	log 2 fold change (sRNA/MS 2 only control)	Experi-mental ranking	**Gene**	log 2 fold change (sRNA/MS 2 only control)	Region	inTherAcc ranking	Extreme	Region BLAST ranking	Experi-mental ranking
ZMO0388	-17.31	36	0.461669079	15	**ZMO1457**	0.530862219	47	3	high	2	8
ZMO0387	-17.31	35	0.45167408	17	**ZMO0912**	0.382186525	20	1	low	1	34
ZMO1997	-31.34	2	0.33258842	60	**ZMO1697**	0.354698975	10	1	high	2	45
ZMO1288	-17.75	25	0.318958168	77	**ZMO1544**	0.282591912	7	2	low	5	109
ZMO1916	-17.36	33	0.303760789	89	**ZMO0192**	0.270830197	28	5	high	2	124
ZMO1781	-17.16	41	0.214281476	196	**ZMO0225**	0.248841992	47	3	high	5	149
ZMO0546	-19.6	9	0.206369273	208	**ZMO0967**	0.246587498	62	2	high	3	151
ZMO0307	-17.01	46	0.188779734	238	**ZMO0307**	0.188779734	20	1	low	5	238
ZMO0569	-19.02	12	0.165079385	289	**ZMO0655**	0.181809881	68	3	low	3	252
ZMO0419	-19.1	11	0.146259702	335	**ZMO0372**	0.147817173	20	1	low	2	326
ZMO1237	-18.39	18	0.1442617	345	**ZMO2008**	0.1400246	16	4	low	5	361

Zms-6											
intaRNA					inTherAcc						
Gene	Energy	IntaRNA ranking	log 2 fold change (sRNA/MS 2 only control)	Experi-mental ranking	**Gene**	log 2 fold change (sRNA/MS 2 only control)	Region	inTherAcc ranking	Extreme	Region BLAST ranking	Experi-mental ranking
			81			65					
ZMO0864	-18.45	17	0.132002716	384	**ZMO0982**	0.138152154	16	4	low	5	371
ZMO1796	-17.55	30	0.125882158	399	**ZMO1461**	0.130097782	10	1	high	4	387
ZMO0975	-17.94	22	0.120905348	420	**ZMO1751**	0.092782579	68	3	low	5	485
ZMO0992	-18.55	16	0.117264557	426	**ZMO1173**	0.086976338	68	3	low	2	507
ZMO1915	-17.36	34	0.115433431	431	**ZMO1221**	0.080154702	7	2	low	3	534
ZMO1442	-23.52	5	0.106134164	452	**ZMO1923**	0.062063524	25	4	high	4	594
ZMO0581	-19.99	8	0.080791084	531	**ZMO1417**	0.030820878	28	5	high	1	714
ZMO0695	-20.52	7	0.080542892	532	**ZMO0072**	0.016196899	68	3	low	1	769
ZMO1167	-17.22	39	0.074835207	552	**ZMO1056**	0.007487742	47	3	high	1	806
ZMO1223	-17.29	37	0.065937172	579	**ZMO0672**	0.006812157	7	2	low	1	808

Zms-6											
intaRNA					inTherAcc						
Gene	Energy	IntaRNA ranking	log 2 fold change (sRNA/MS 2 only control)	Experi-mental ranking	**Gene**	log 2 fold change (sRNA/MS 2 only control)	Region	inTherAcc ranking	Extreme	Region BLAST ranking	Experi-mental ranking
ZMO0938	-18.84	15	0.0598131	600	**ZMO1545**	0.003467185	7	2	low	4	825
ZMO0478	-18.03	20	0.054768634	618							
ZMO0902	-17	50	0.049684077	643							
ZMO0503	-18.19	19	0.046448918	650							
ZMO1527	-17.61	28	0.038042379	686							
ZMO1777	-17.94	23	0.023998877	743							
ZMO0460	-190.03	1	0.004364413	821							

Match

Top 18% of mRNAs experimentally enriched (sRNA purification vs MS2 only control

Matches between both computational approaches (inTherAcc, IntaRNA)

APPENDIX C: SUPPLEMENTARY DATA FOR CHAPTER FOUR

Supplementary figures for Chapter Four

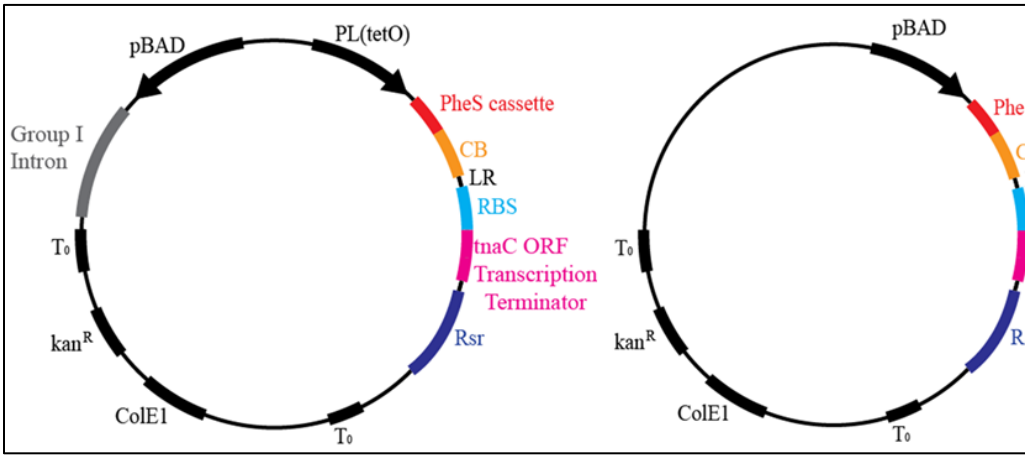
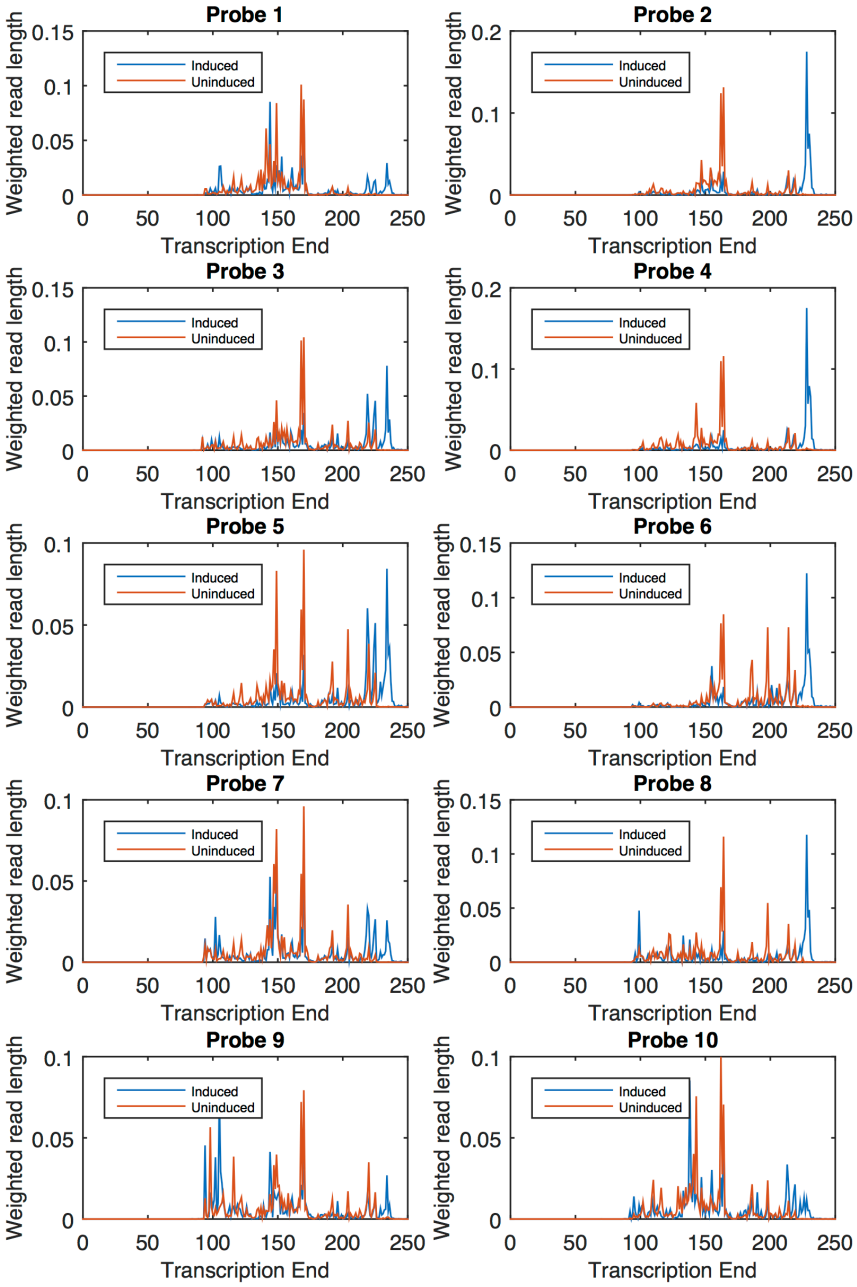
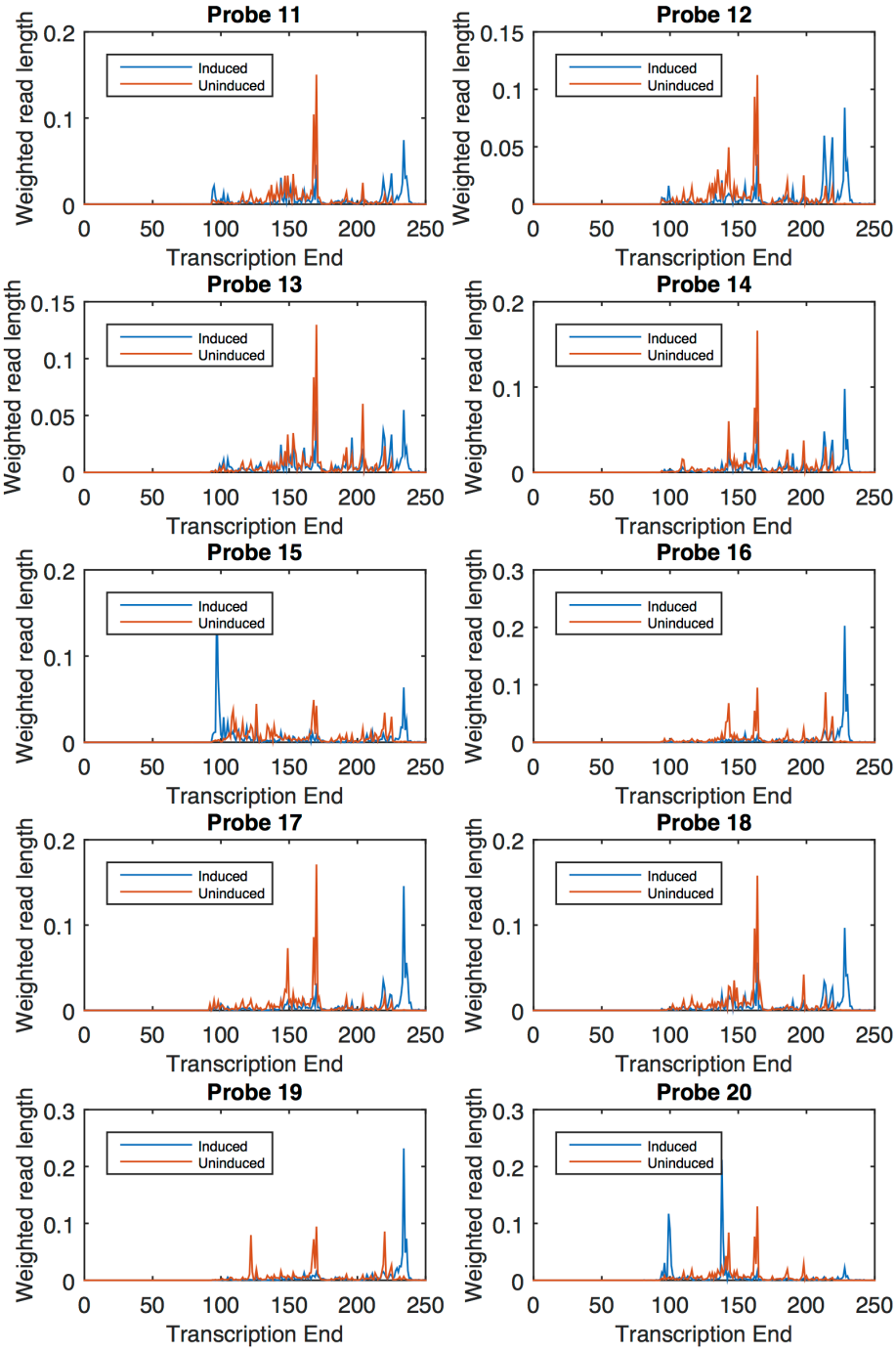


Figure C.1. INTERFACE plasmids (left) for heterologously expressed target RNAs and (right) for native target RNAs.

Figure C.2. INTERFACE traces for each region characterized in the group I intron.




Probe 11
Probe 12
Probe 13
Probe 14
Probe 15
Probe 16
Probe 17
Probe 18
Probe 19
Probe 20
Induced
Uninduced
Weighted read length
Transcription End

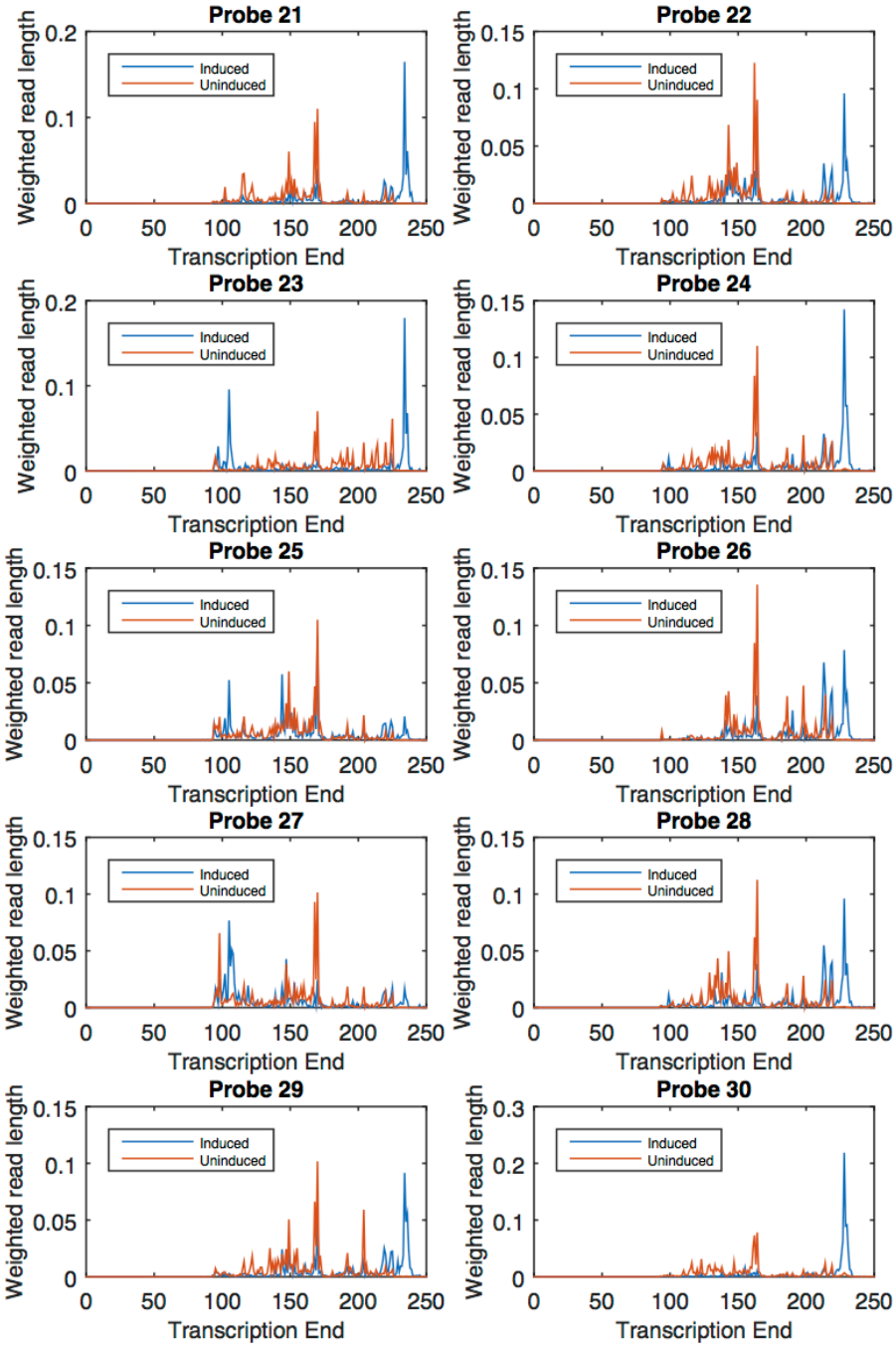
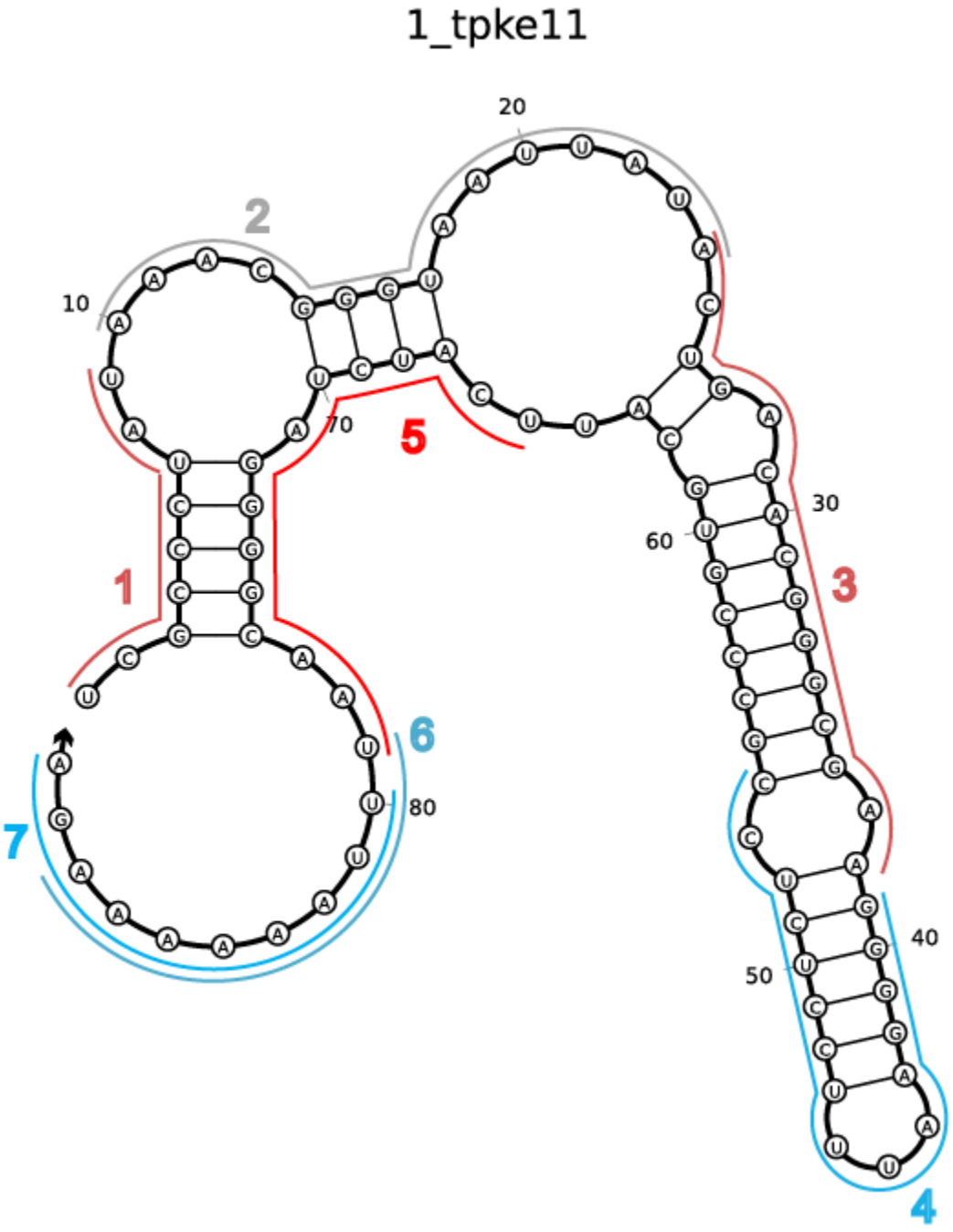
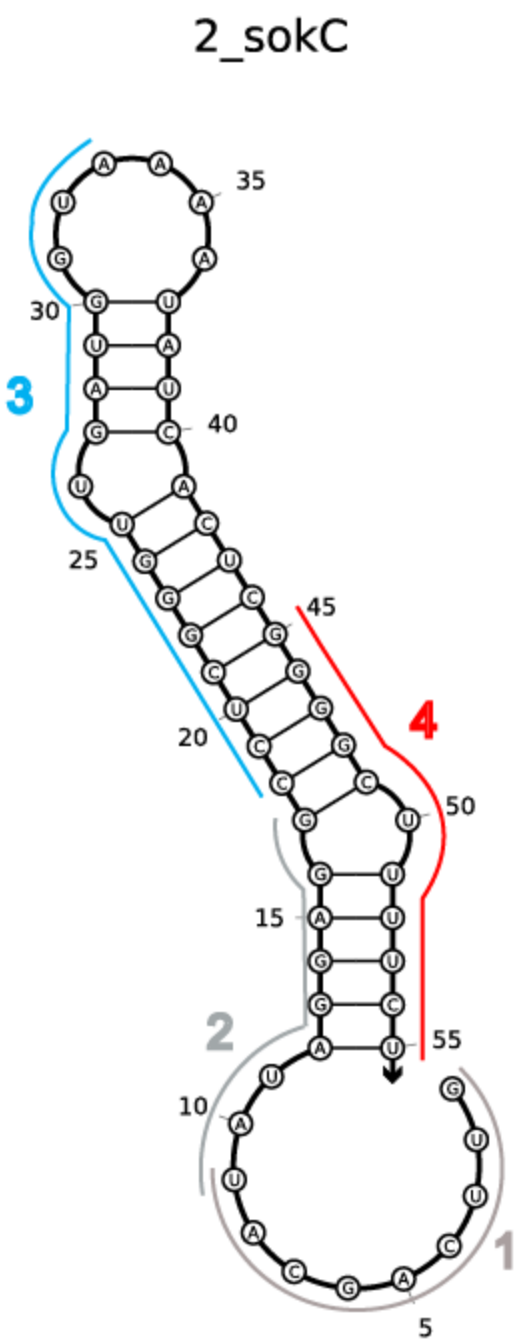
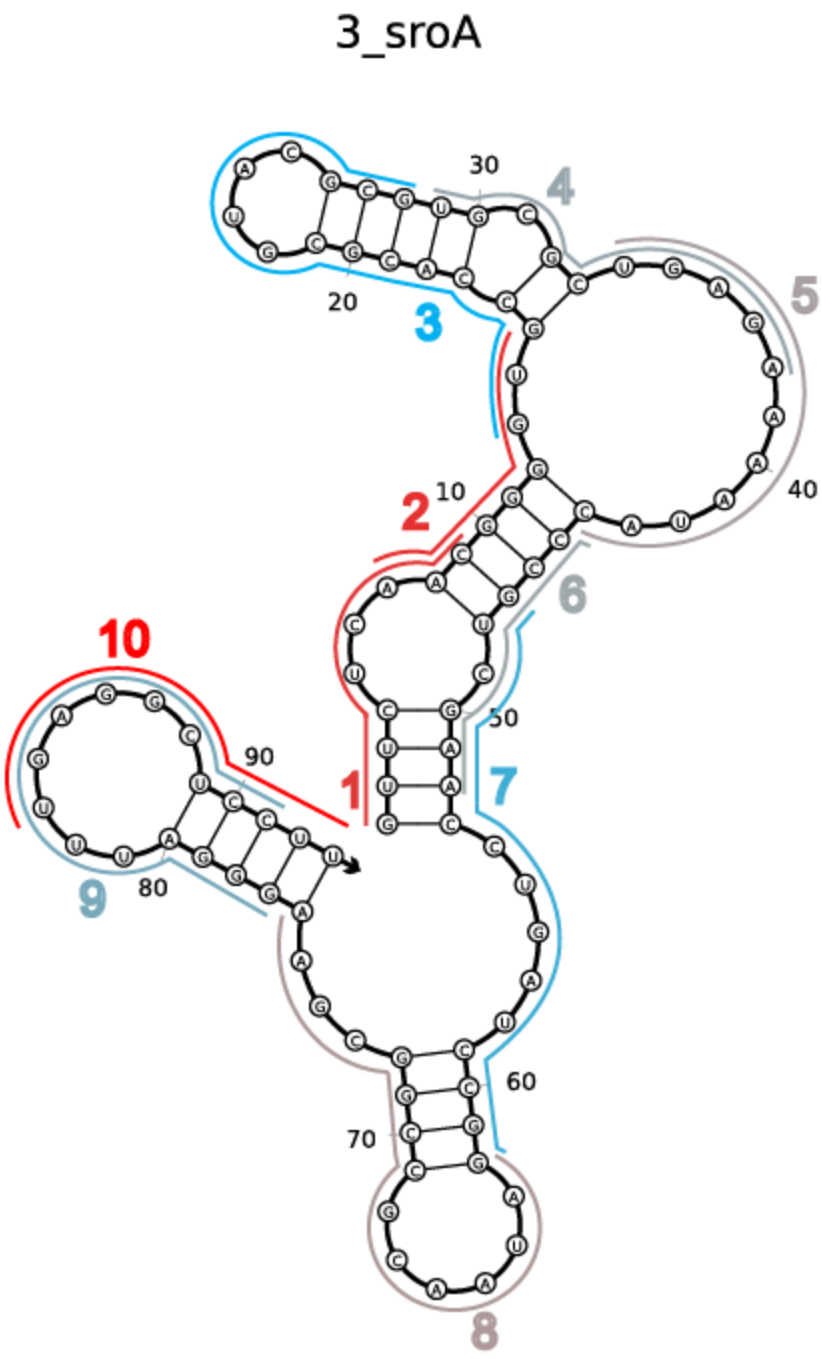

Probe 21
Probe 22
Probe 23
Probe 24
Probe 25
Probe 26
Probe 27
Probe 28
Probe 29
Probe 30
Induced
Uninduced
Weighted read length
Transcription End

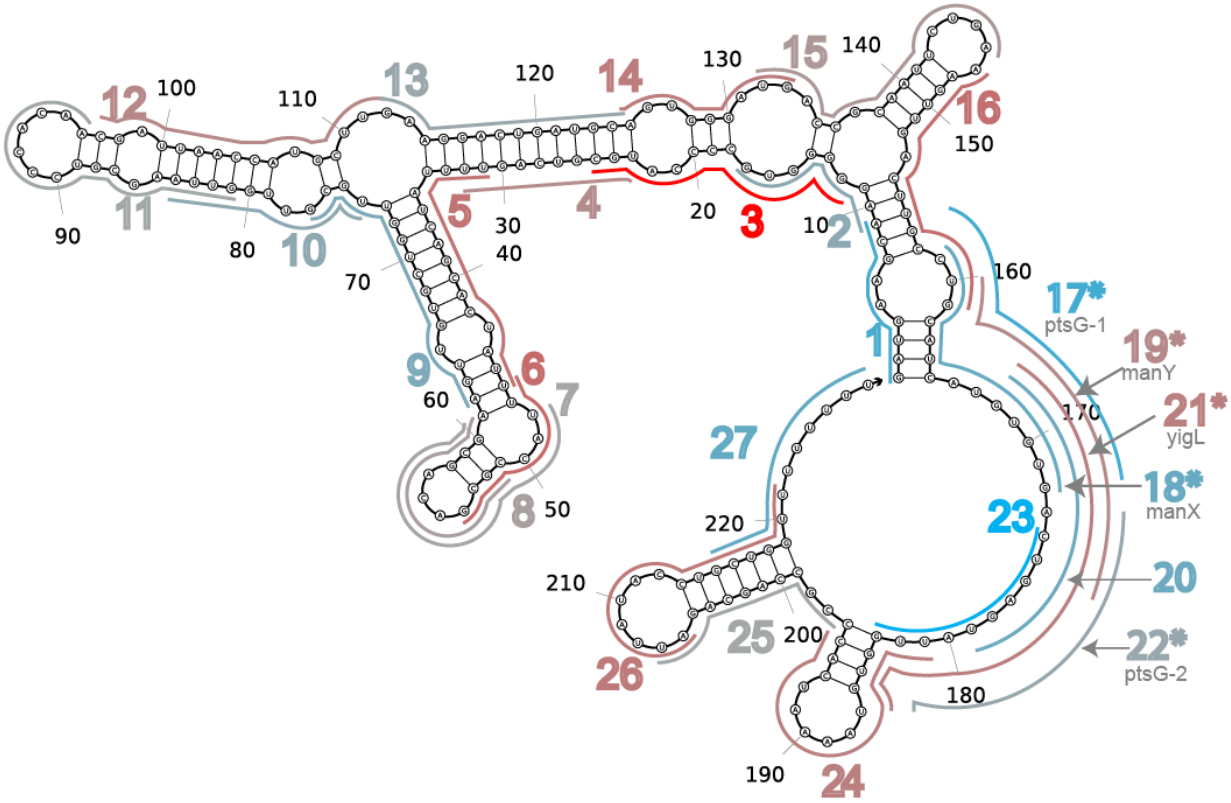
Figure C.3. INTERFACE accessibility heat maps for each sRNA molecule analyzed. Red is accessible, gray is in the middle and blue is inaccessible.

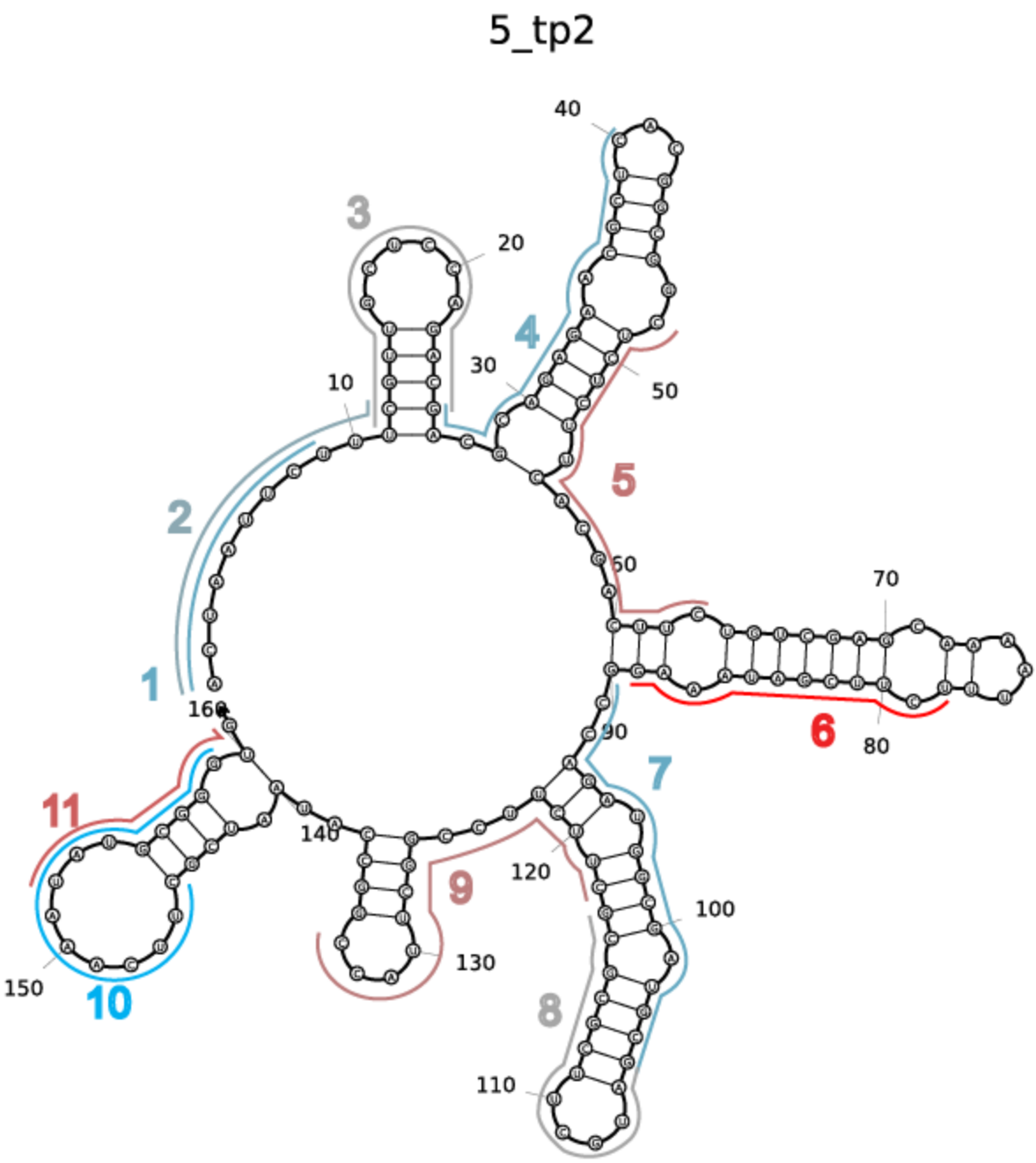


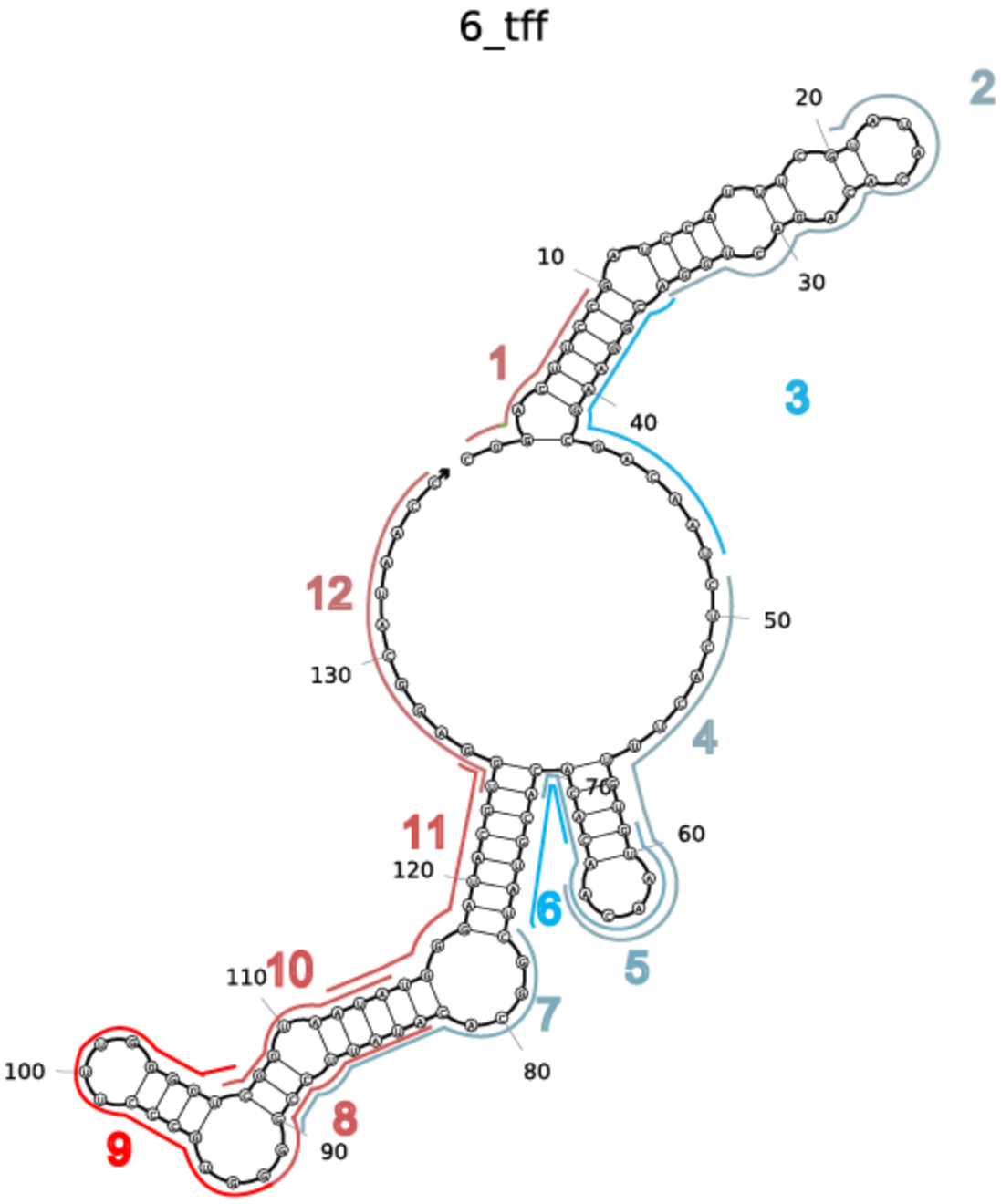

2_sokC
35
30
3
40
25
45
4
20
50
15
2
55
10
1
5

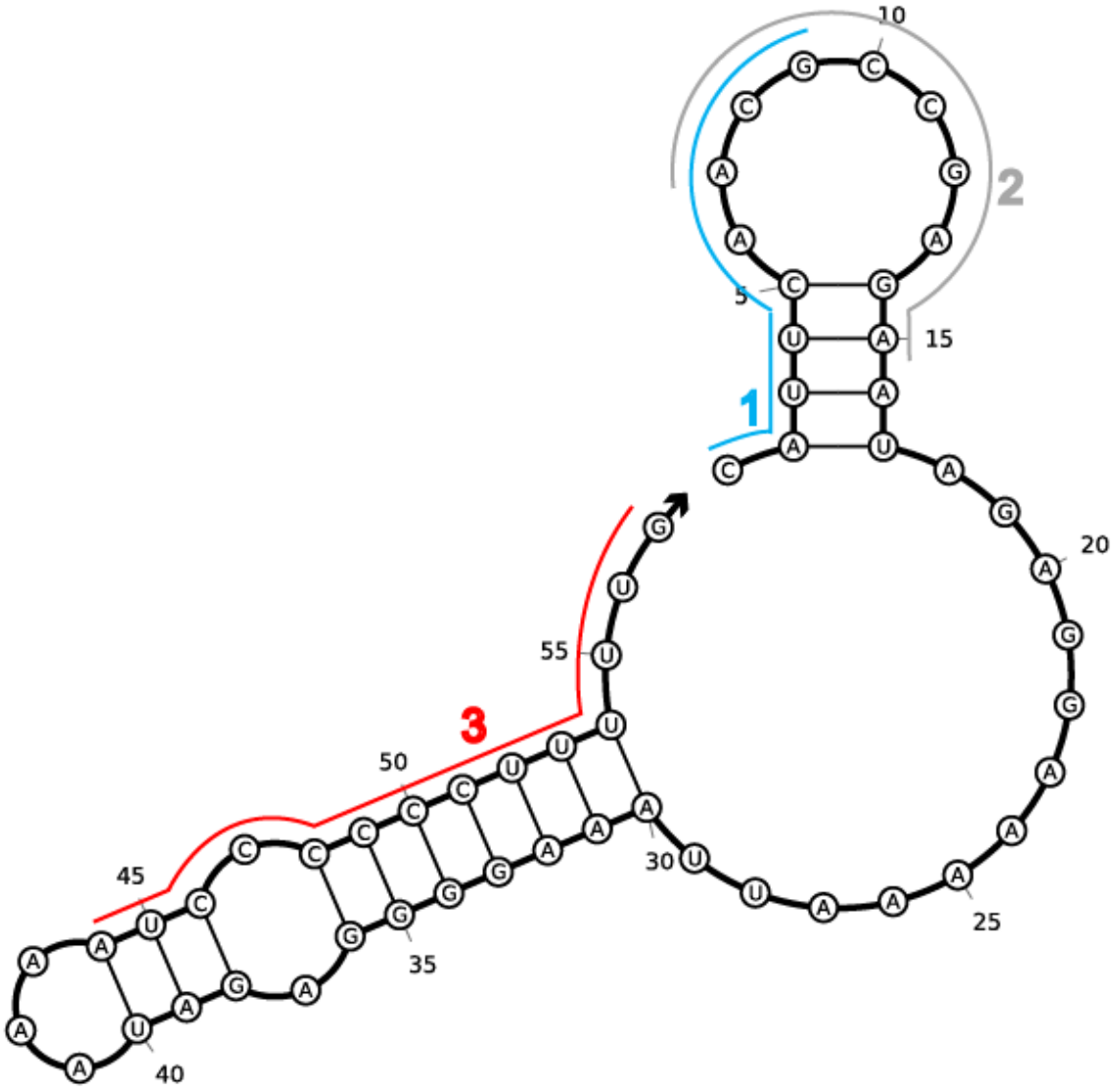

3_sroA
30
4
20
3
5
10
2
40
6
10
50
90
1
7
9
80
60
70
8

4_sgrS

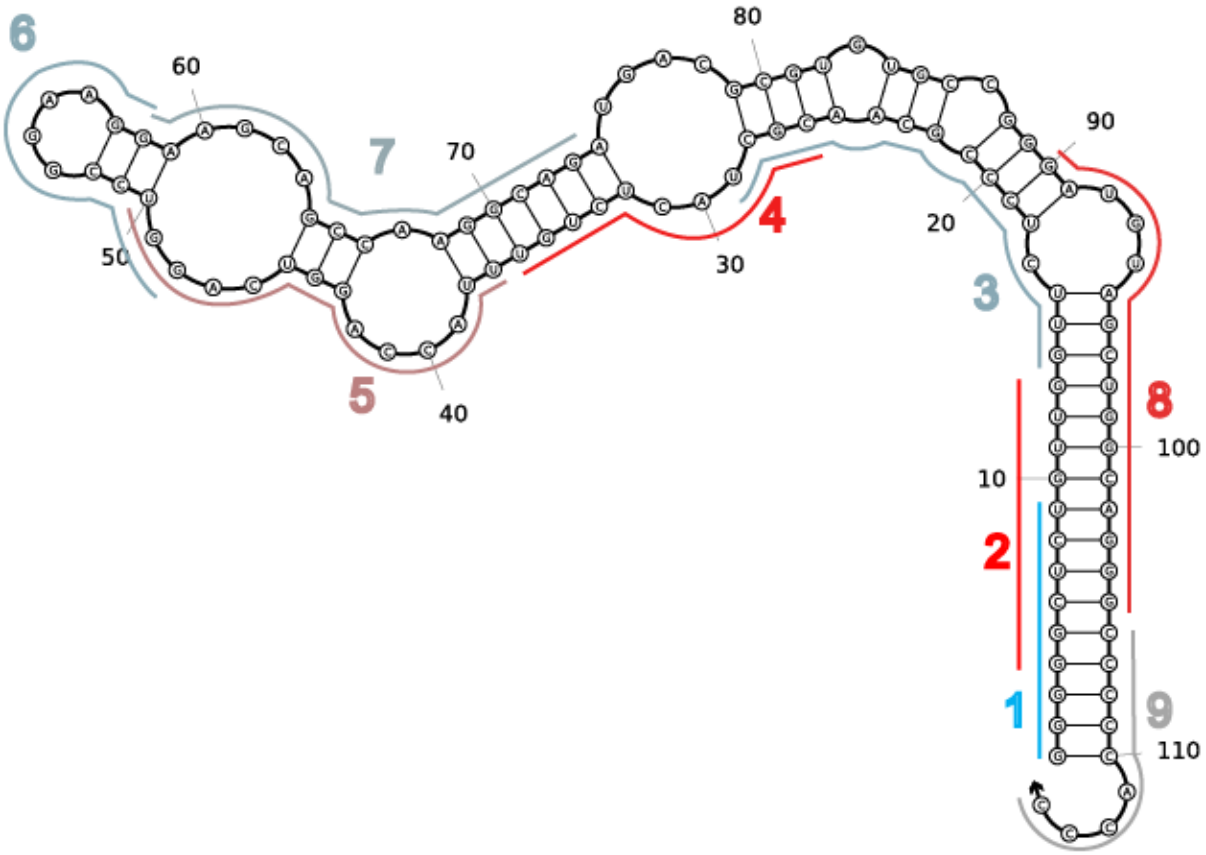

12
100
110
13
120
14
130
15
140
16
150
11
90
80
10
5
30
40
70
4
20
3
10
2
160
17*
ptsG-1
19*
manY
1
9
6
7
60
170
21*
yigL
27
18*
manX
8
50
220
23
210
20
25
200
22*
ptsG-2
26
180
190
24


5_tp2
40
3
20
4
30
10
50
2
5
60
70
1
160
90
6
80
7
11
140
120
9
100
130
150
10
8
110

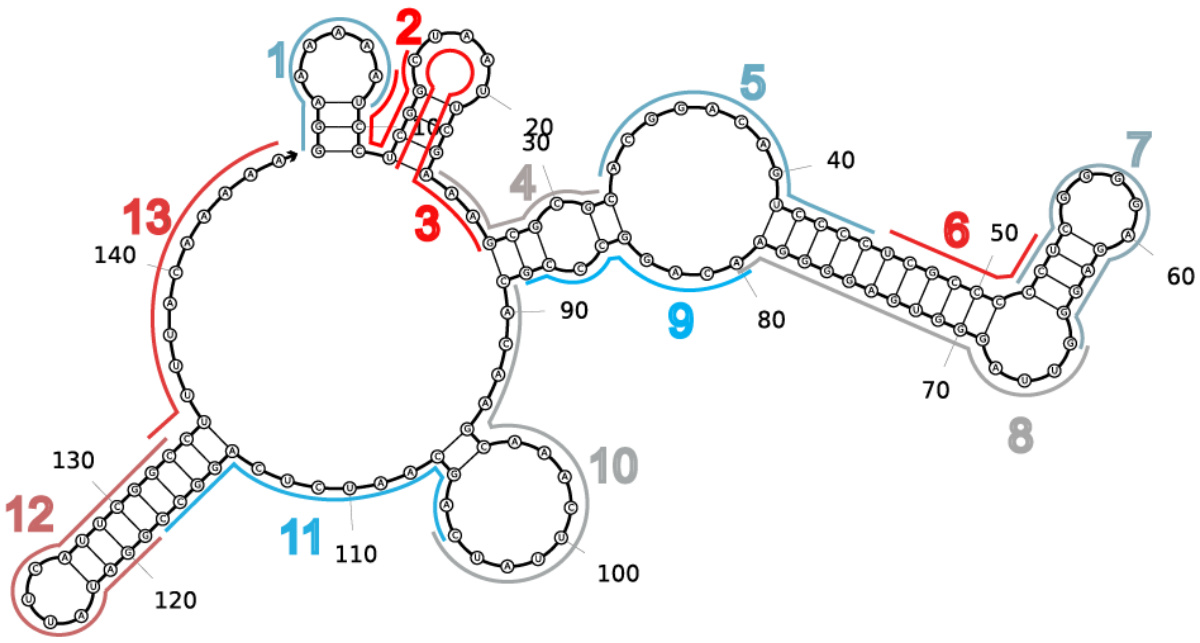

6_tff
20
2
10
30
1
3
40
12
50
130
4
70
11
60
120
6
5
110
10
7
100
80
8
9
90

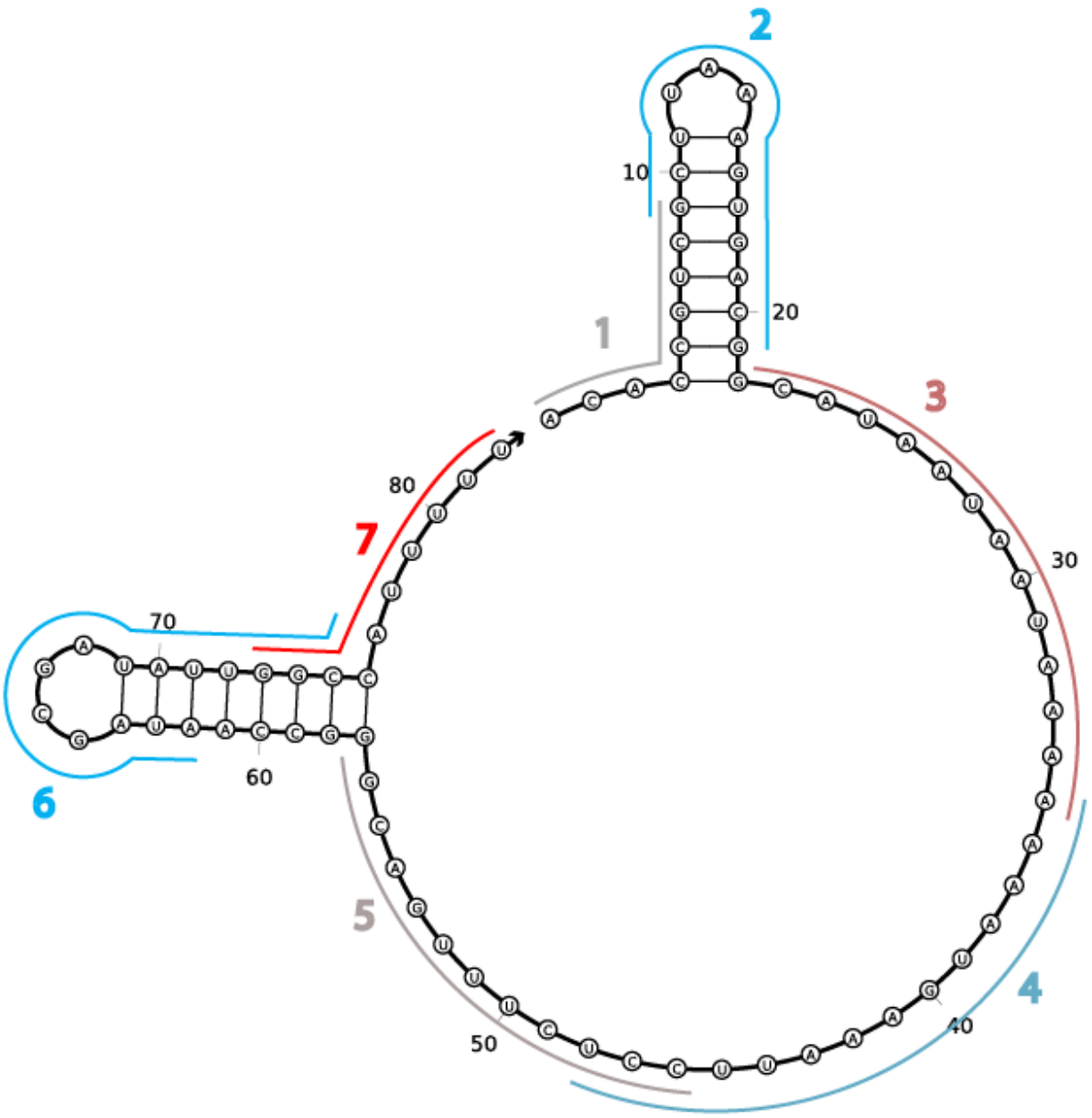

7_sraA
10
2
5
15
1
20
55
3
50
45
30
25
35
40

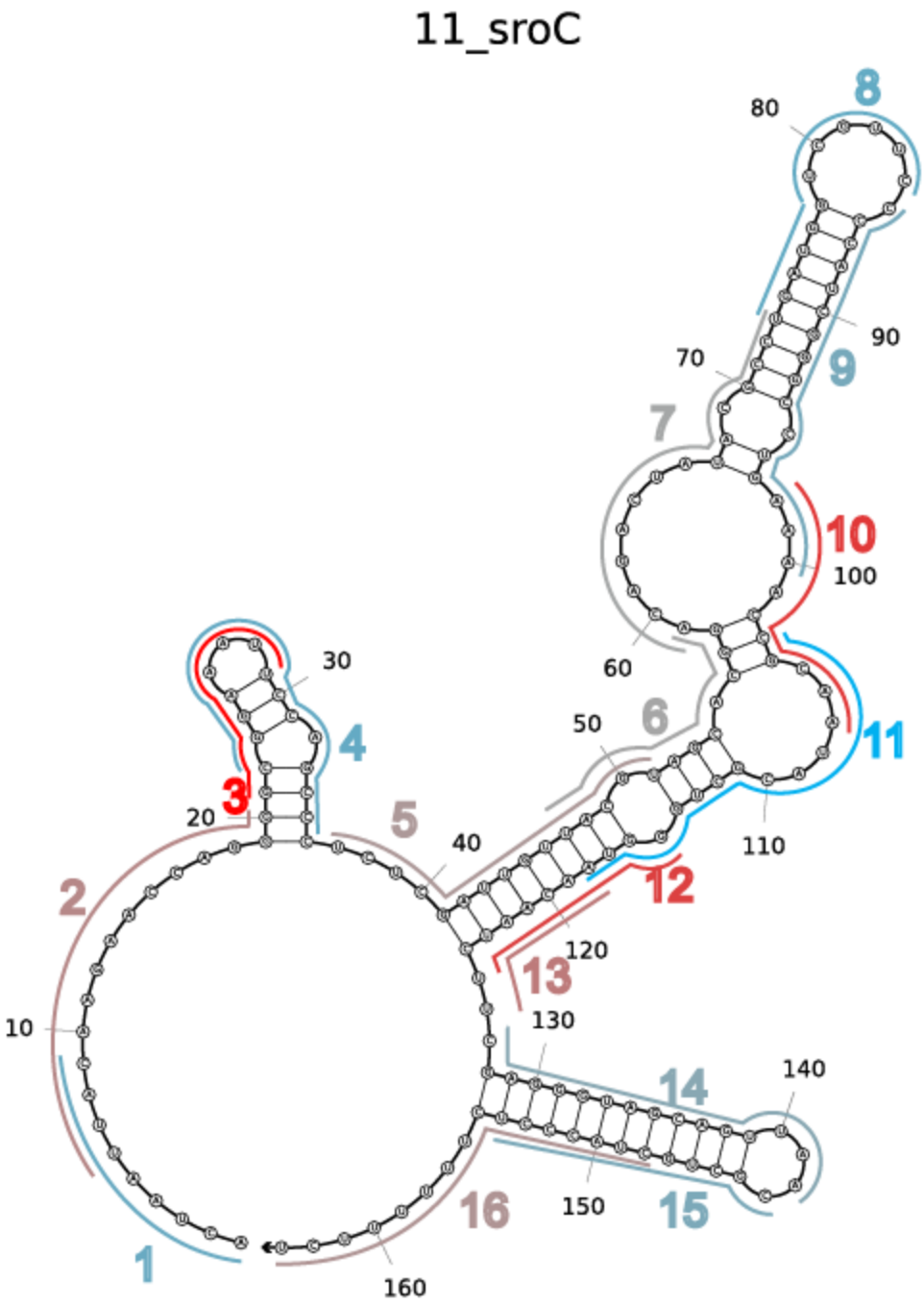
8_ffs

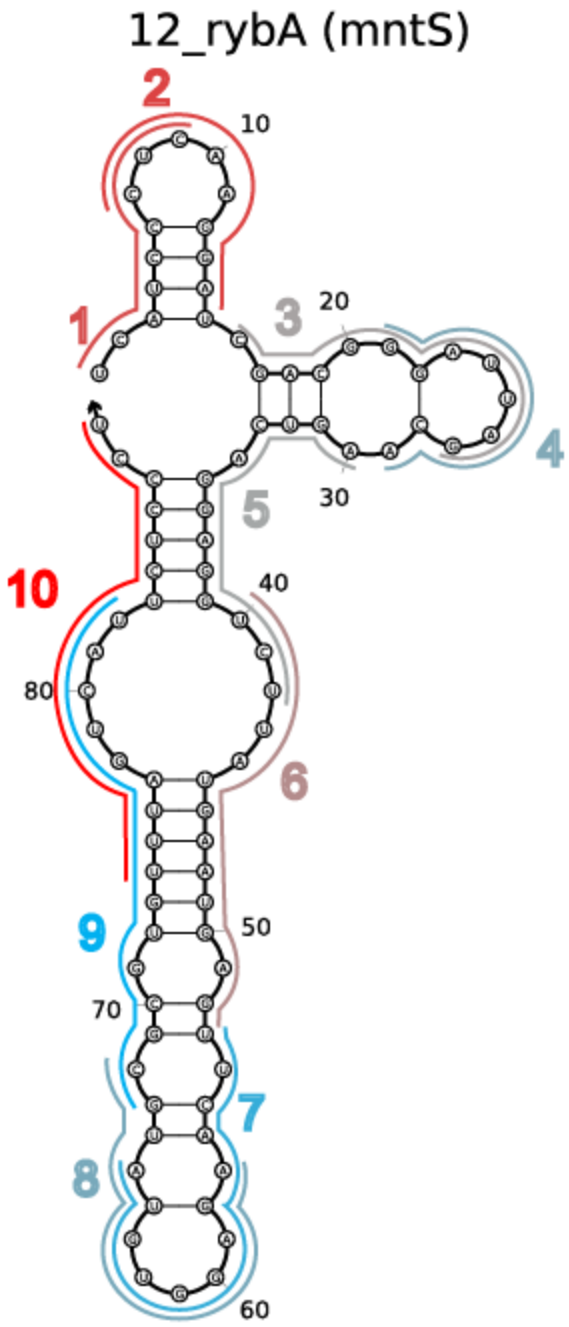


9_nc2

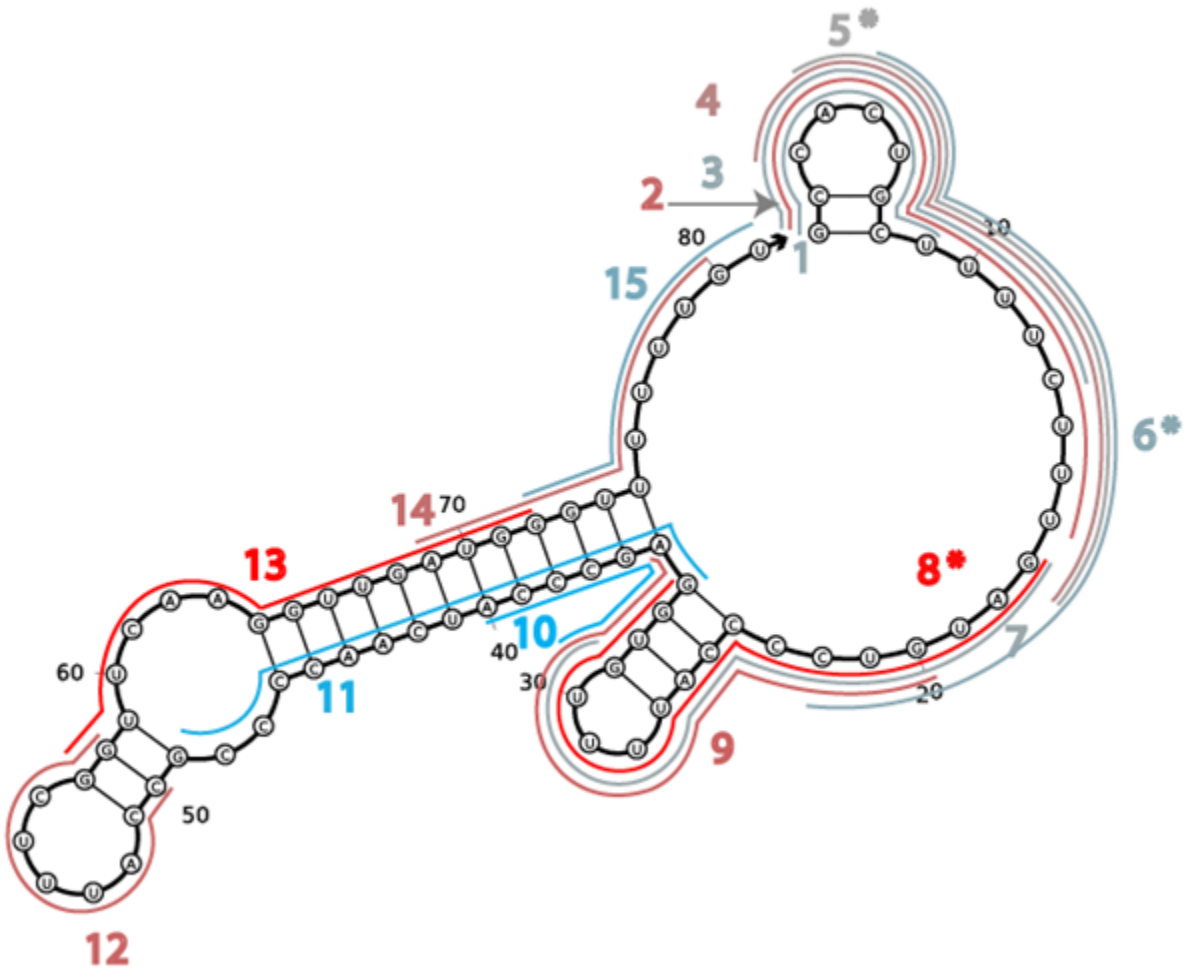


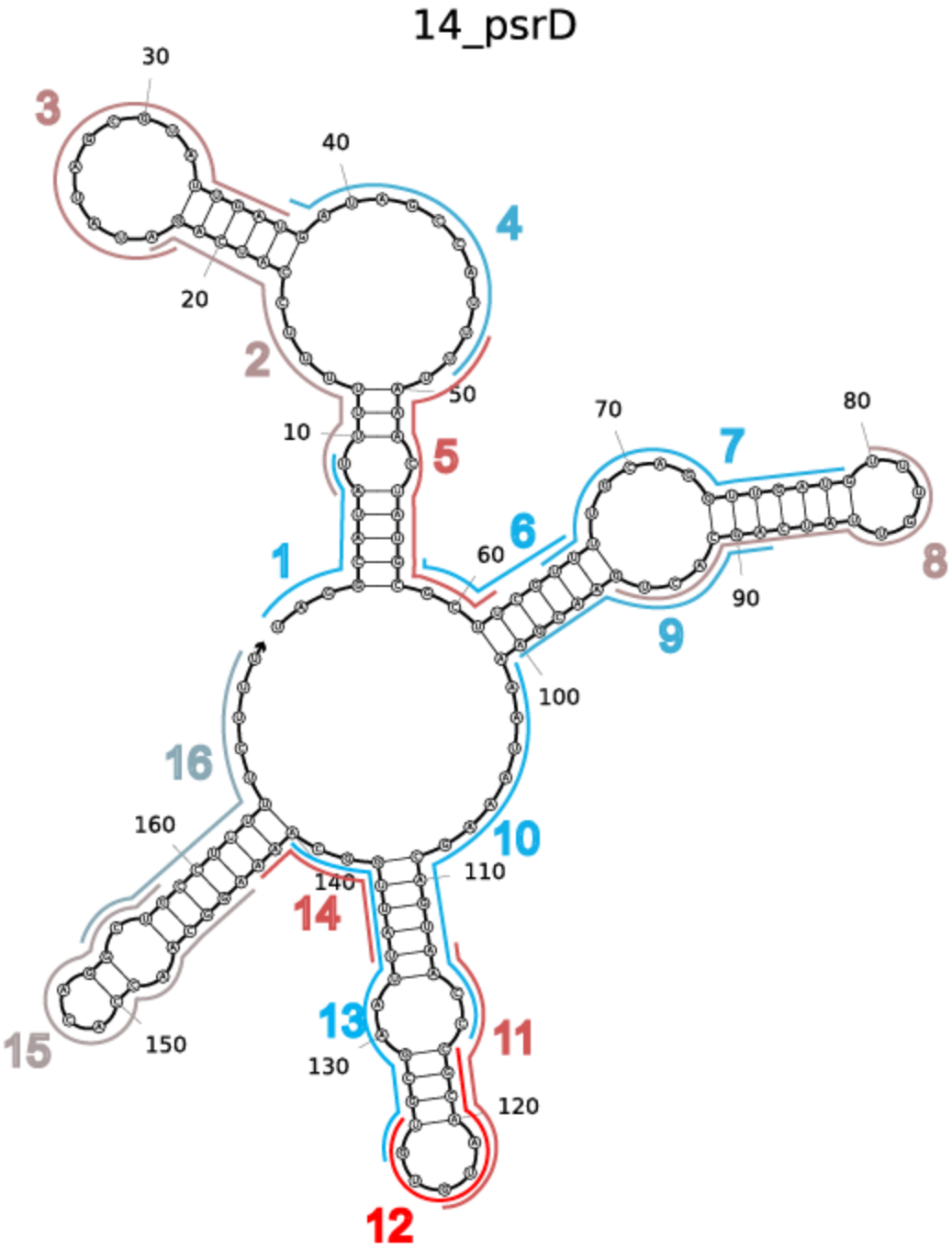

10_sroB (chiX)
1
2
3
4
5
6
7
10
20
30
40
50
60
70
80

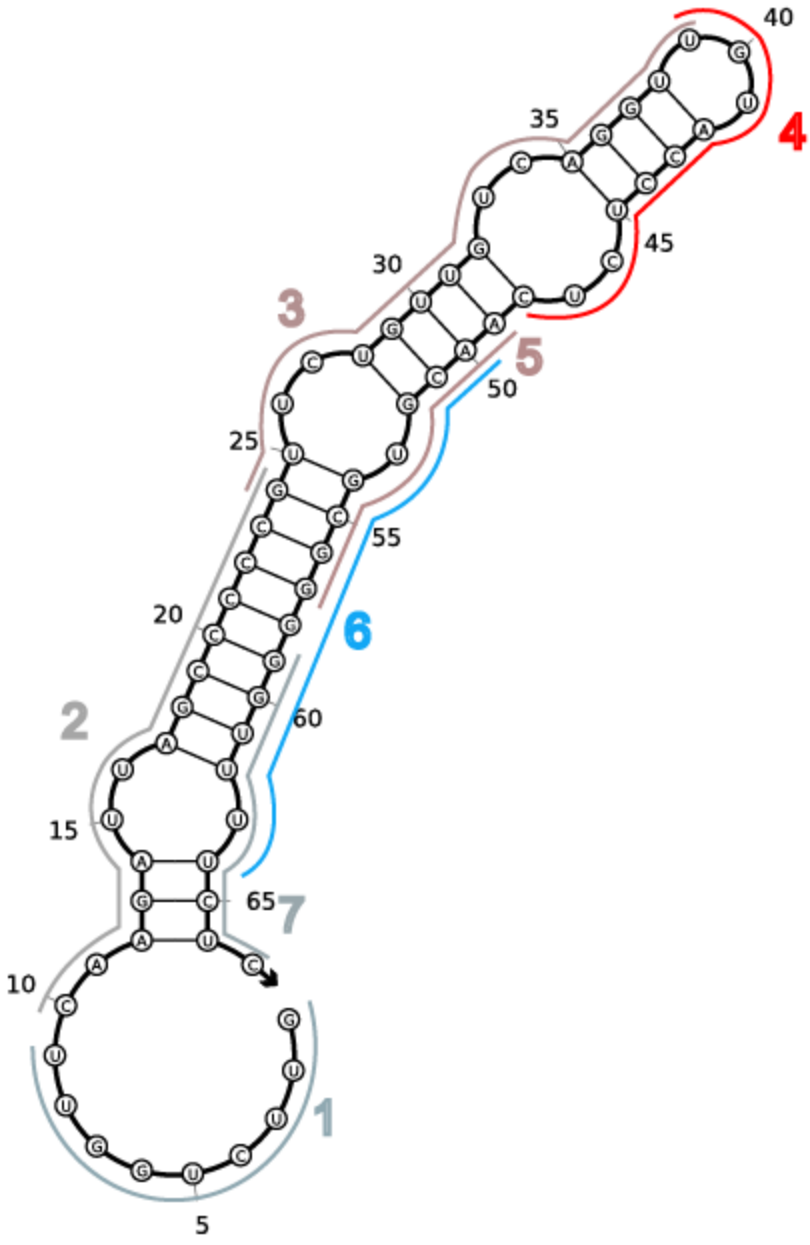

11_sroC
8
80
90
70
9
7
10
100
60
30
4
50
6
11
3
20
5
40
110
2
12
120
13
10
130
140
14
16
150
15
1
160

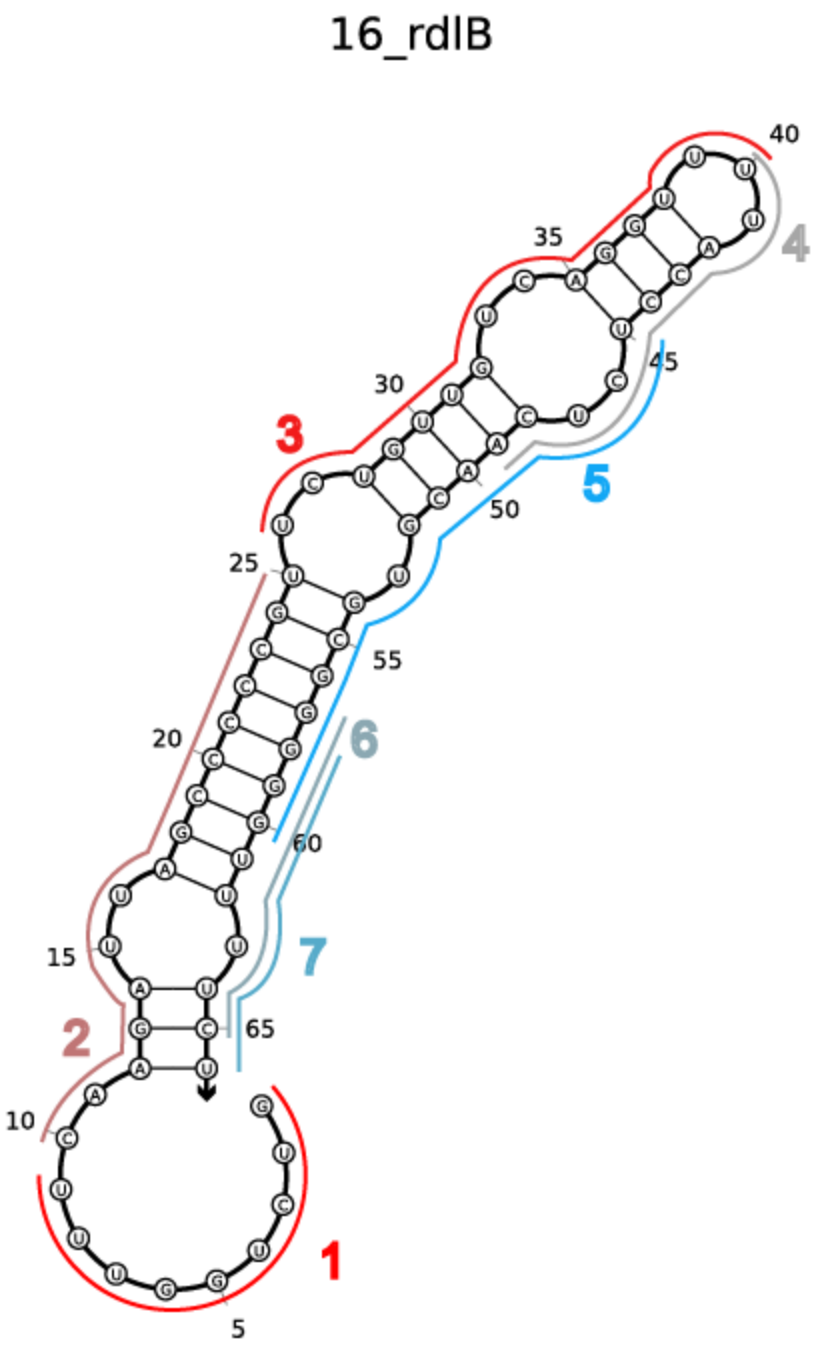

12_rybA (mntS)
2
10
1
3
20
4
30
5
10
40
80
6
9
50
70
7
8
60

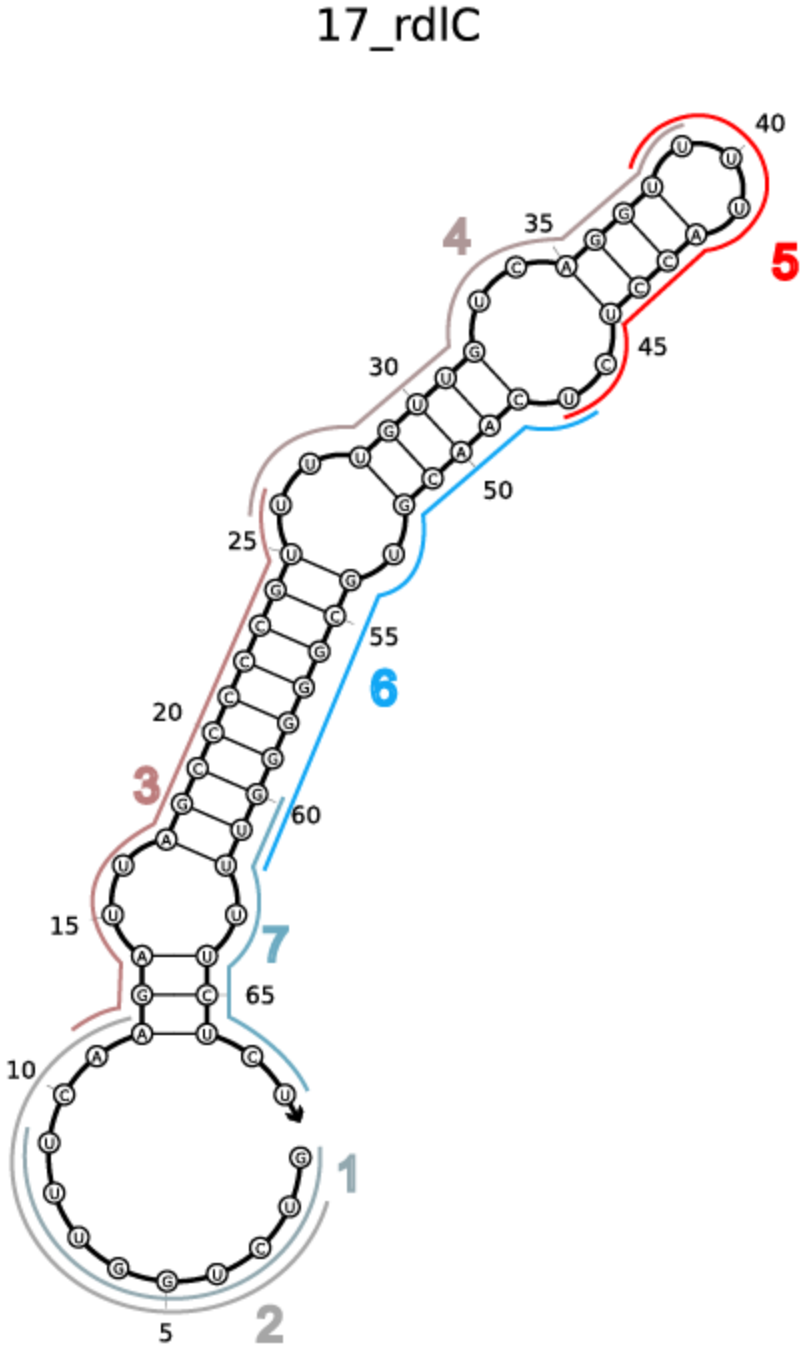
13_rybB



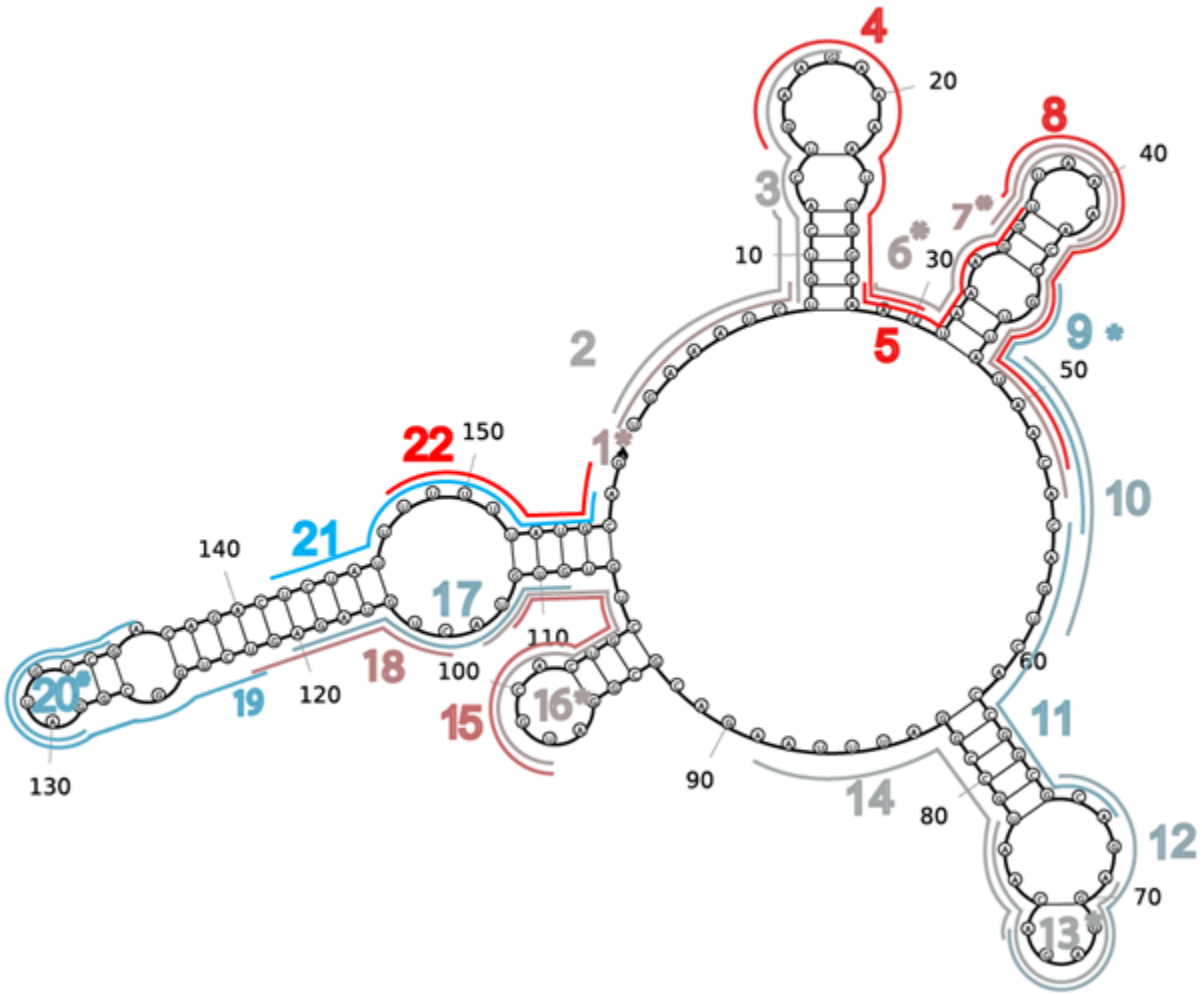

14_psrD
30
3
40
4
20
2
50
10
5
70
7
80
1
6
60
8
90
9
100
16
160
10
110
140
14
13
11
15
150
130
120
12

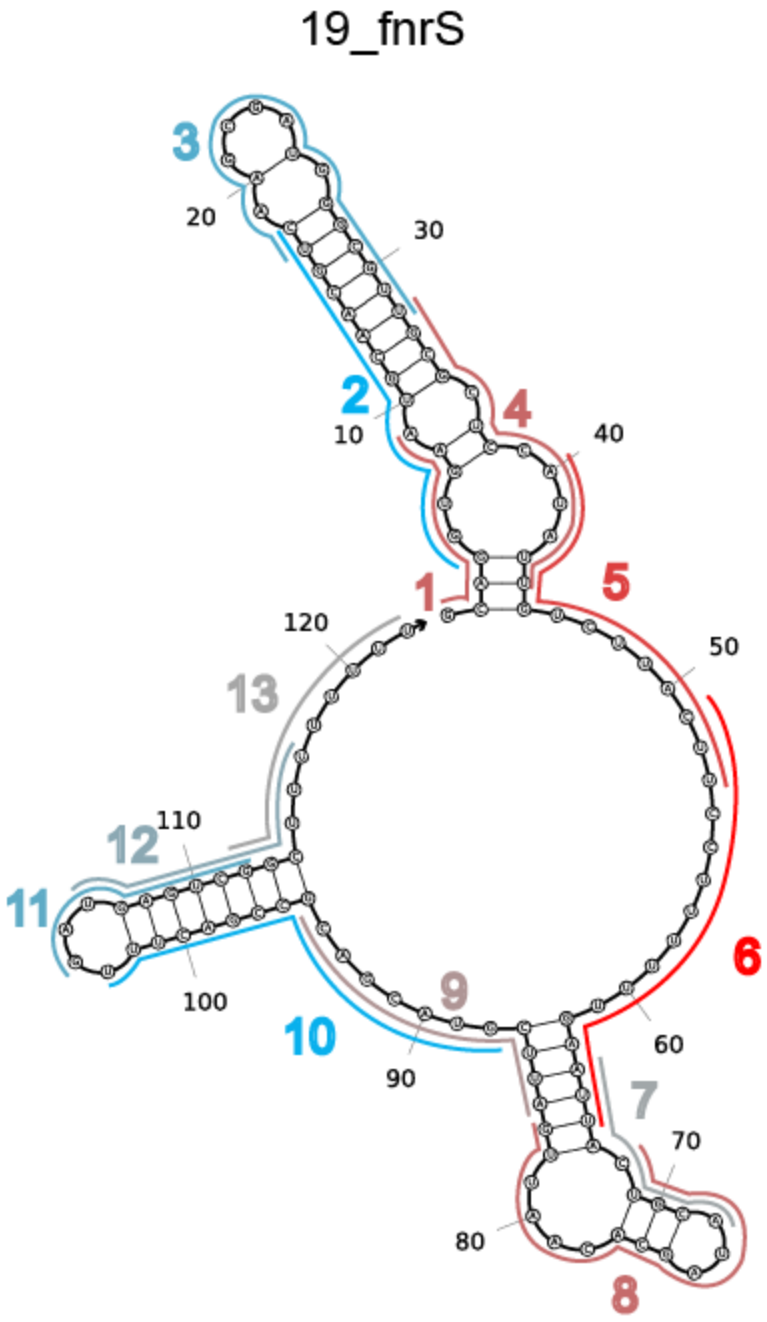

15_rdlA
40
35
4
45
30
3
5
50
25
55
6
20
2
60
15
65
7
10
1
5

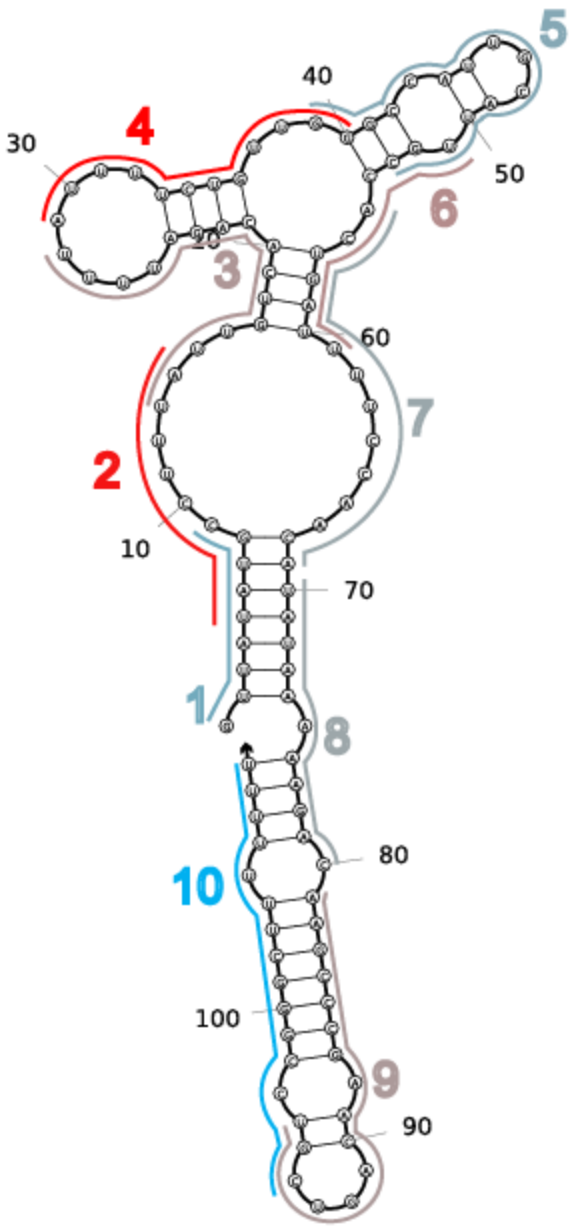

16_rdlB
1
2
3
4
5
6
7
5
10
15
20
25
30
35
40
45
50
55
60
65

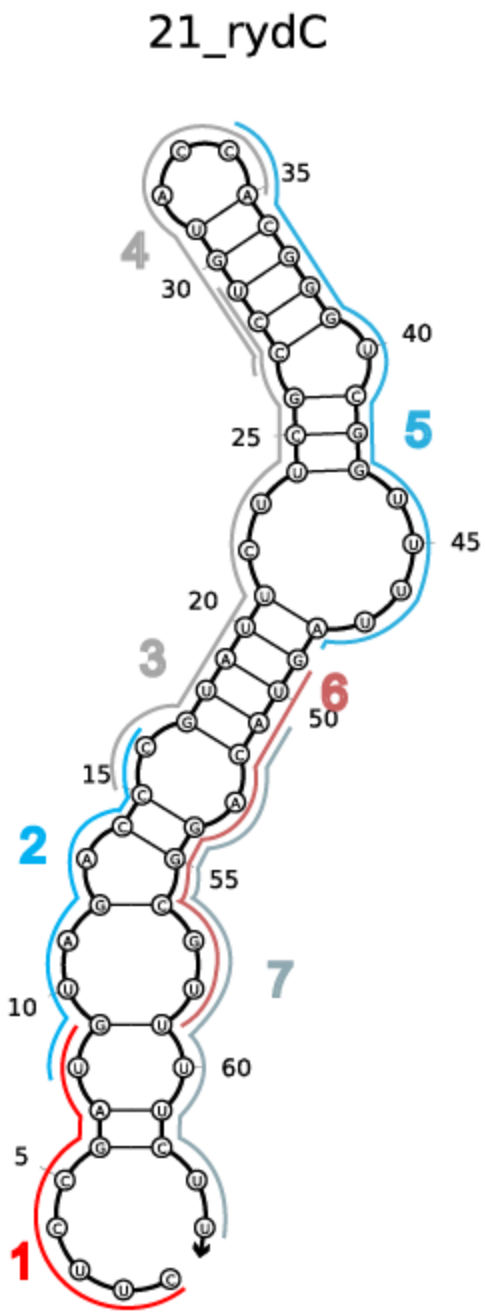

17_rdlC
40
35
4
5
45
30
50
25
55
6
20
3
60
15
7
65
10
1
5
2

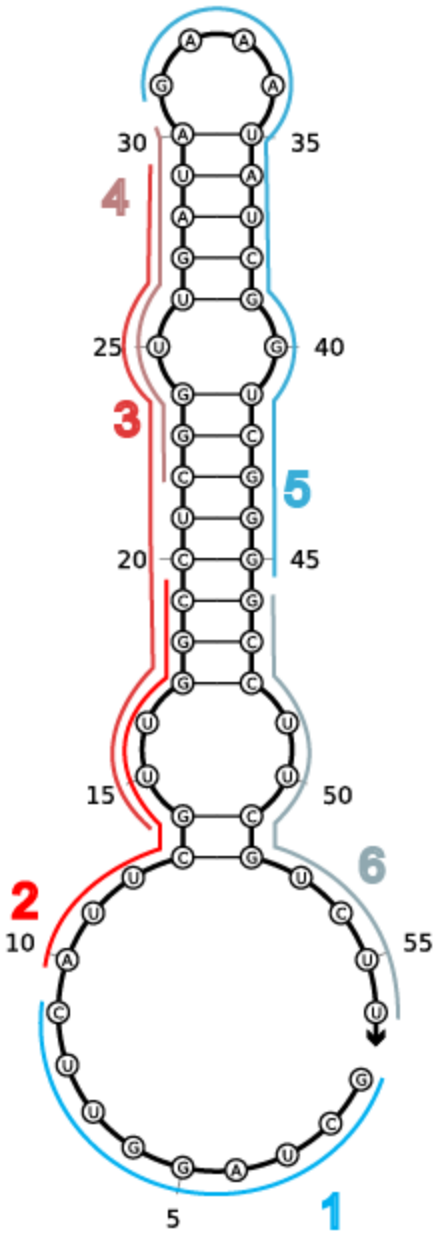
18_mcaS

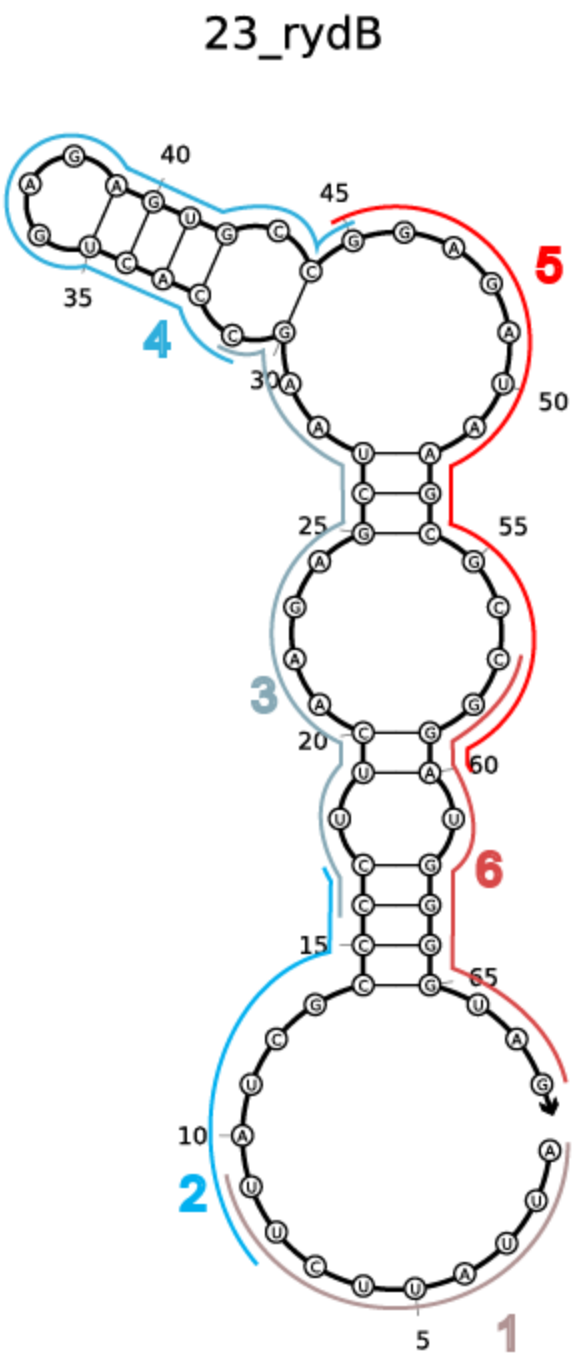

4
20
8
40
3
7*
6*
30
10
2
5
9*
50
22
150
1*
10
140
21
17
110
18
100
60
20*
19
120
15
16*
11
130
90
14
80
12
70
13*

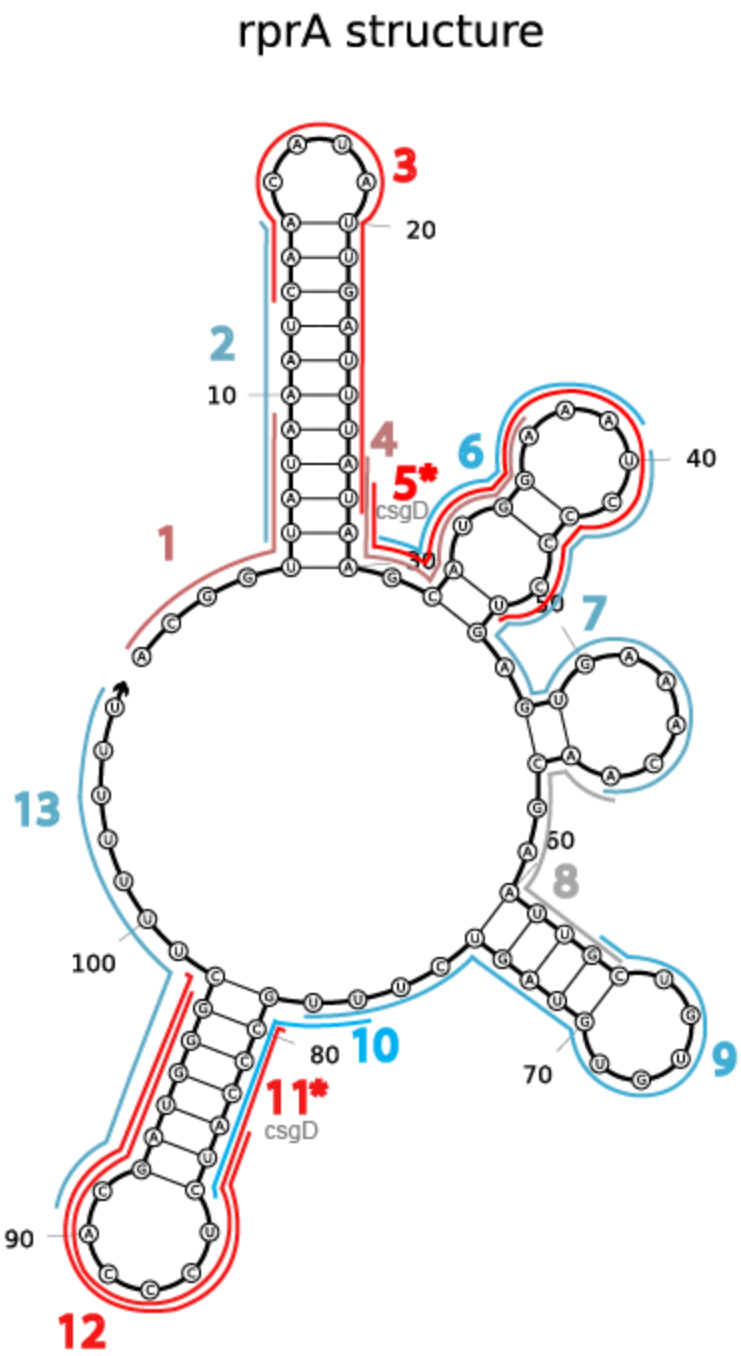

19_fnrS
3
20
30
2
10
4
40
1
5
120
50
13
110
12
11
6
100
9
10
60
90
7
70
80
8


20_micC
5
40
4
30
50
6
3
60
7
2
10
70
1
8
10
80
100
9
90

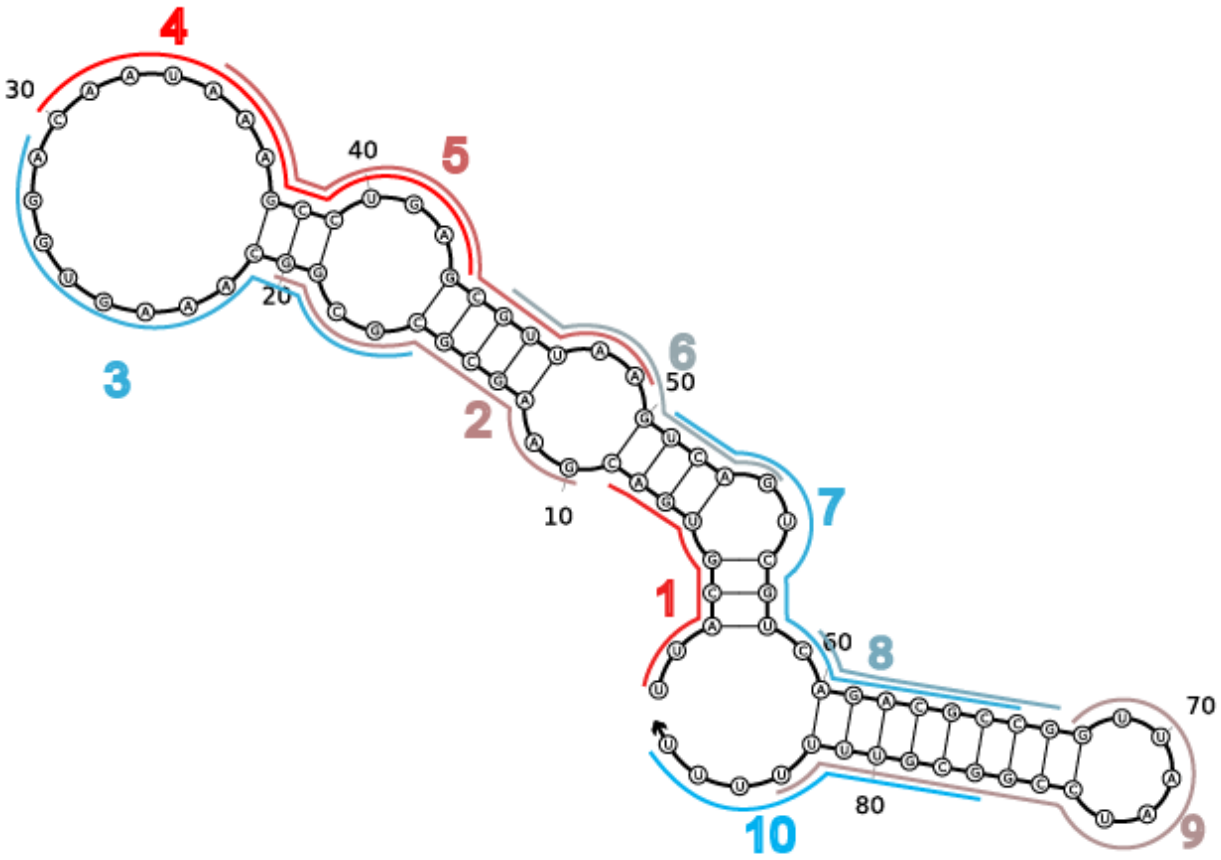

21_rydC
35
4
30
40
5
25
45
20
3
6
50
15
2
55
7
10
60
5
1


22_sokB
30
35
4
25
40
3
5
20
45
15
50
6
2
10
55
5
1

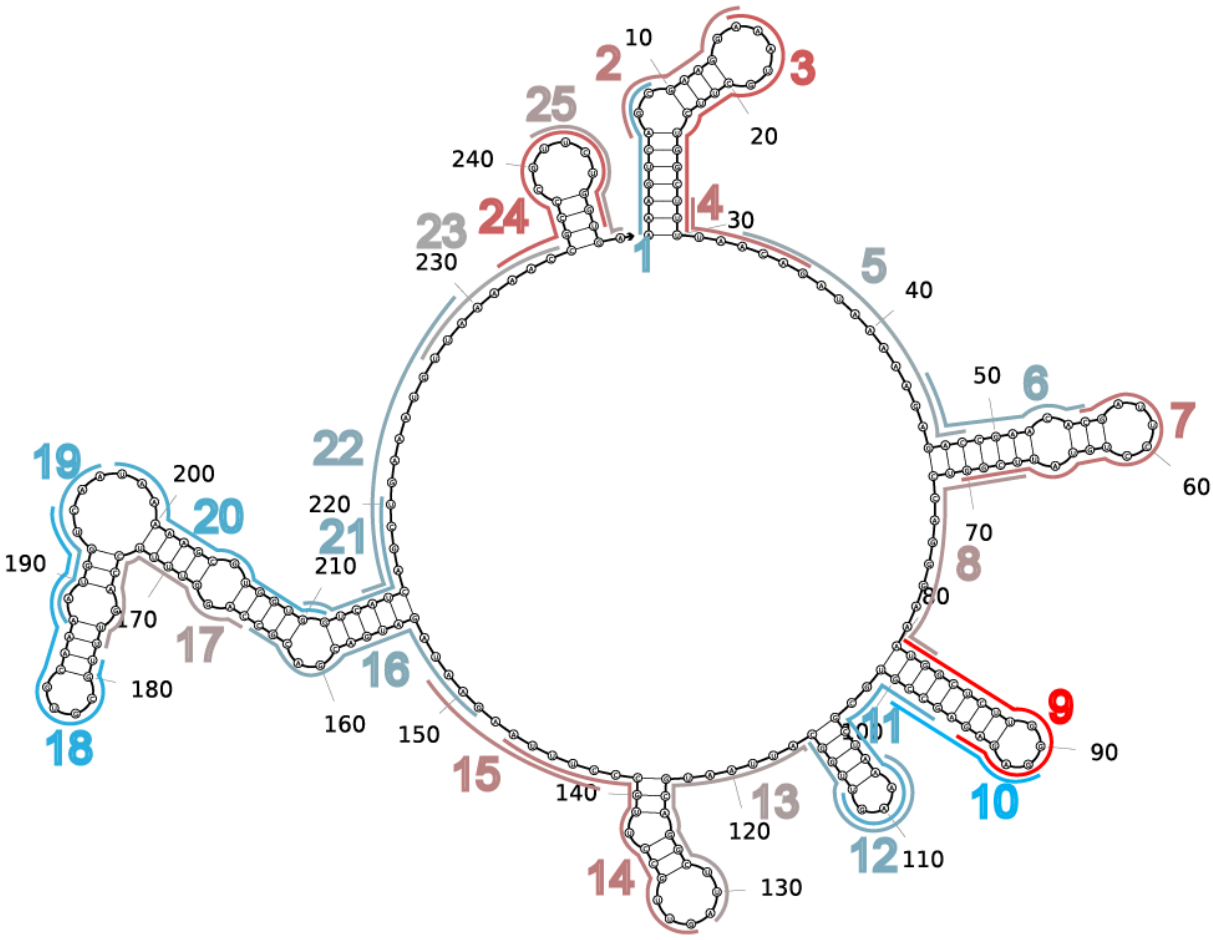

23_rydB
40
45
5
35
4
50
30
25
55
3
20
60
6
15
65
10
2
5
1

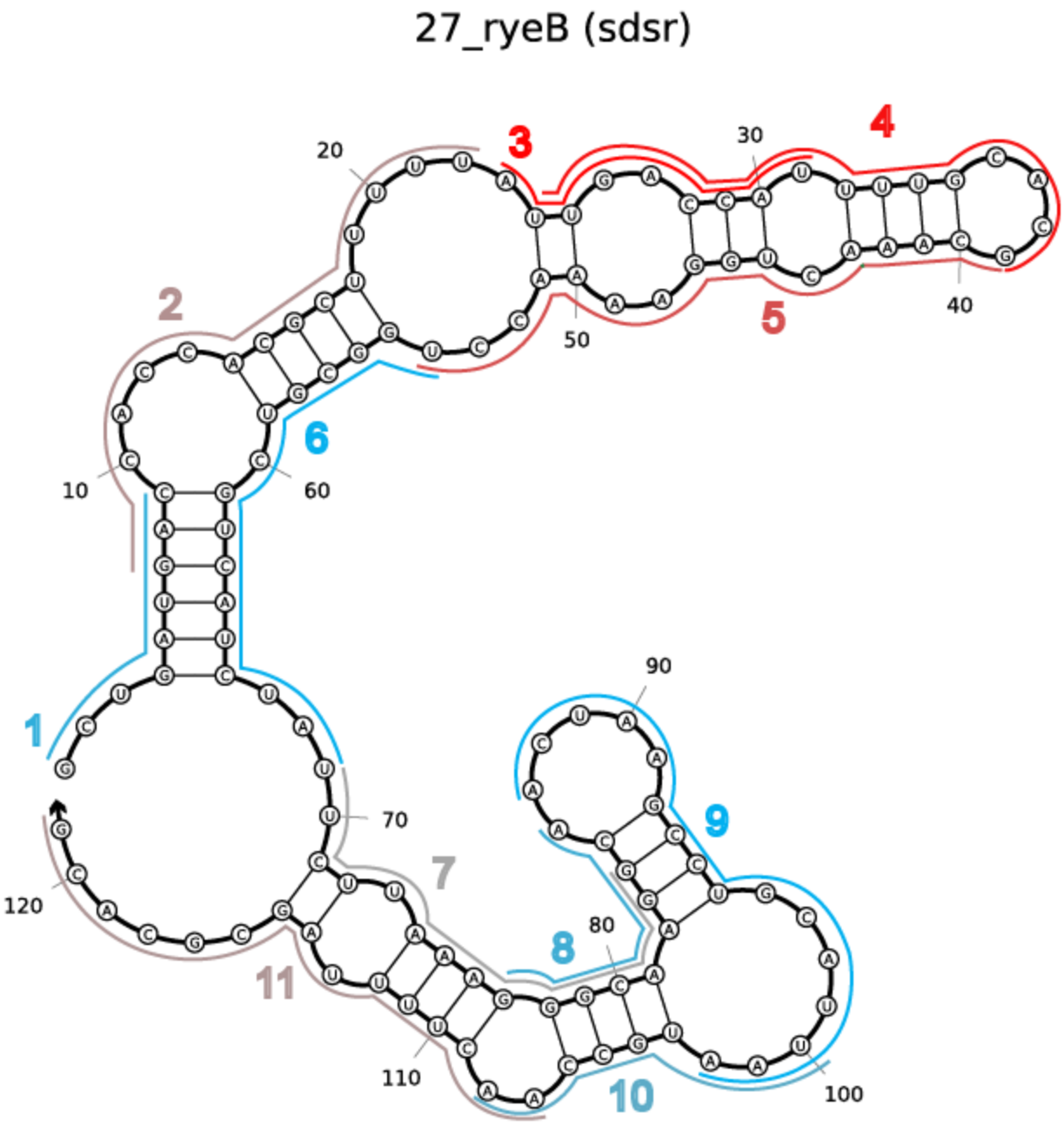

rprA structure
3
20
2
10
4
5*
csgD
6
40
1
30
7
13
50
8
100
10
80
11*
csgD
70
9
90
12

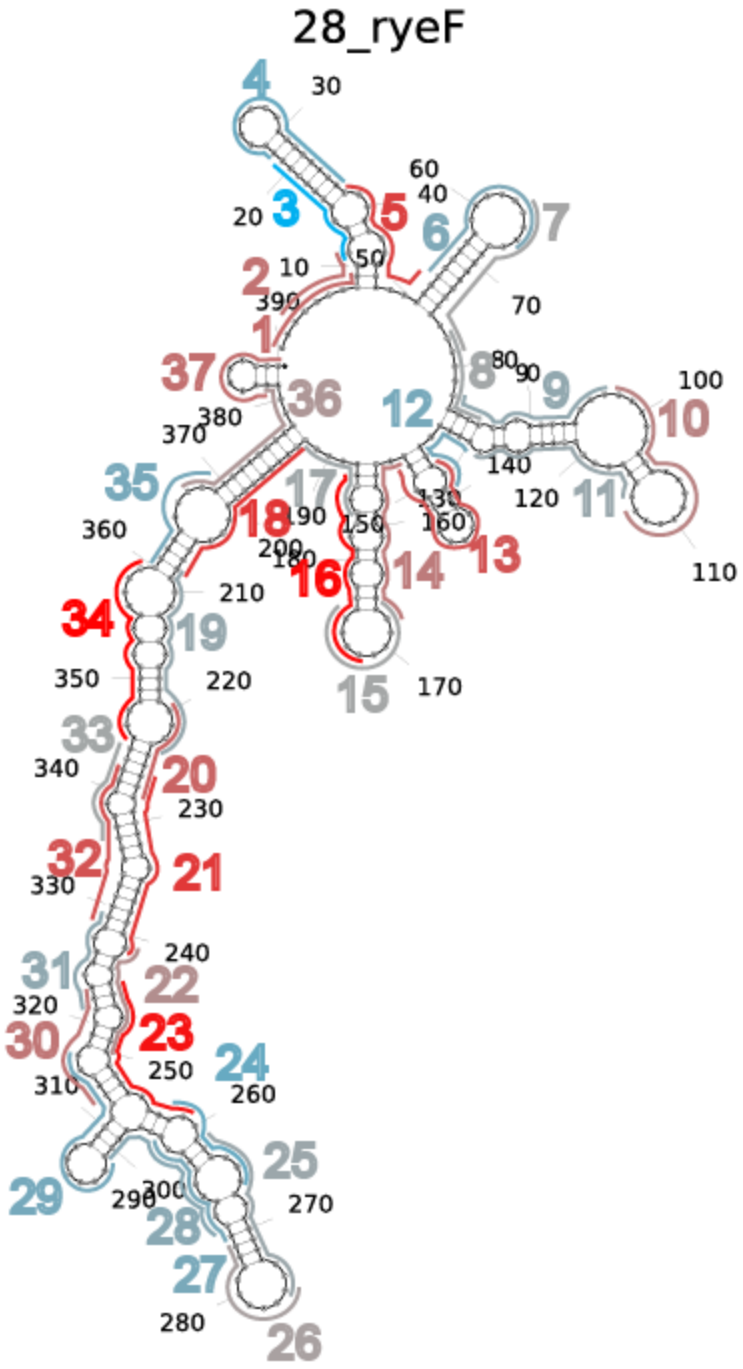
25_sroD



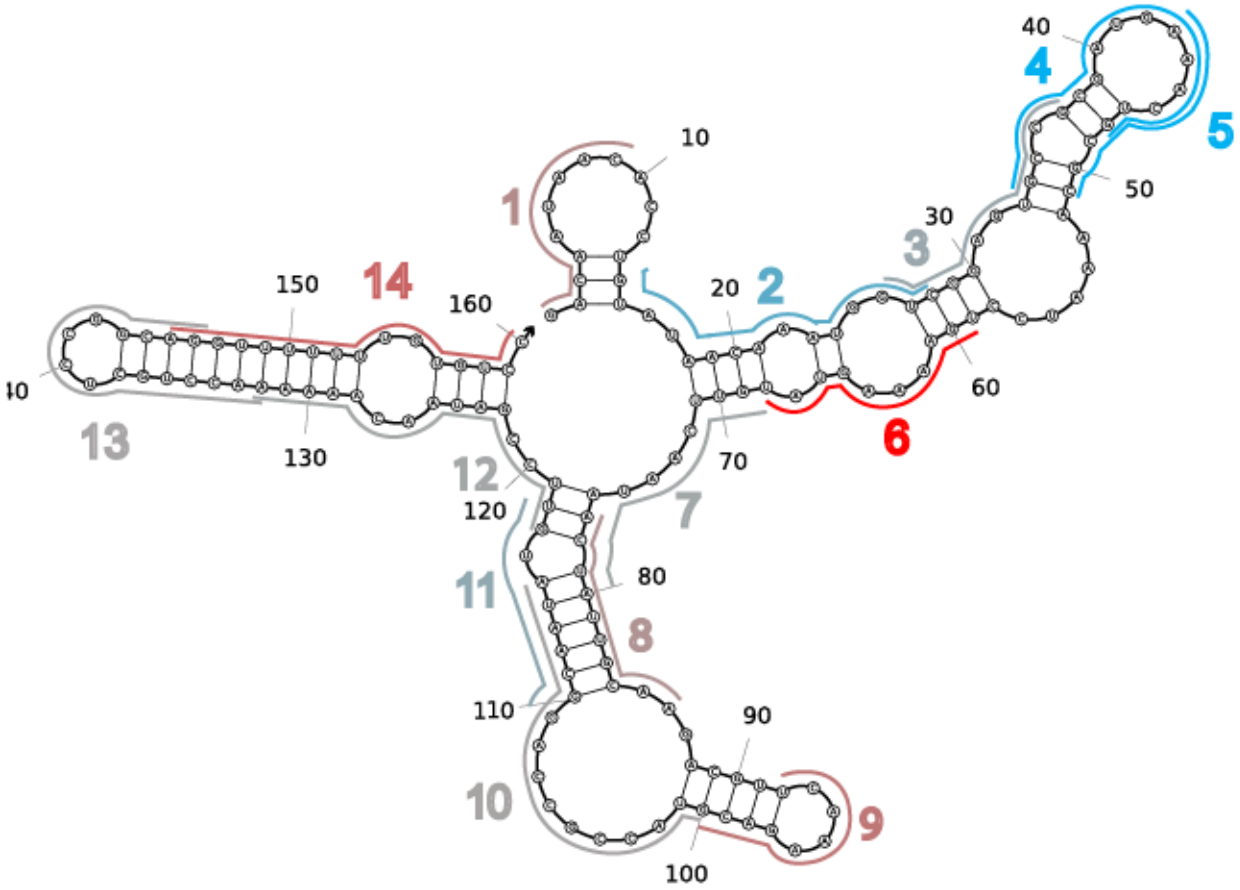
26_ryeA

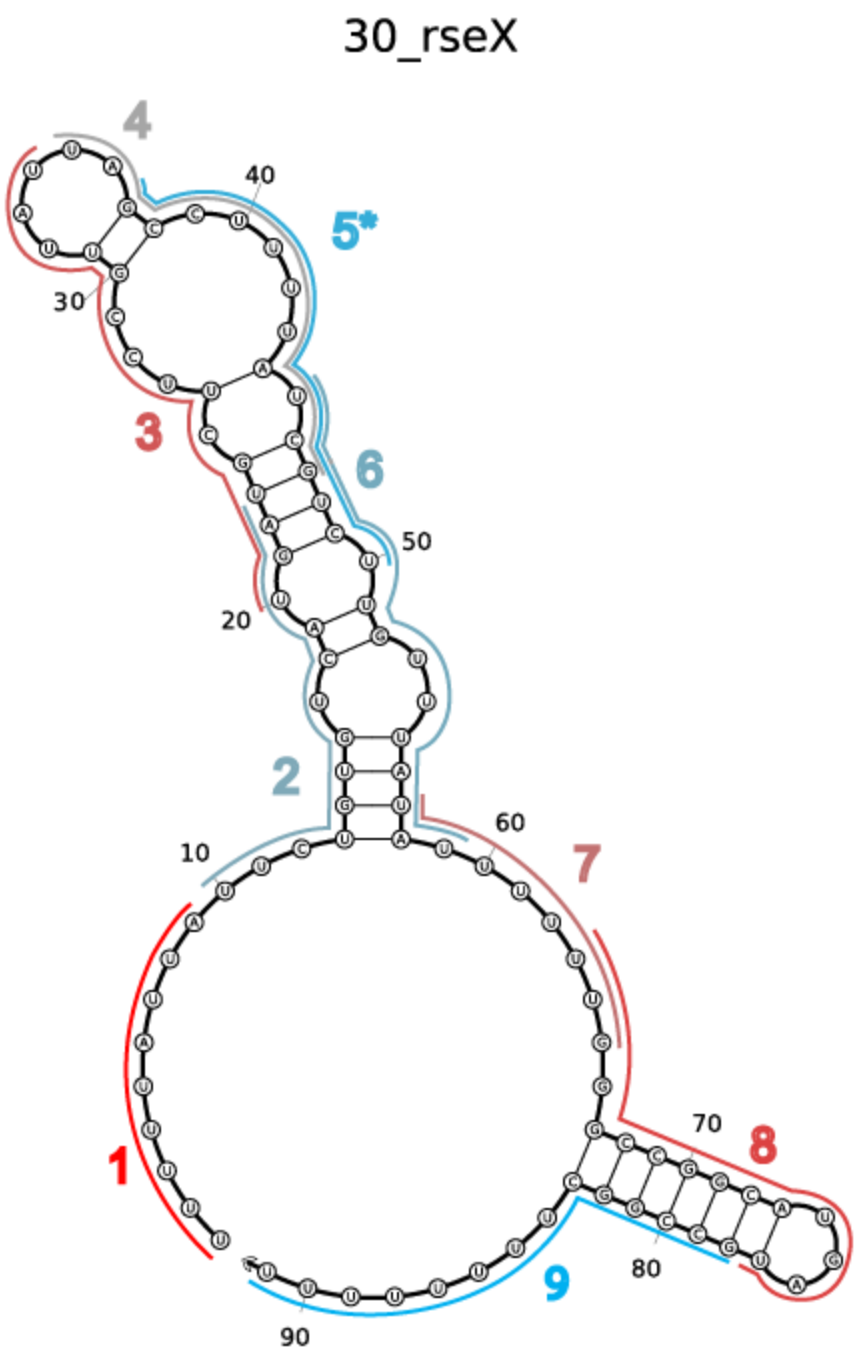

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240

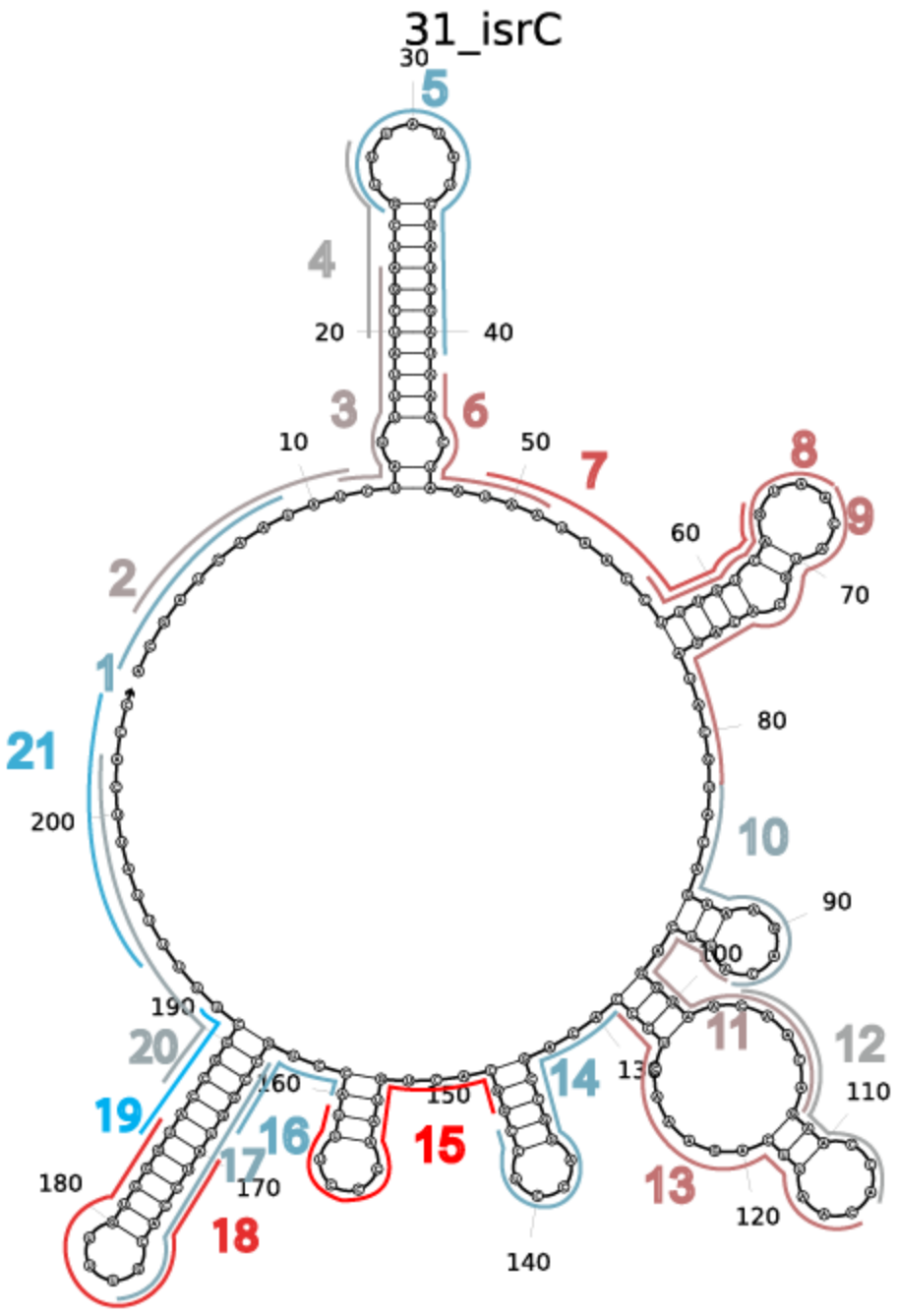

27_ryeB (sdsr)
1
2
3
4
5
6
7
8
9
10
11
10
20
30
40
50
60
70
80
90
100
110
120

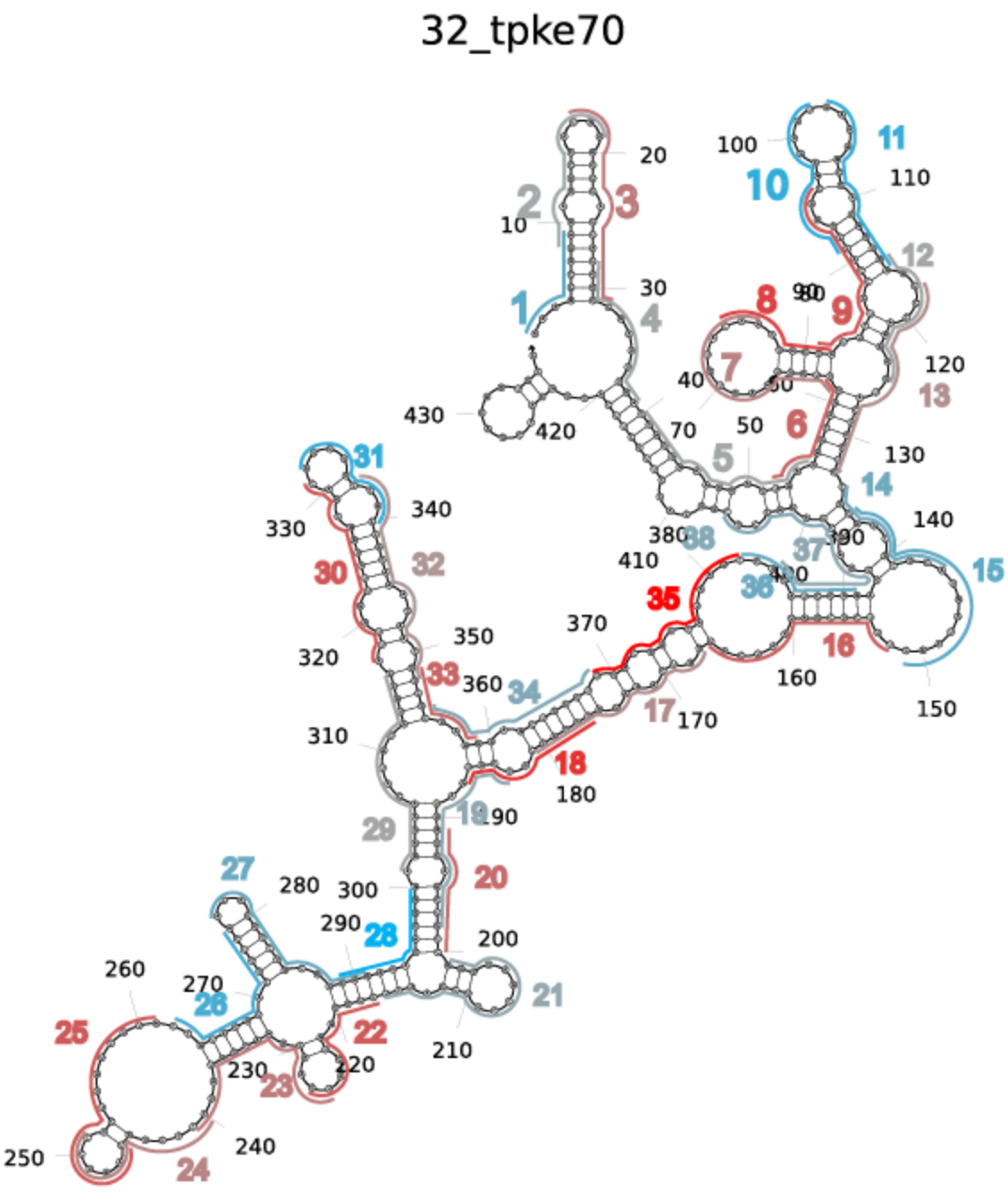

28_ryeF
4
30
60
40
20
3
5
6
7
2
10
50
390
70
1
37
8
80
90
100
36
12
9
10
380
370
140
35
17
120
11
18
190
150
130
160
360
200
180
13
110
16
14
210
34
19
350
220
15
170
33
340
20
230
32
330
21
240
31
22
320
30
23
250
24
310
260
25
29
300
290
270
28
27
280
26

29_isrB

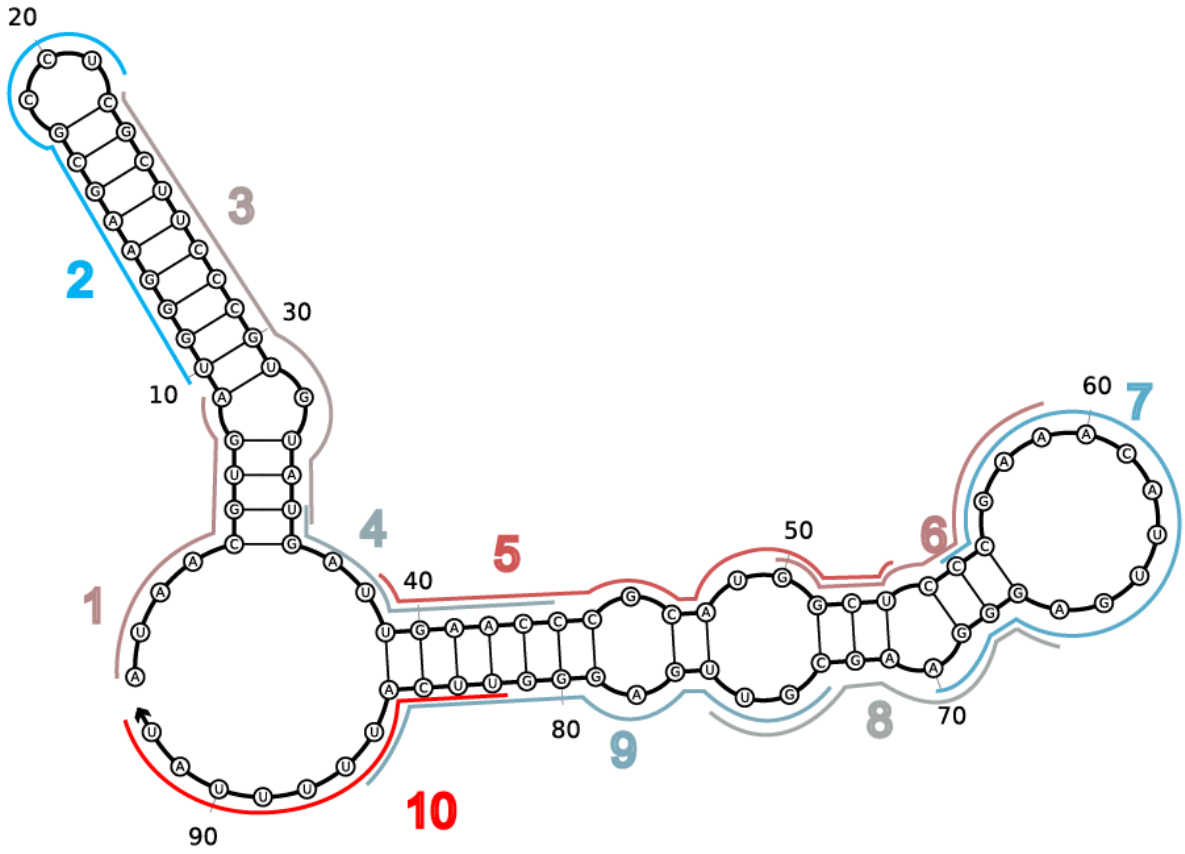

40
4
5
10
1
50
30
3
150
14
160
20
2
60
10
13
130
6
12
70
120
7
11
80
8
110
90
10
9
100


30_rseX
4
40
5*
30
3
6
50
20
2
60
10
7
70
8
1
80
9
90

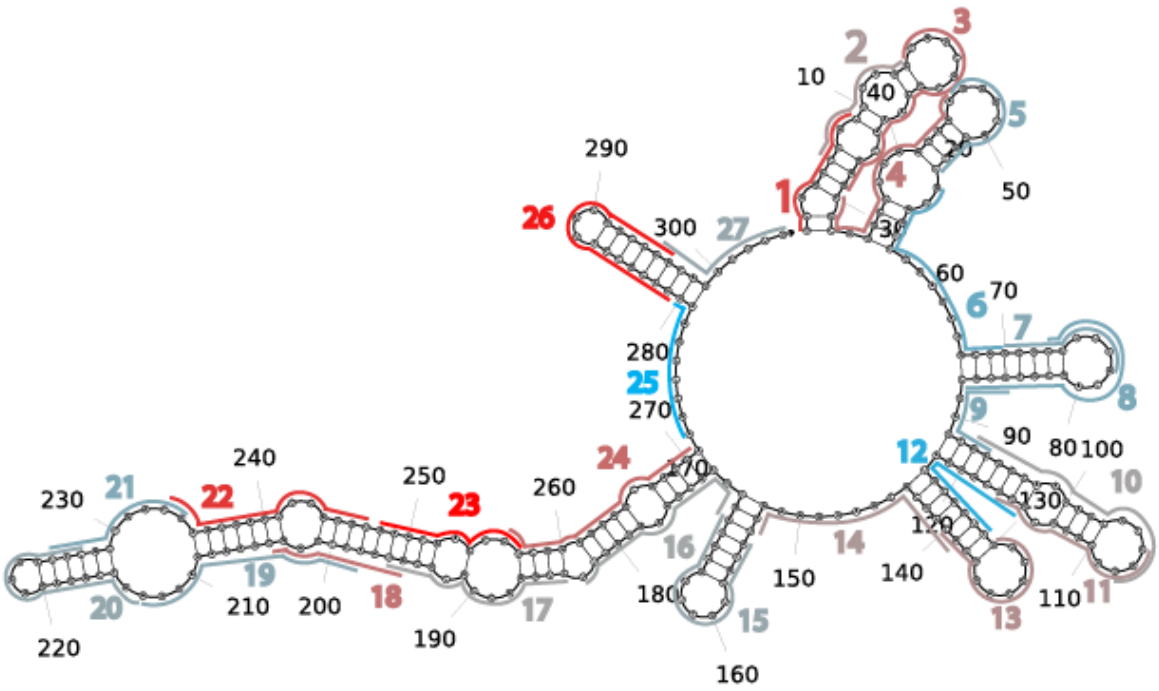

31_isrC
30
5
4
20
40
3
6
10
50
7
8
9
60
70
2
1
80
21
200
10
90
100
190
20
11
12
160
150
14
13
110
19
15
16
17
180
170
13
120
18
140


32_tpke70

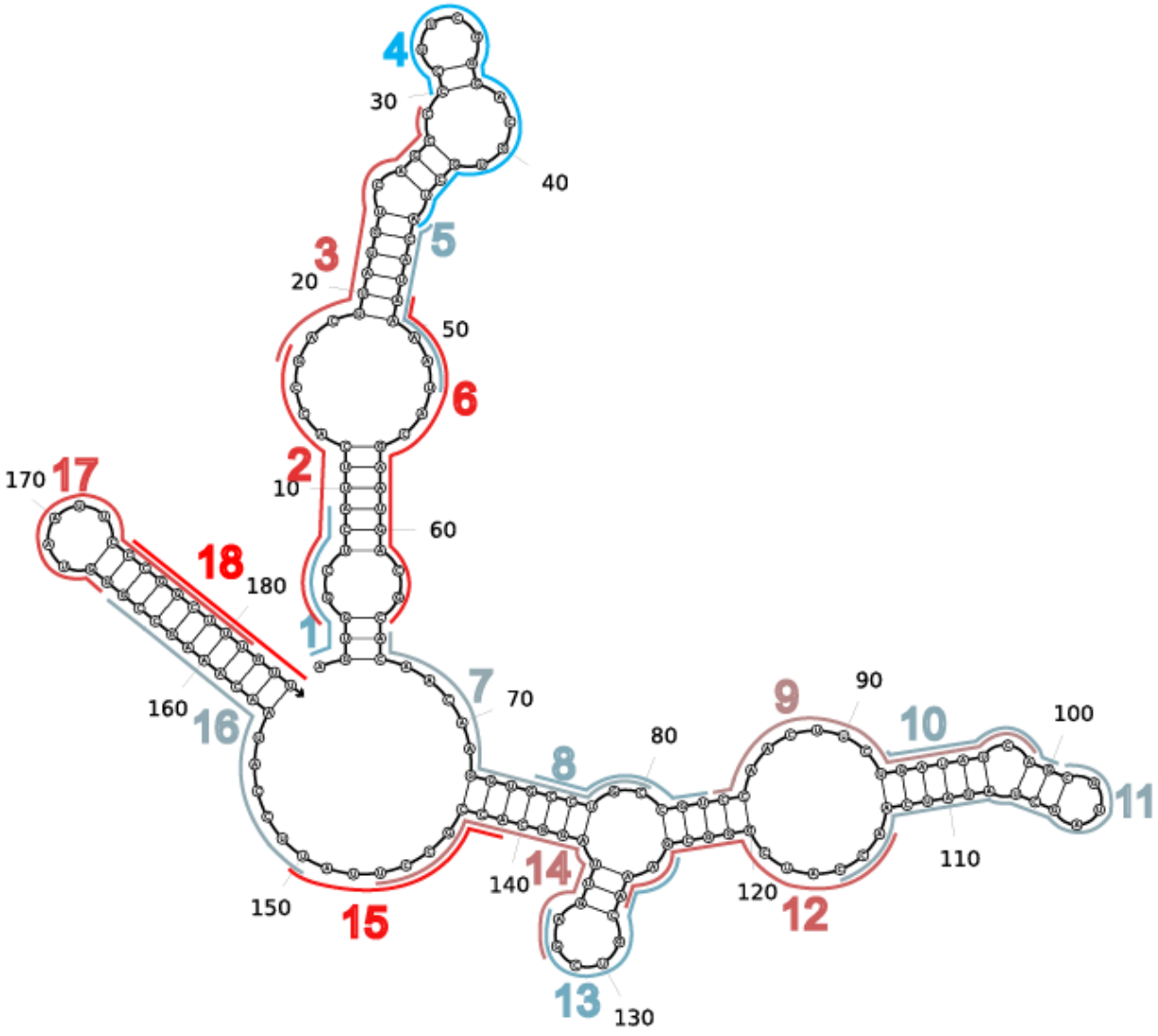
33_sroE



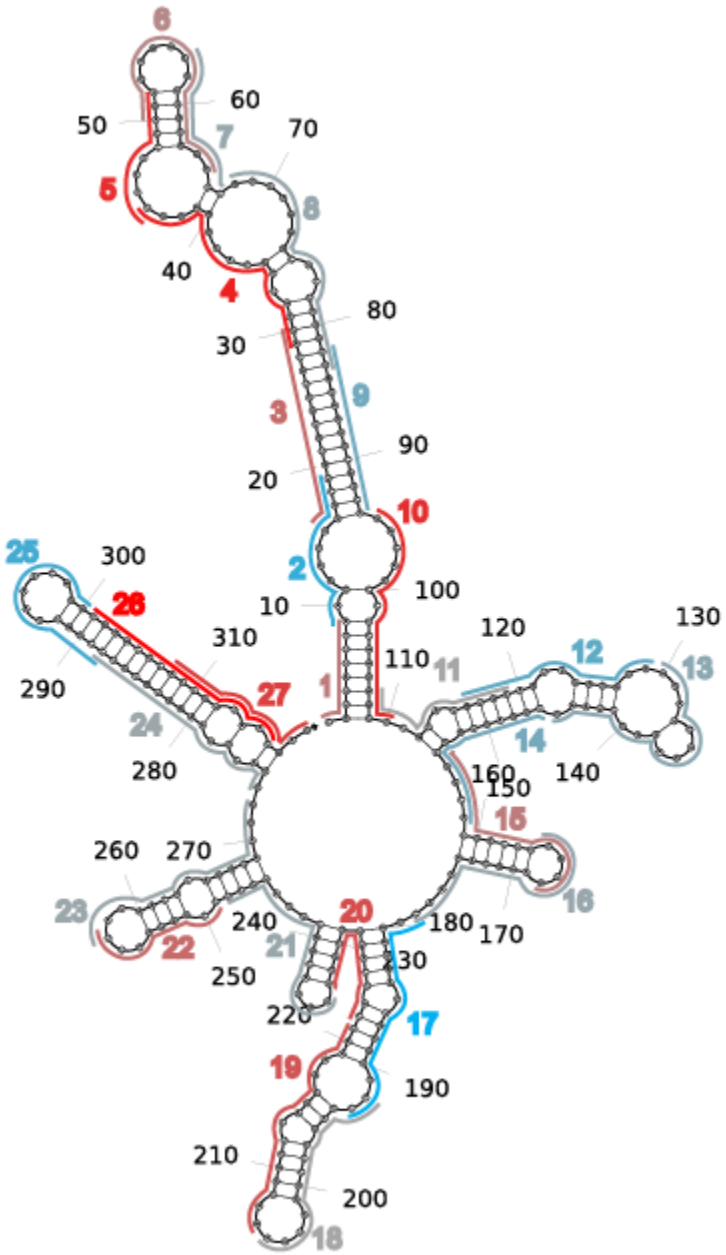
34_ryfA


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
10
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300

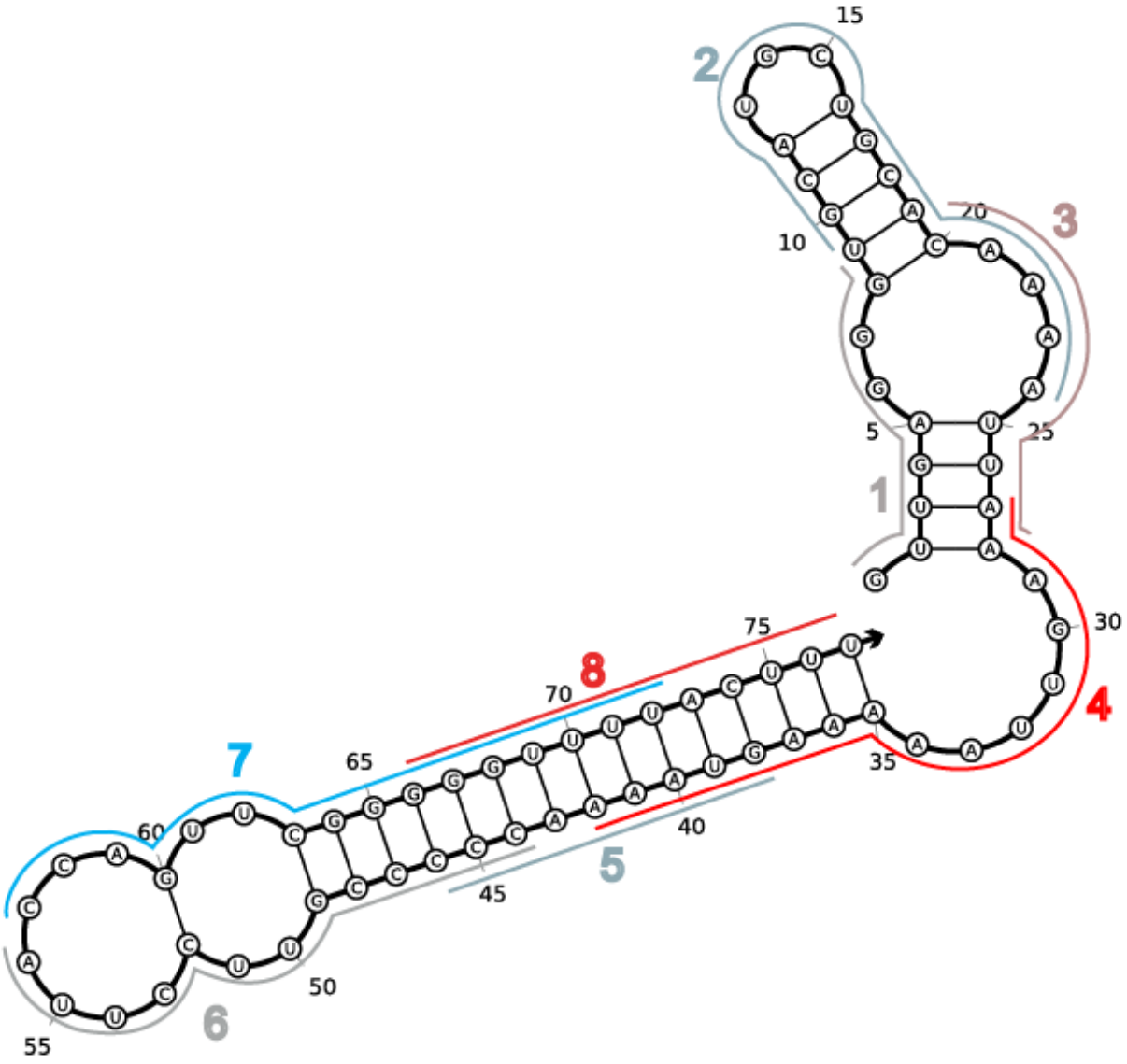
35_glmY

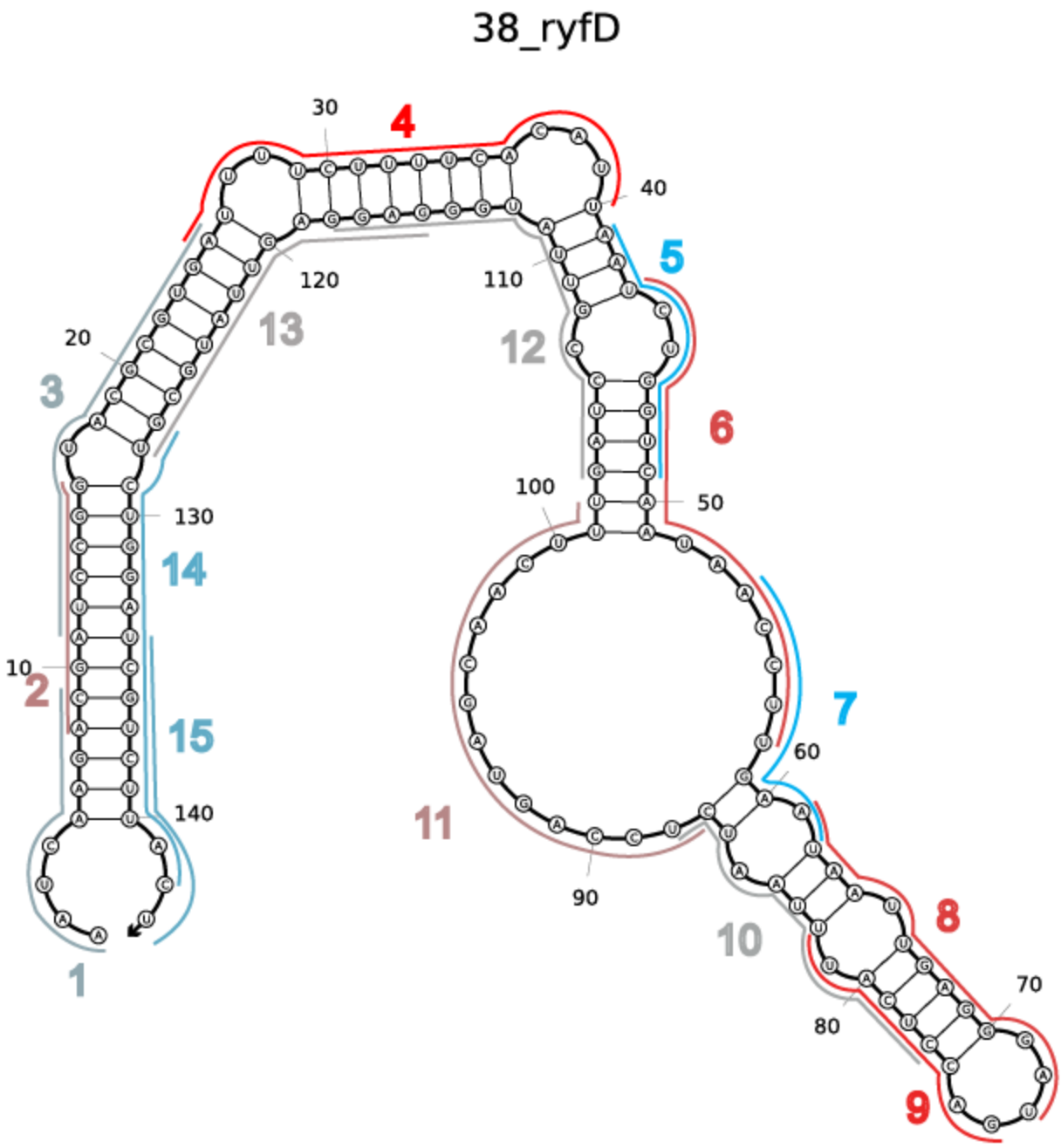

4
30
40
3
5
20
50
6
2
10
60
170
17
18
180
1
160
16
7
70
80
8
9
90
10
100
11
110
120
12
140
14
150
15
13
130

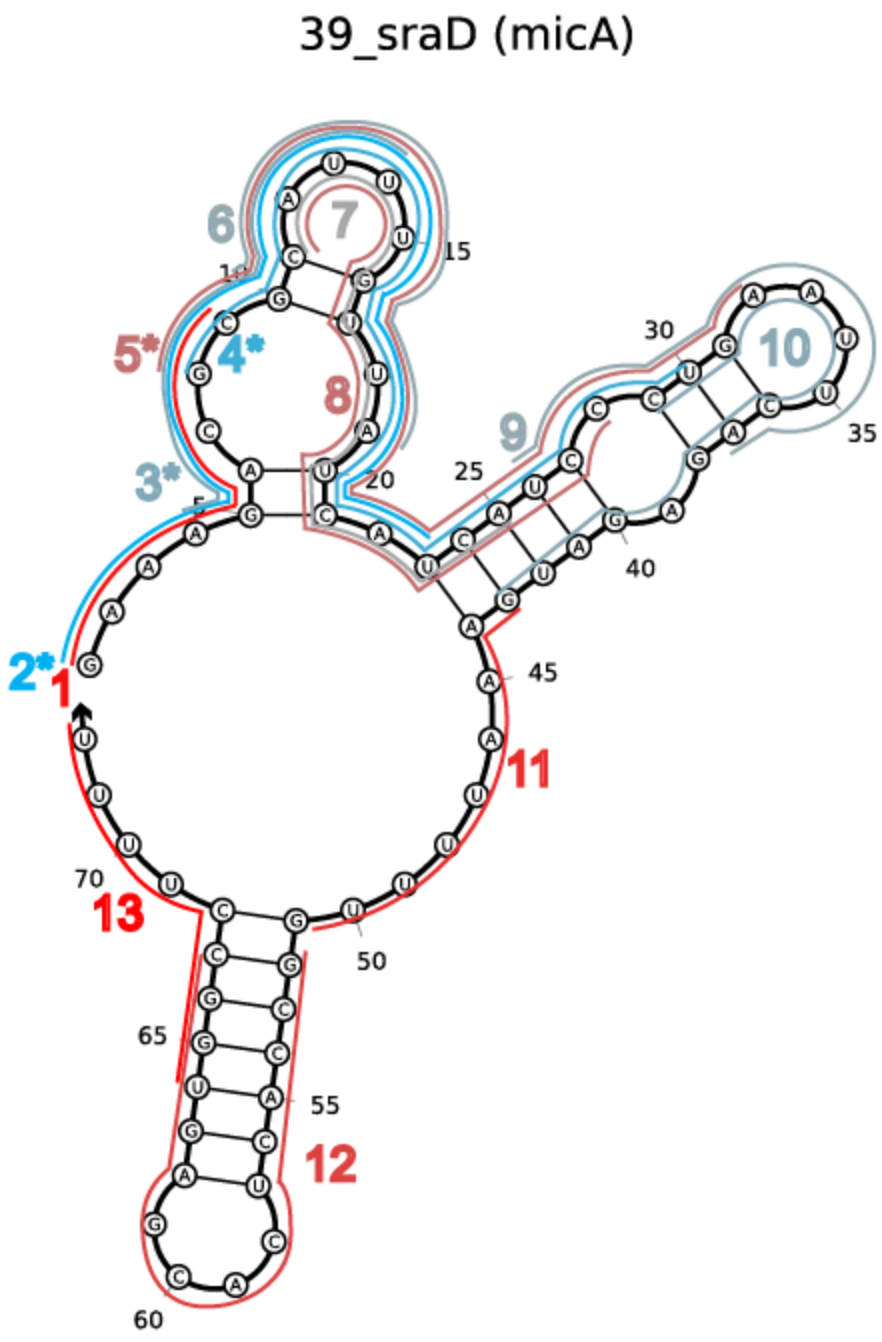
36_ryfB

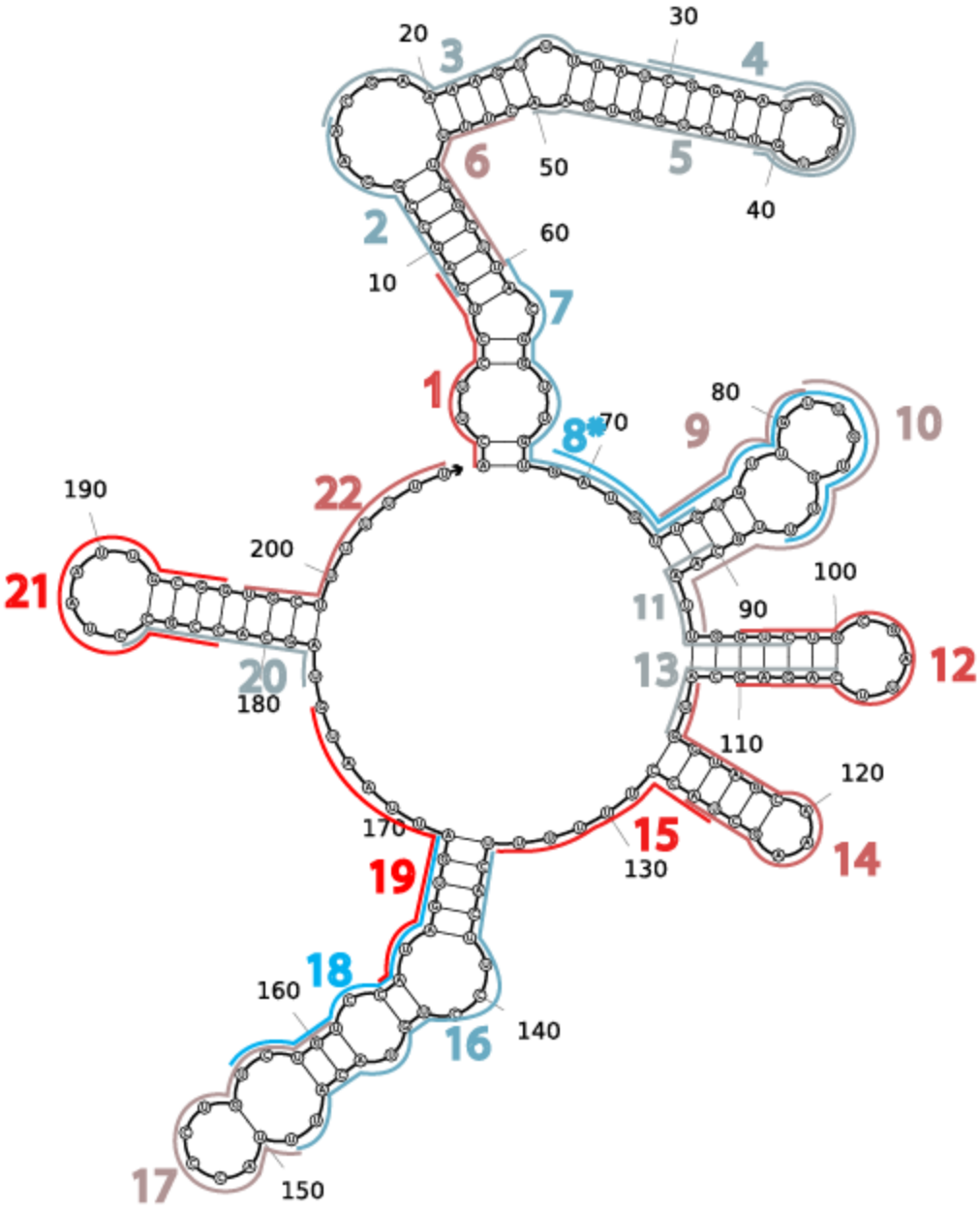

6
60
50
7
70
5
8
40
4
80
30
9
3
90
20
10
25
300
2
100
26
10
120
130
310
110
12
290
27
1
11
13
24
14
280
160
150
140
15
260
270
16
23
240
20
180
170
22
21
250
230
220
17
19
190
210
200
18

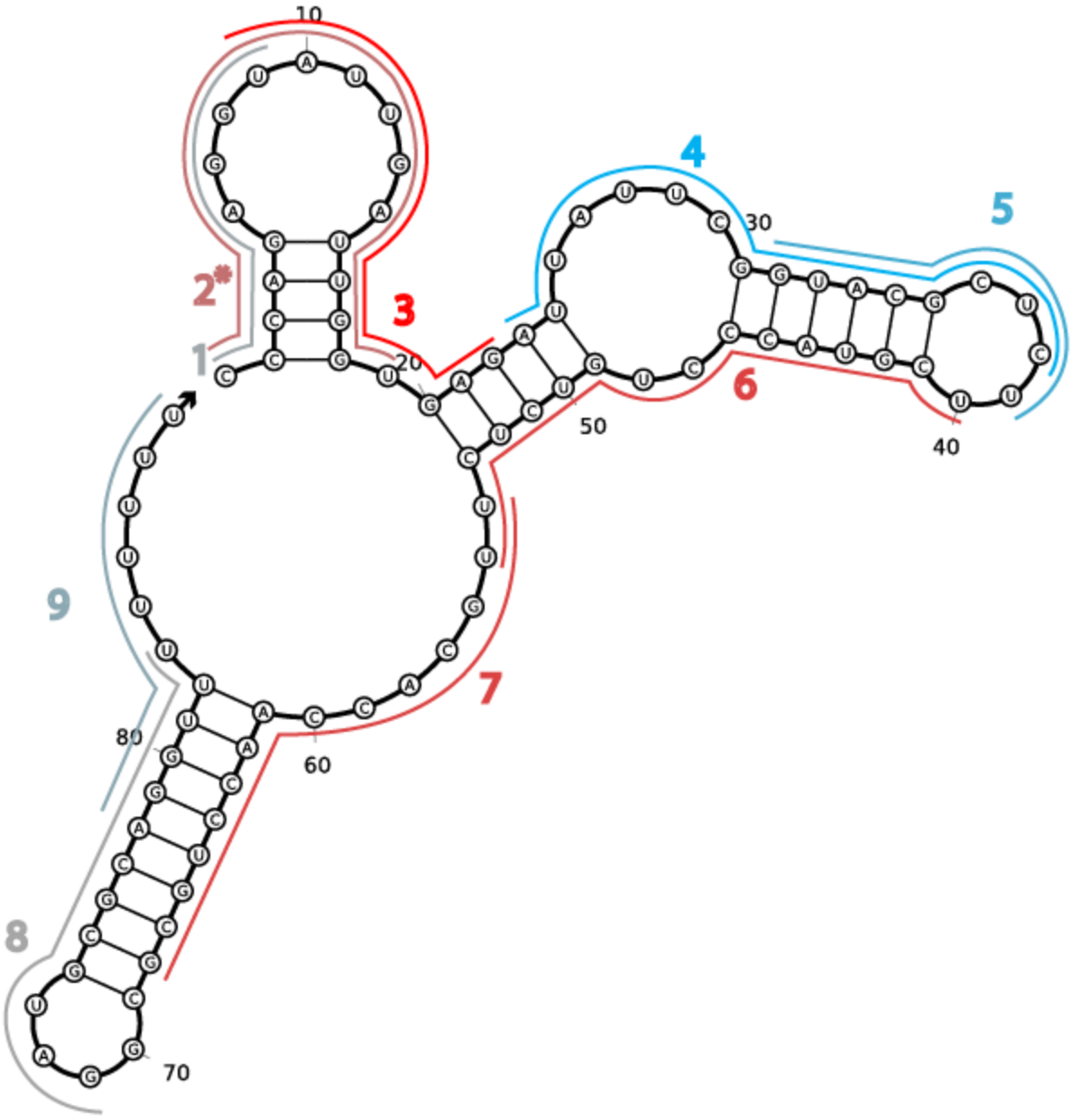
37_ryfC (ohsC)

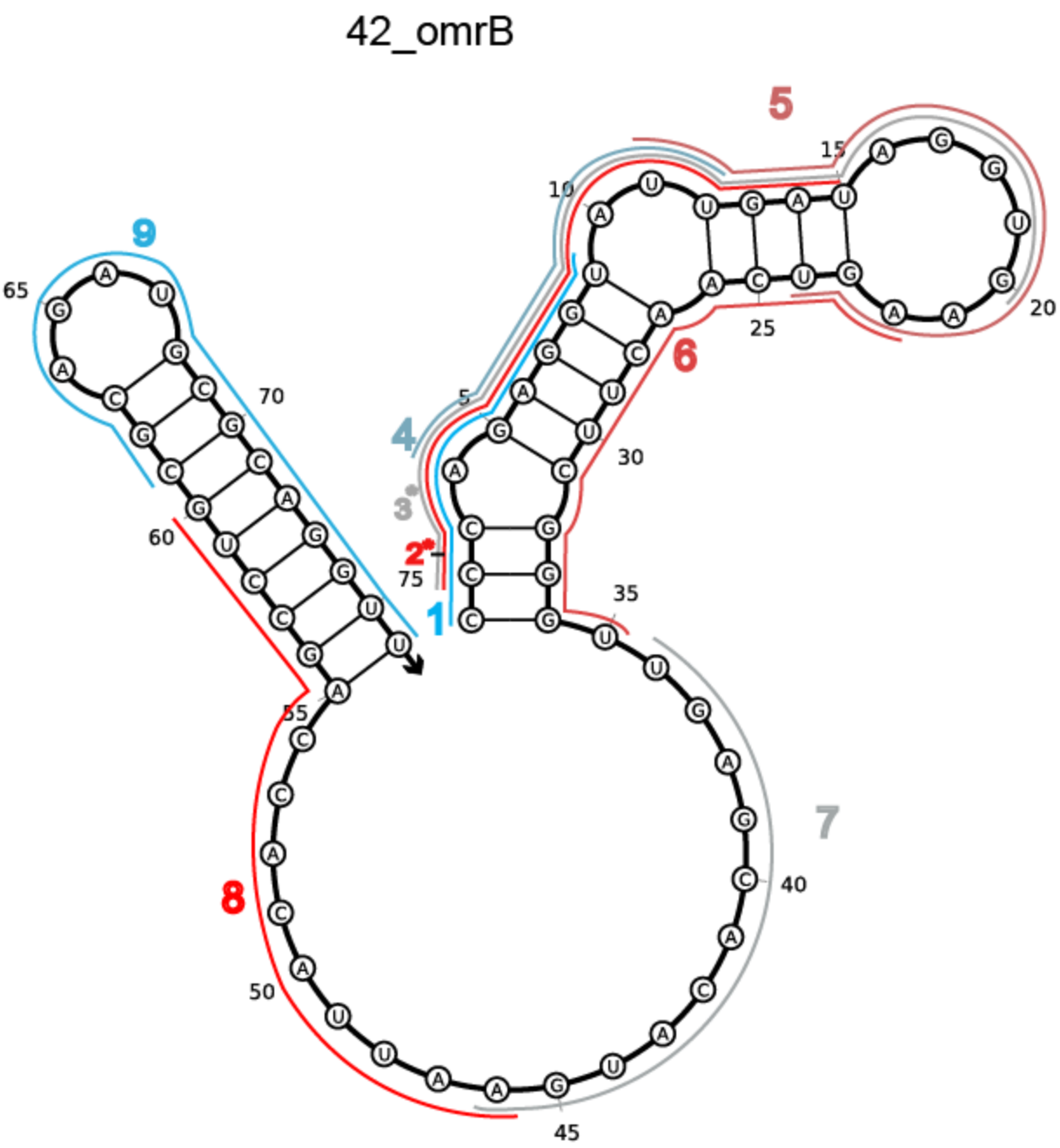


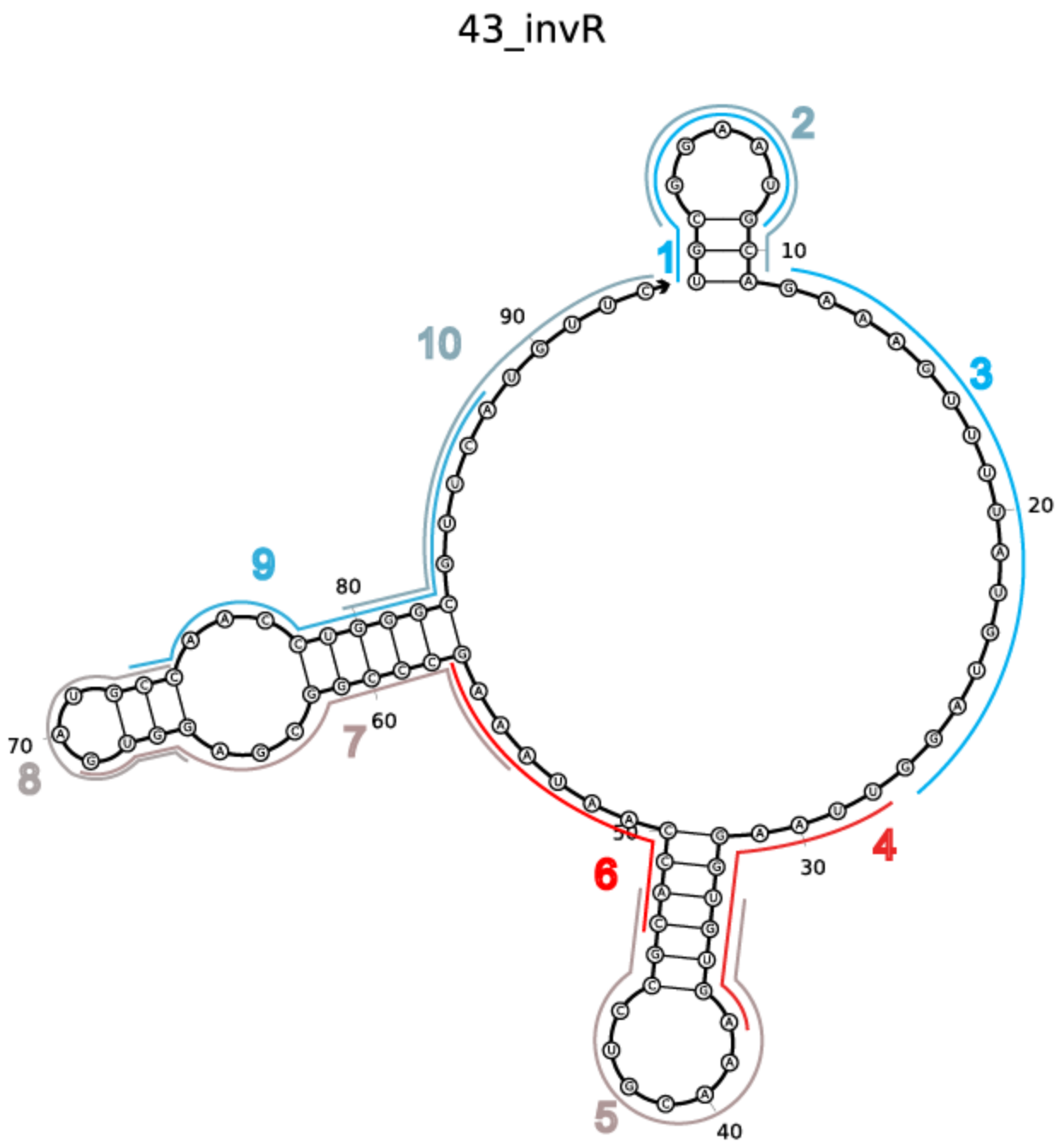

38_ryfD
30
4
40
5
120
110
13
12
20
3
6
100
50
130
14
10
2
7
15
60
11
140
90
10
8
70
80
1
9

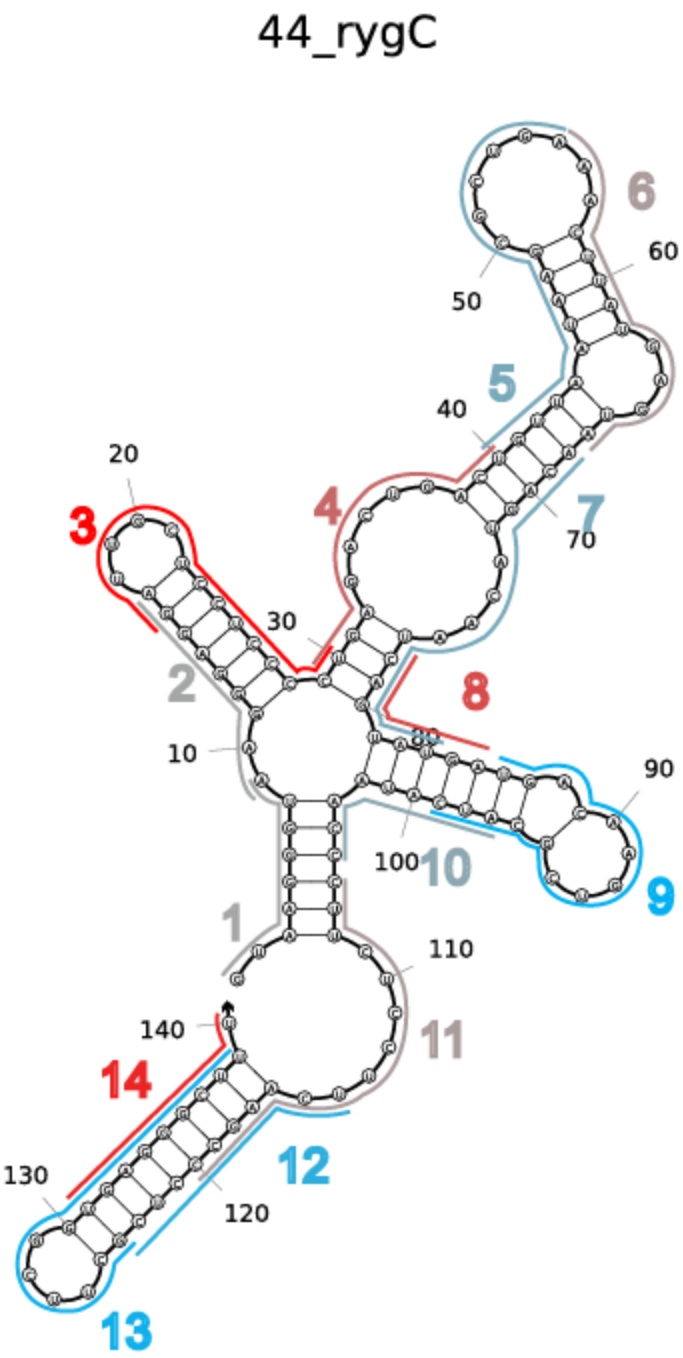

39_sraD (micA)
6
7
15
5*
4*
8
30
10
35
9
3*
20
25
40
2*
1
45
11
70
13
50
65
55
12
60


40 gcvB
20
30
3
4
6
50
5
40
2
60
10
7
1
80
8*
70
9
10
190
22
200
100
21
11
90
12
20
13
180
110
120
170
15
14
19
130
18
160
140
16
17
150

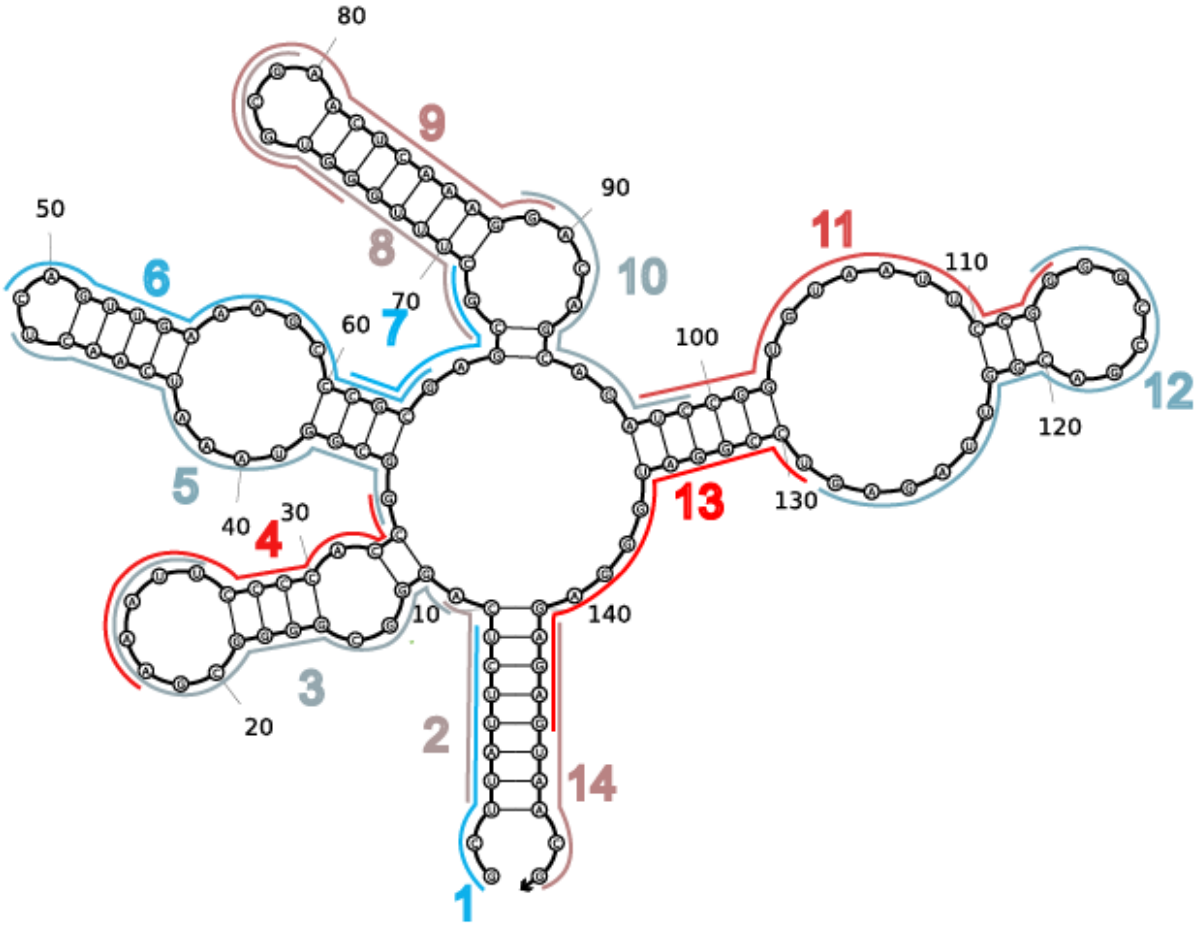

41_omrA
10
20
30
40
50
60
70
80
1
2*
3
4
5
6
7
8
9

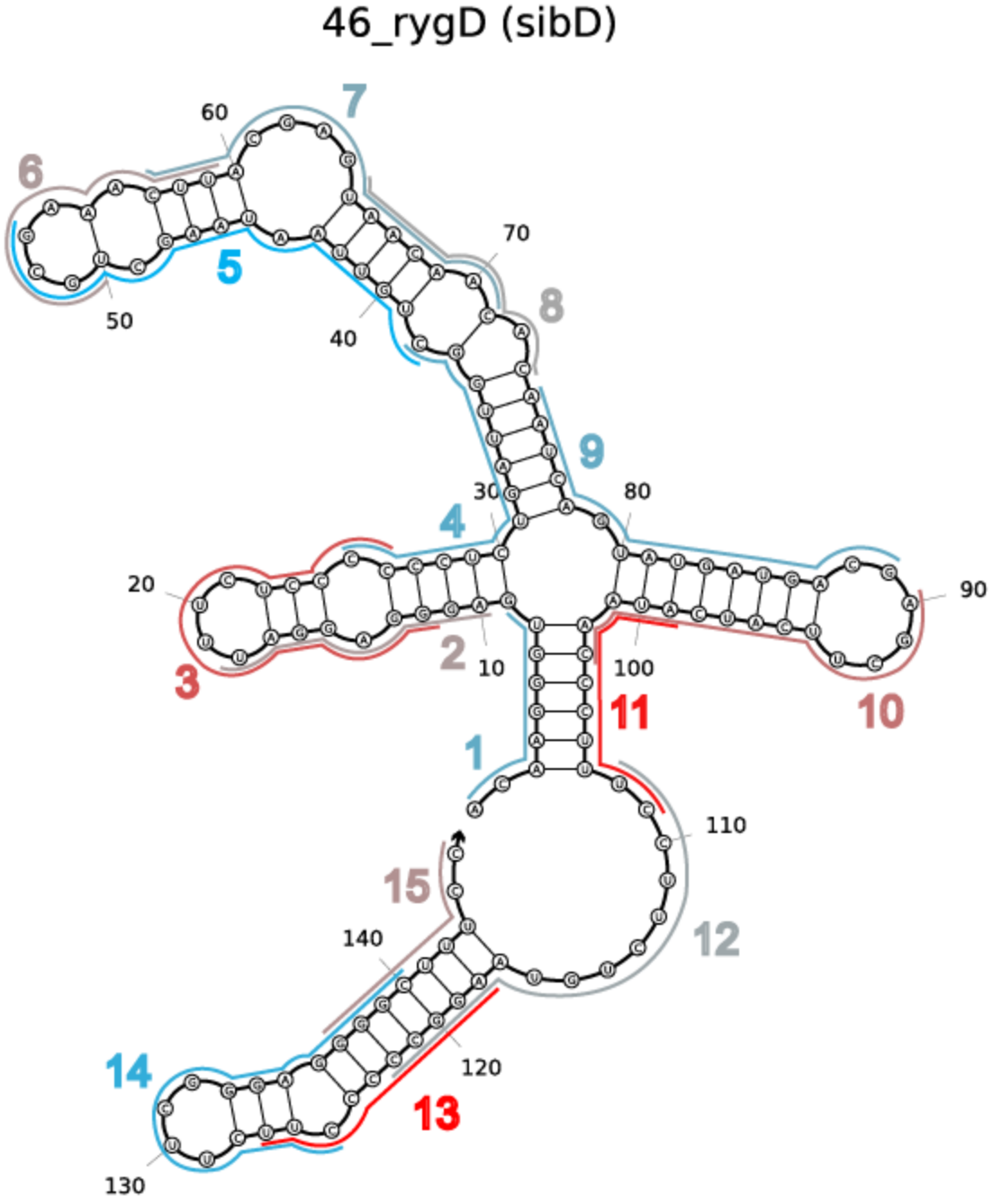

42_omrB
5
15
10
9
65
70
4
5
6
25
20
30
3
60
2
75
1
35
55
7
8
40
50
45

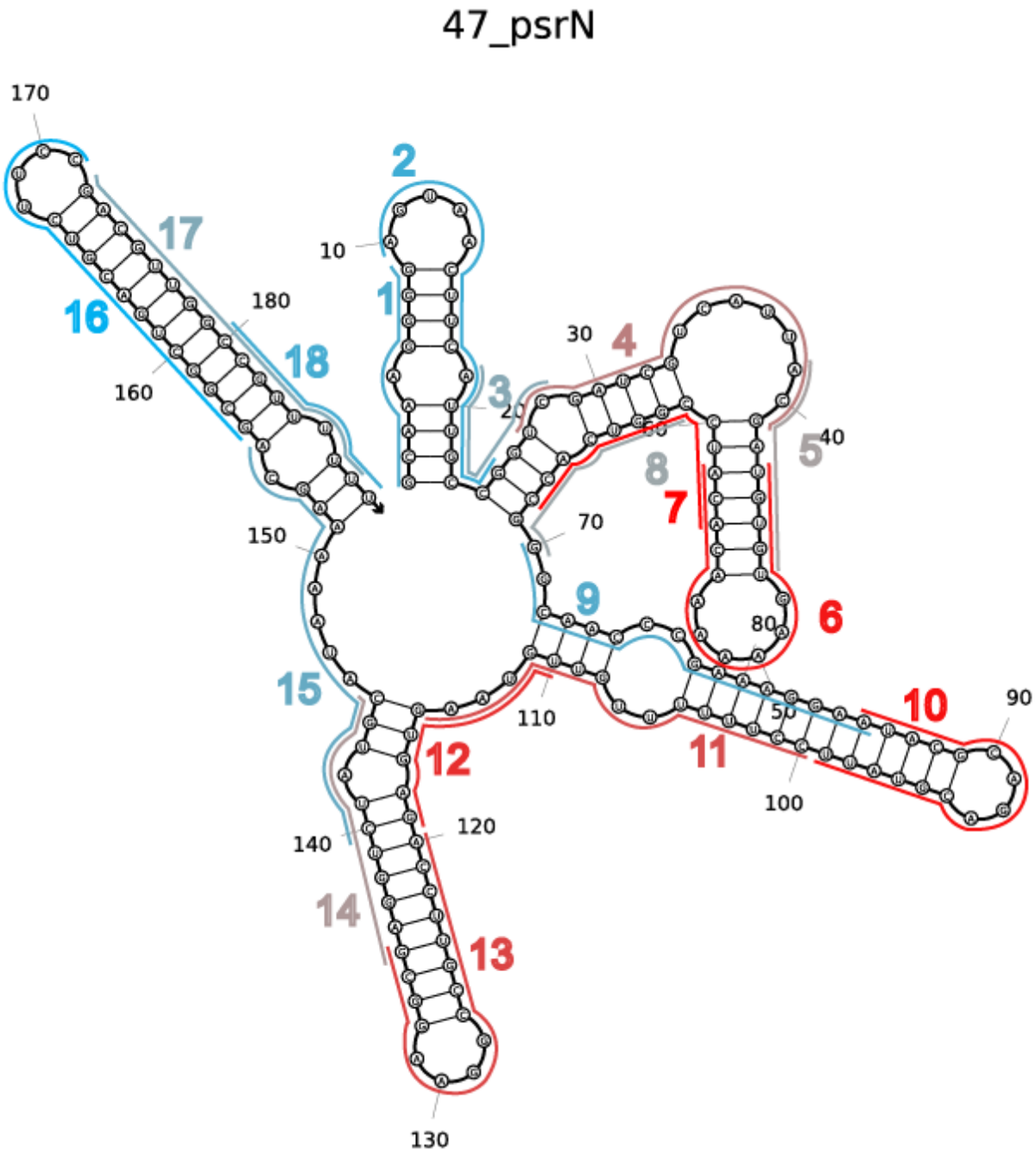

43_invR
1
2
3
4
5
6
7
8
9
10
10
20
30
40
50
60
70
80
90

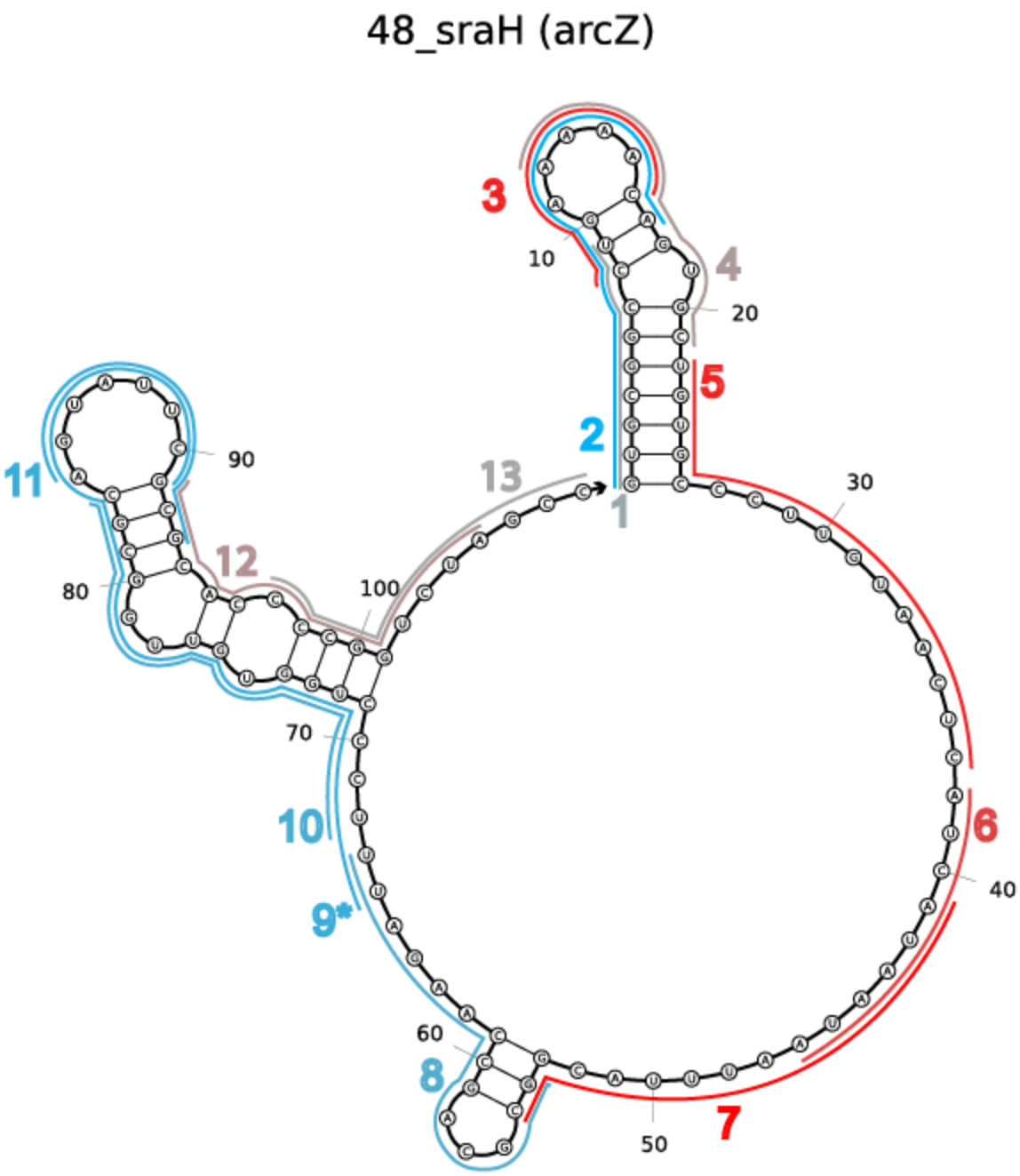

44_rygC
6
60
50
5
40
20
4
7
3
70
30
2
8
10
90
100
10
9
1
110
140
11
14
12
130
120
13

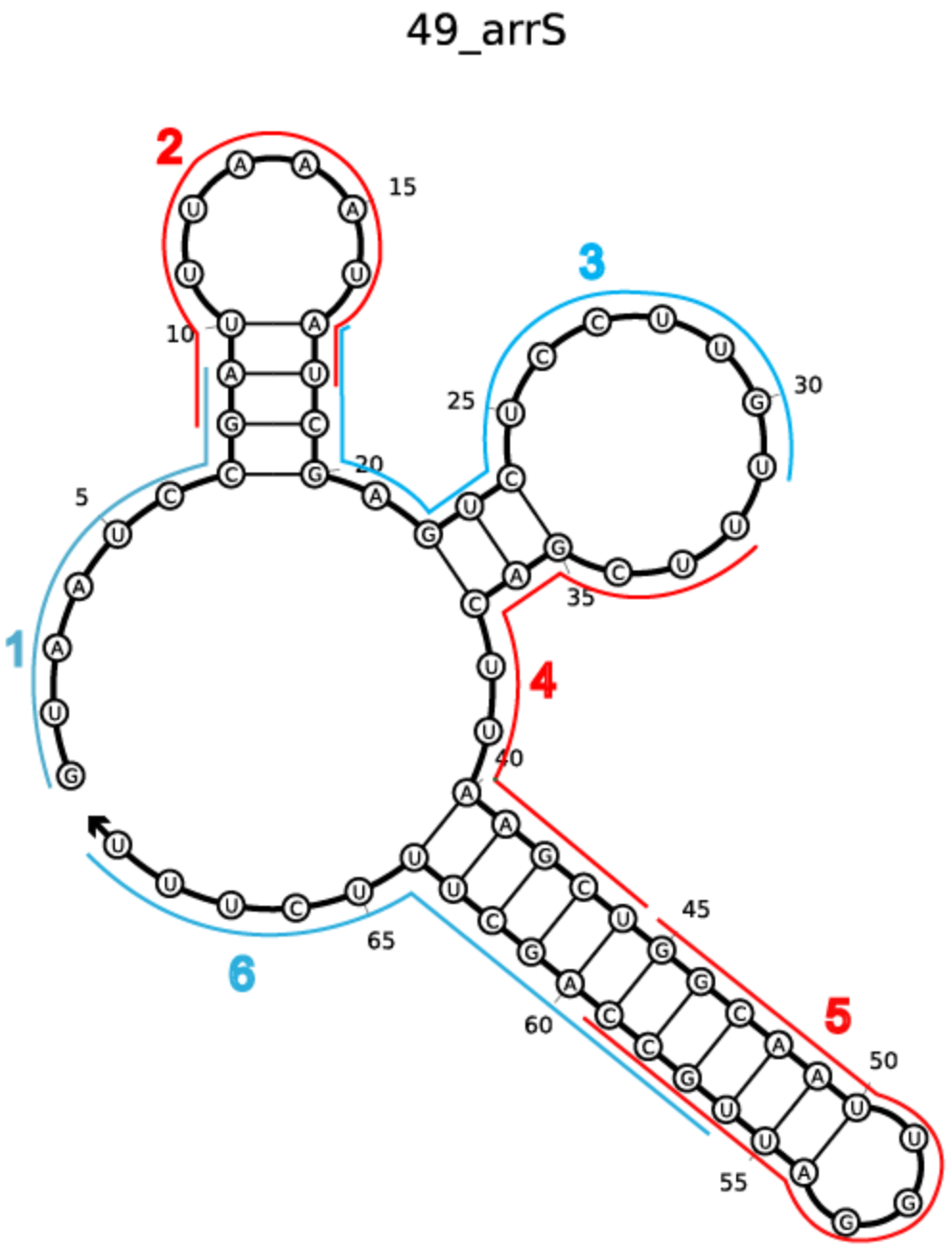
45_sroG

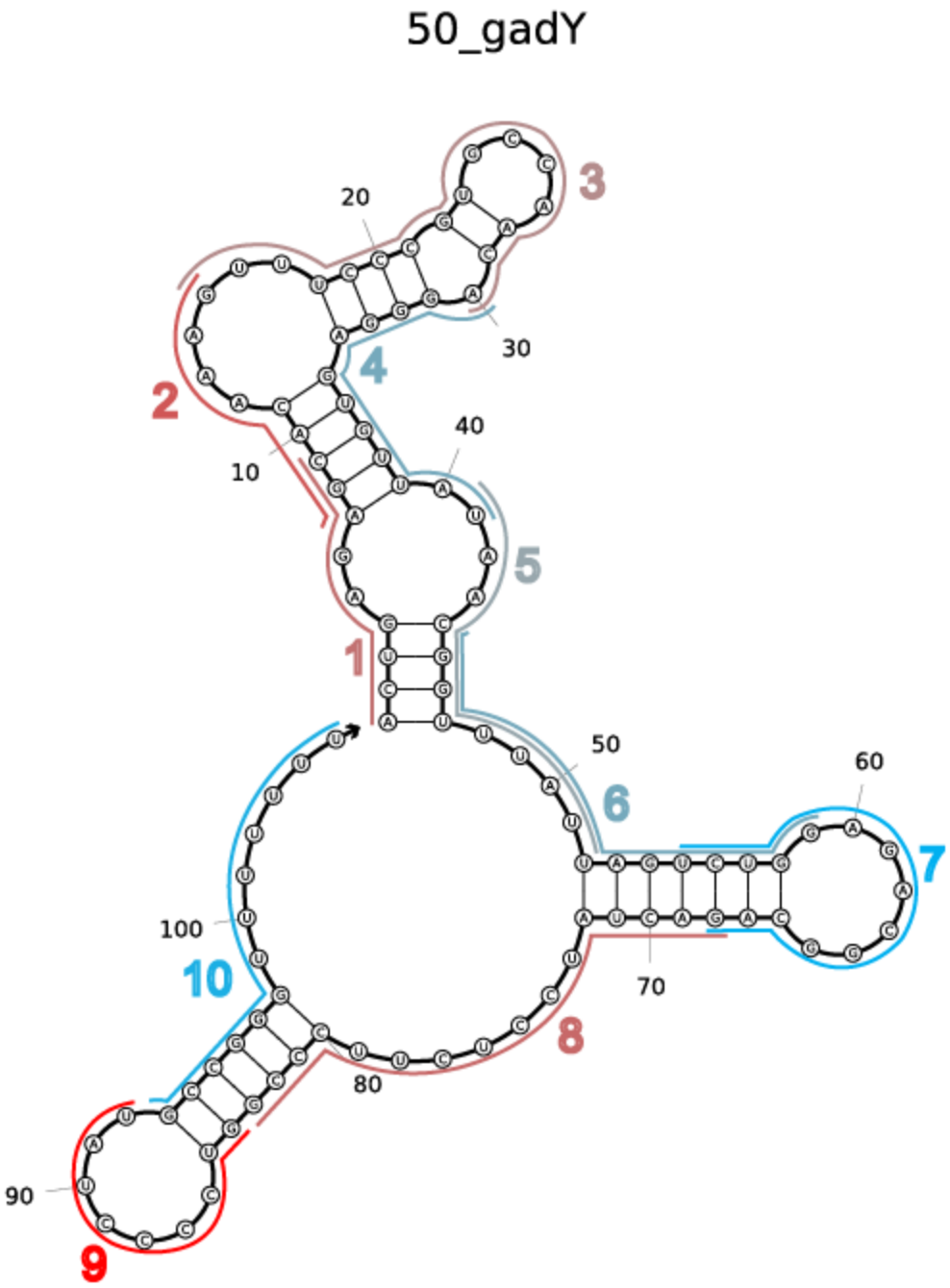

80
9
90
50
8
11
6
10
110
70
7
60
100
12
120
5
13
130
40
30
4
10
140
3
20
2
14
1

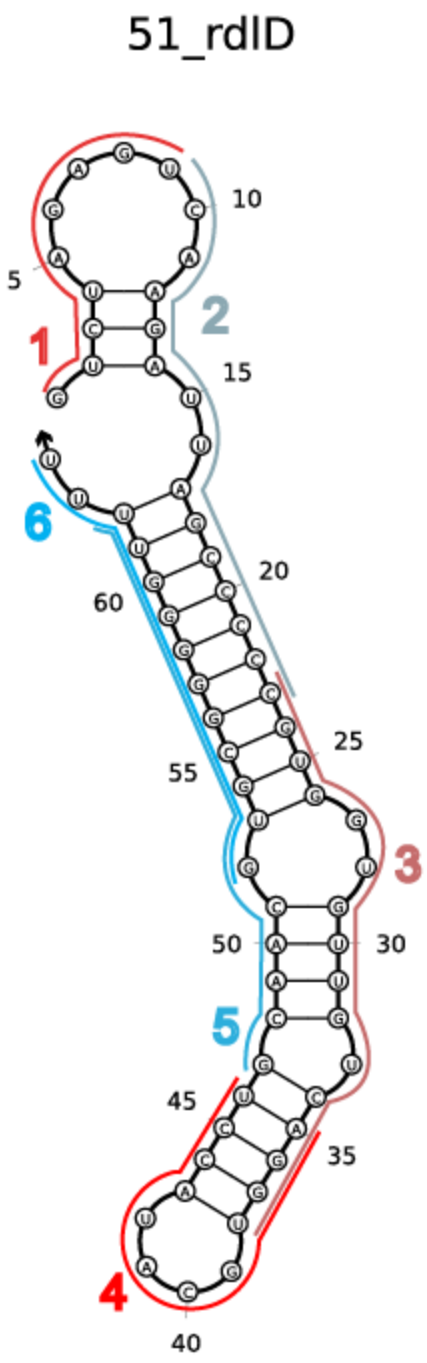

46_rygD (sibD)
7
60
6
70
5
8
50
40
9
30
80
4
20
90
2
10
100
3
11
10
1
110
15
12
140
14
120
13
130

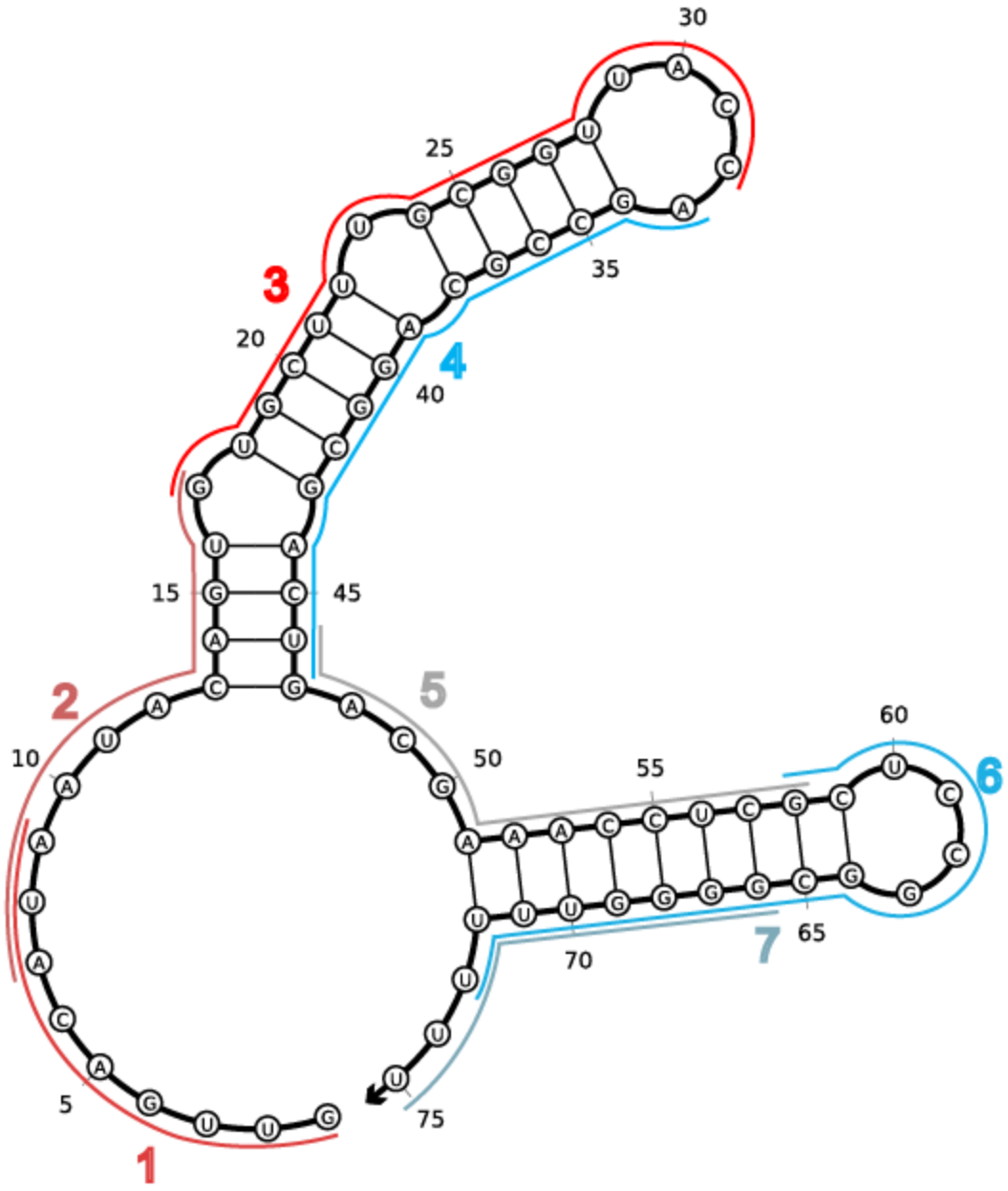

47_psrN
170
2
17
10
1
16
180
18
160
30
4
3
20
40
5
8
7
70
150
9
6
80
15
110
10
90
12
11
50
100
140
120
14
13
130

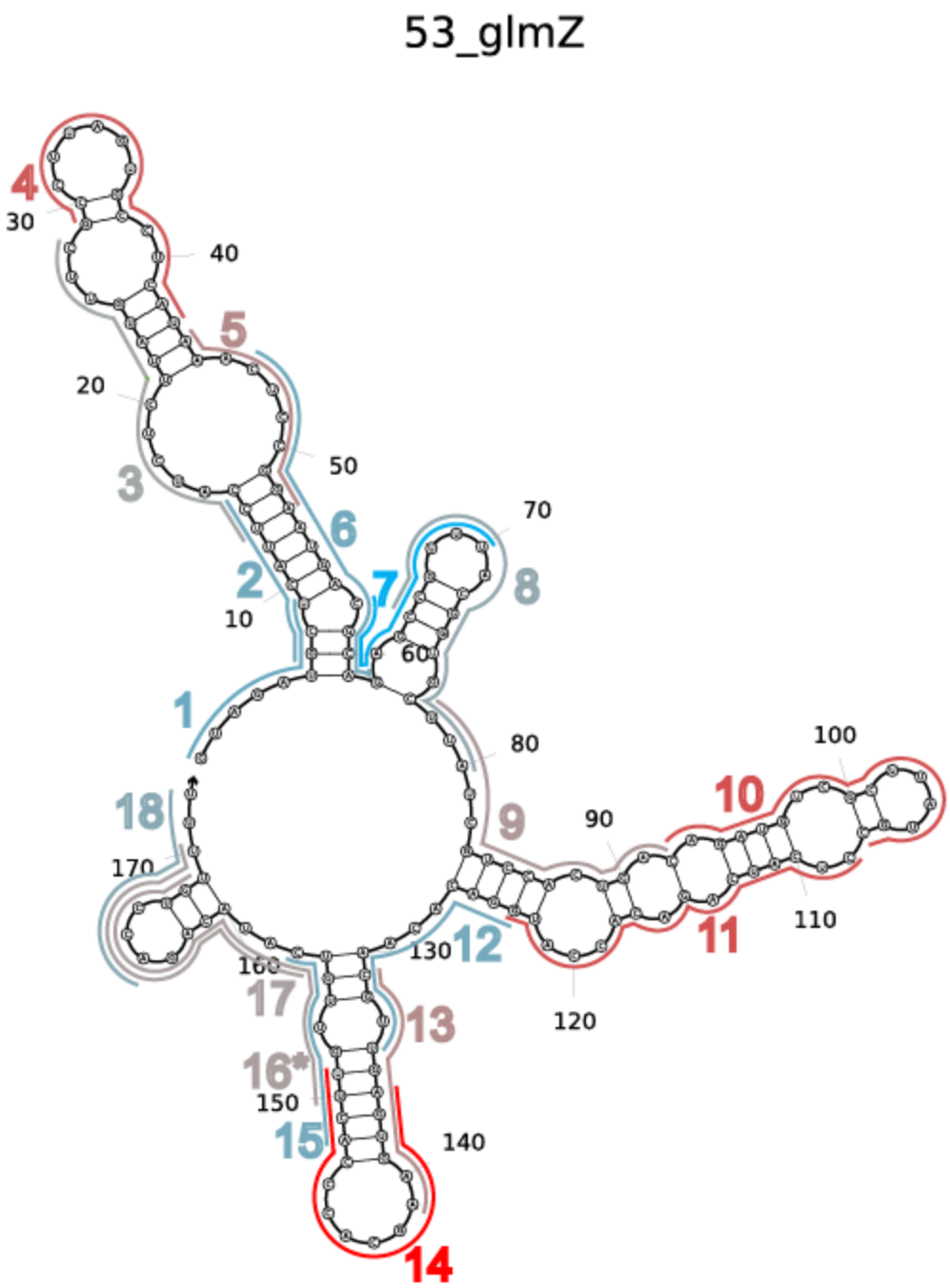

48_sraH (arcZ)
3
10
4
20
5
2
13
1
30
11
90
12
80
100
70
10
6
40
9*
60
8
50
7


49_arrS
1
2
3
4
5
6
5
10
15
20
25
30
35
40
45
50
55
60
65

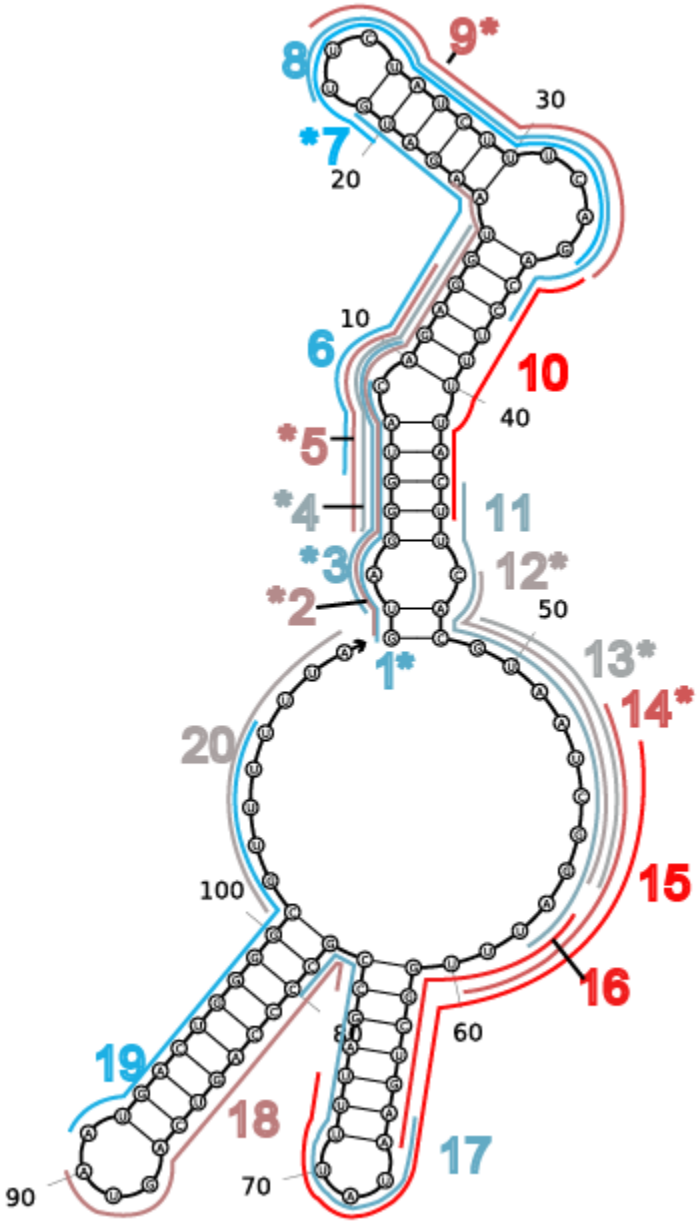

50_gadY
20
3
30
4
2
40
10
5
1
50
60
6
7
100
10
70
8
80
90
9

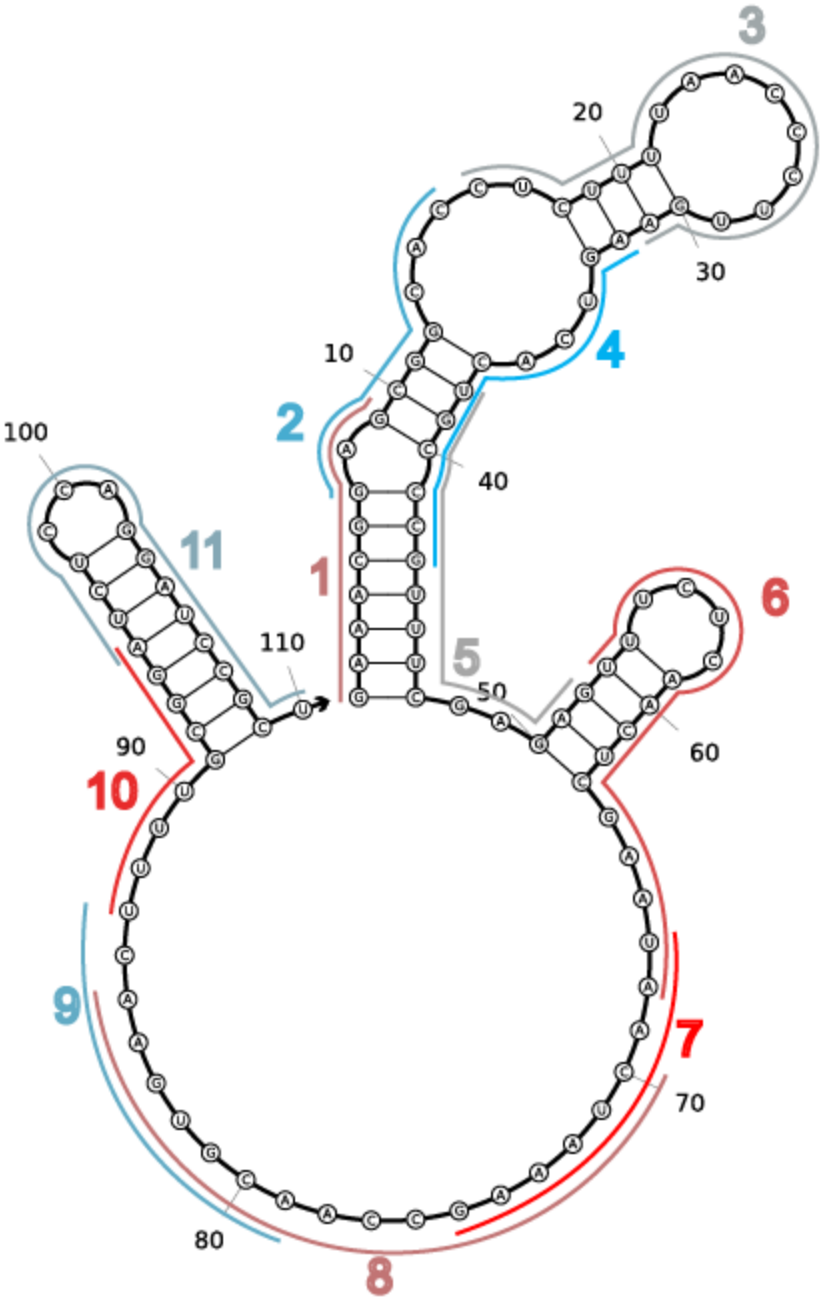

51_rdlD
10
5
2
1
15
6
20
60
25
55
3
50
30
5
45
35
4
40

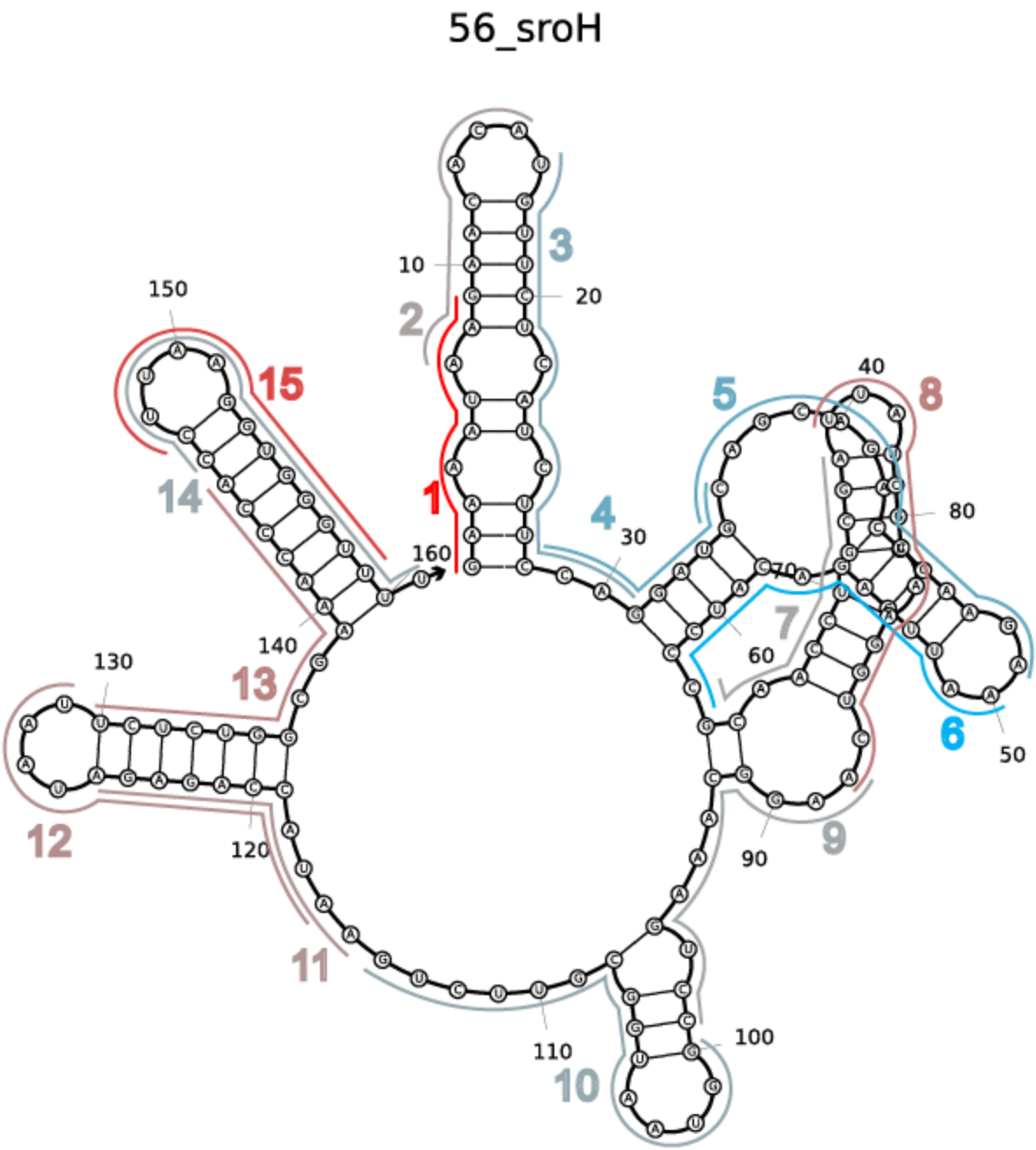

52_istR-1, istR-2
1
2
3
4
5
6
7
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75

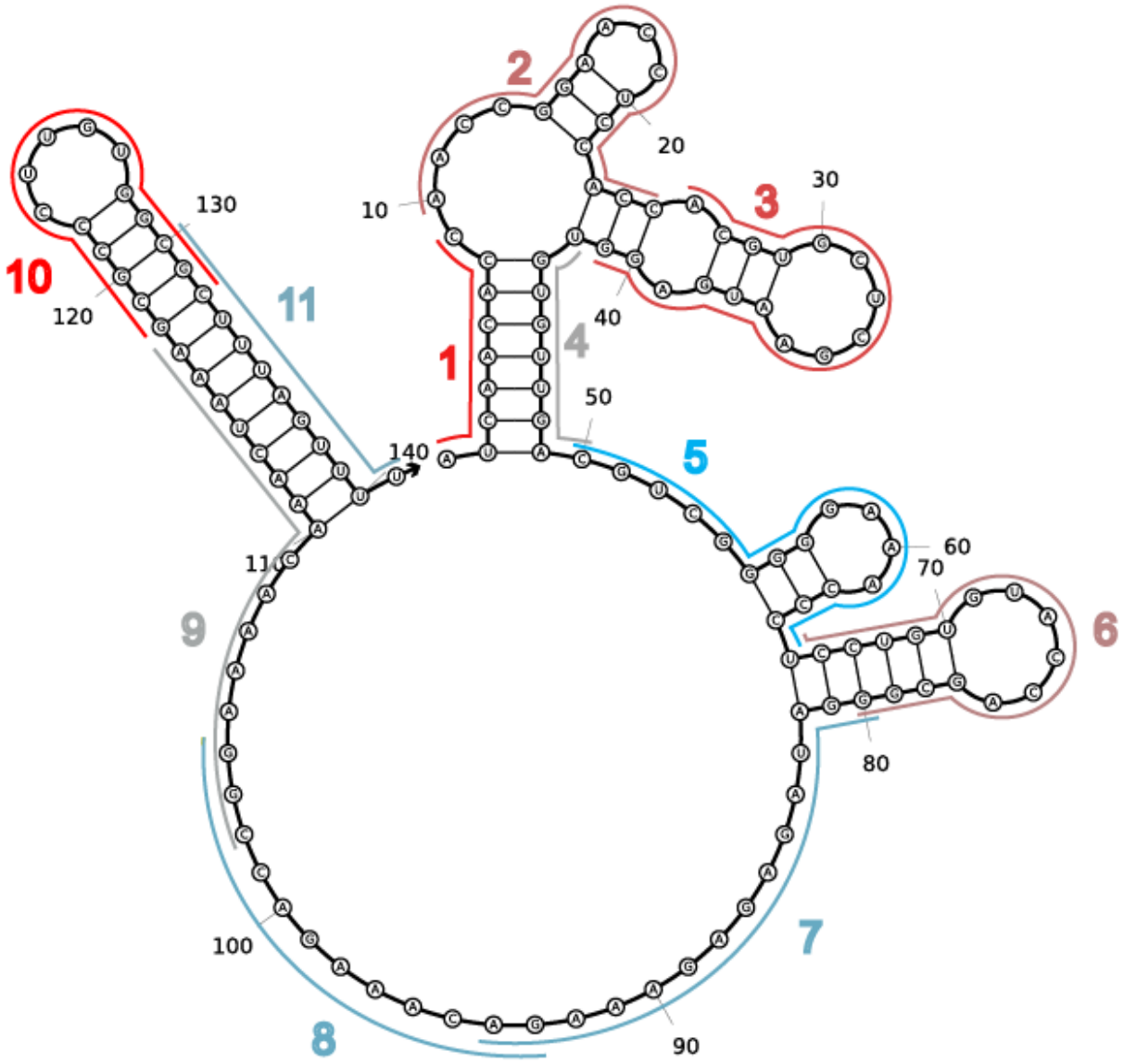

53_glmZ
4
30
40
5
20
3
50
6
70
2
7
8
10
60
1
80
100
18
9
10
90
170
110
11
130
12
17
13
120
16*
150
15
140
14

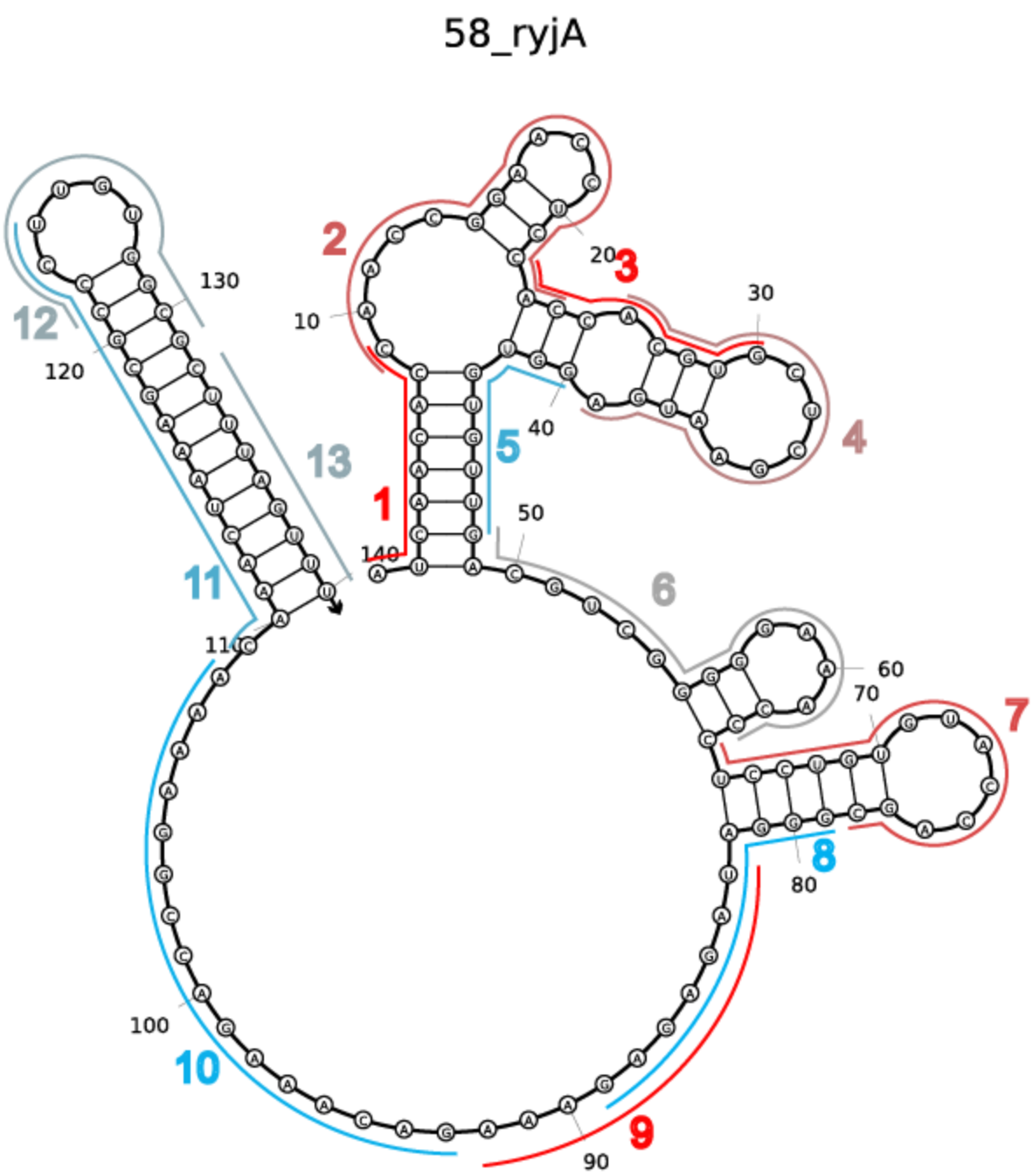
54_Spot42


8
9*
30
*7
20
10
6
10
40
*5
*4
11
*3
12*
*2
50
1*
13*
14*
20
100
15
16
60
19
18
17
90
70

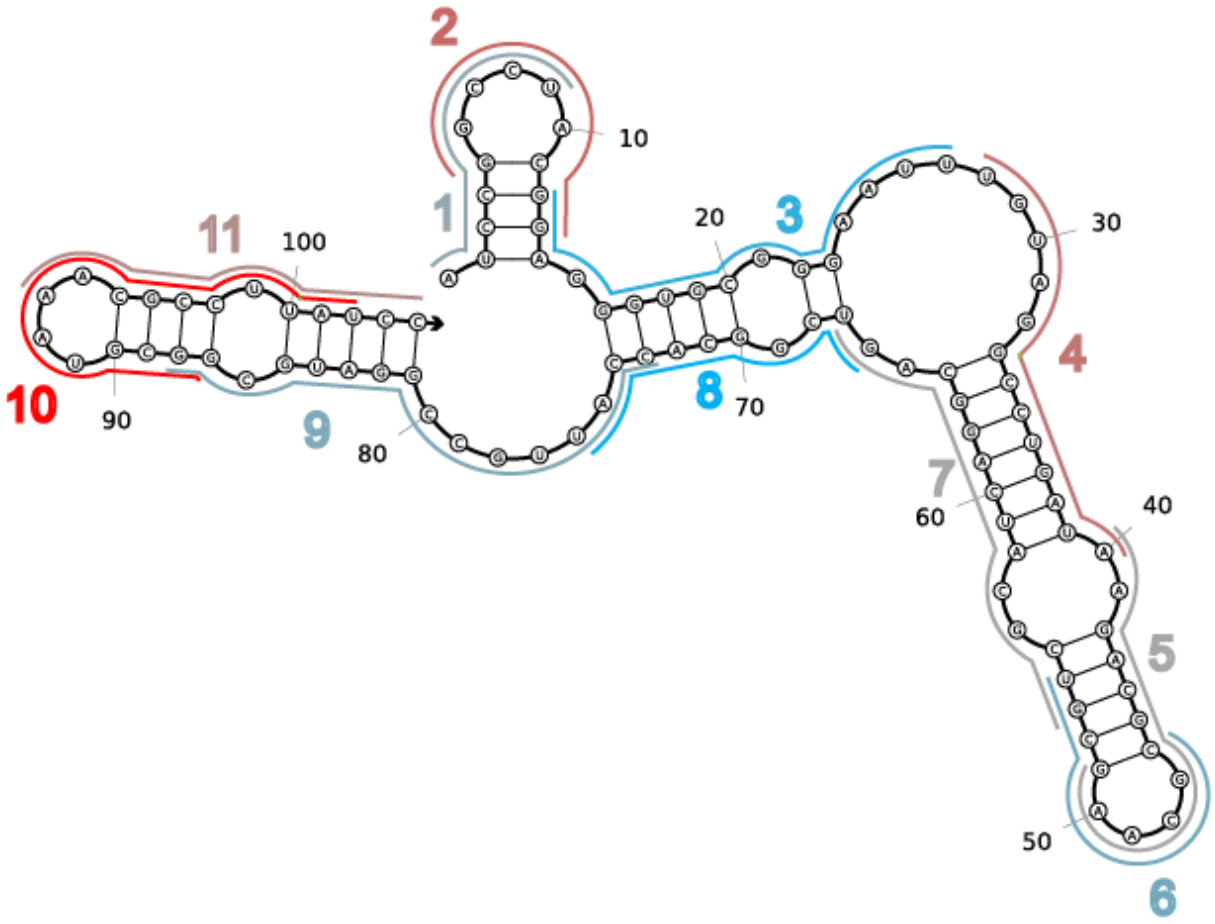

55_oxyS
3
20
30
4
10
2
40
100
11
1
6
110
5
50
60
90
10
9
7
70
80
8

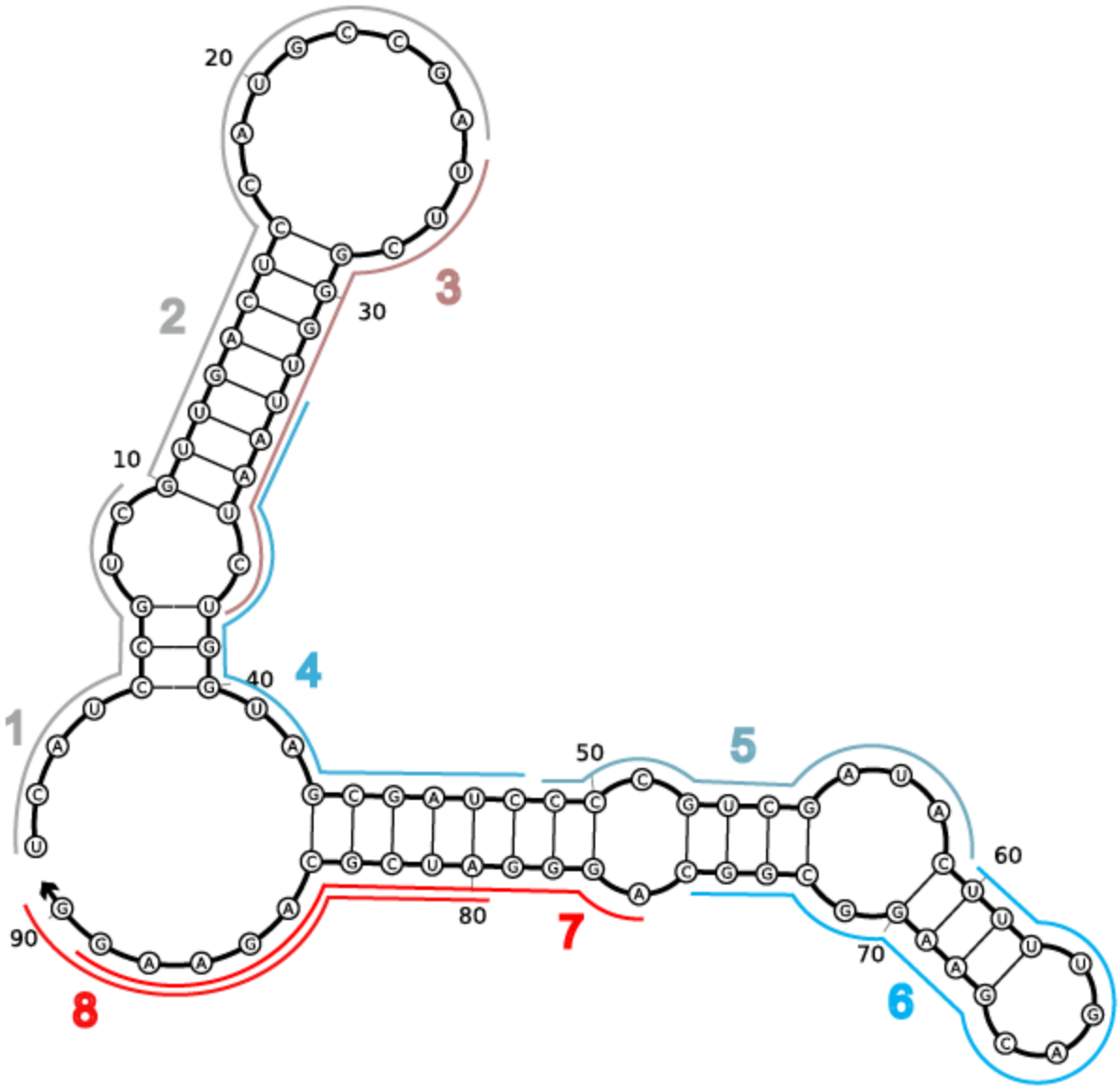

56_sroH
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

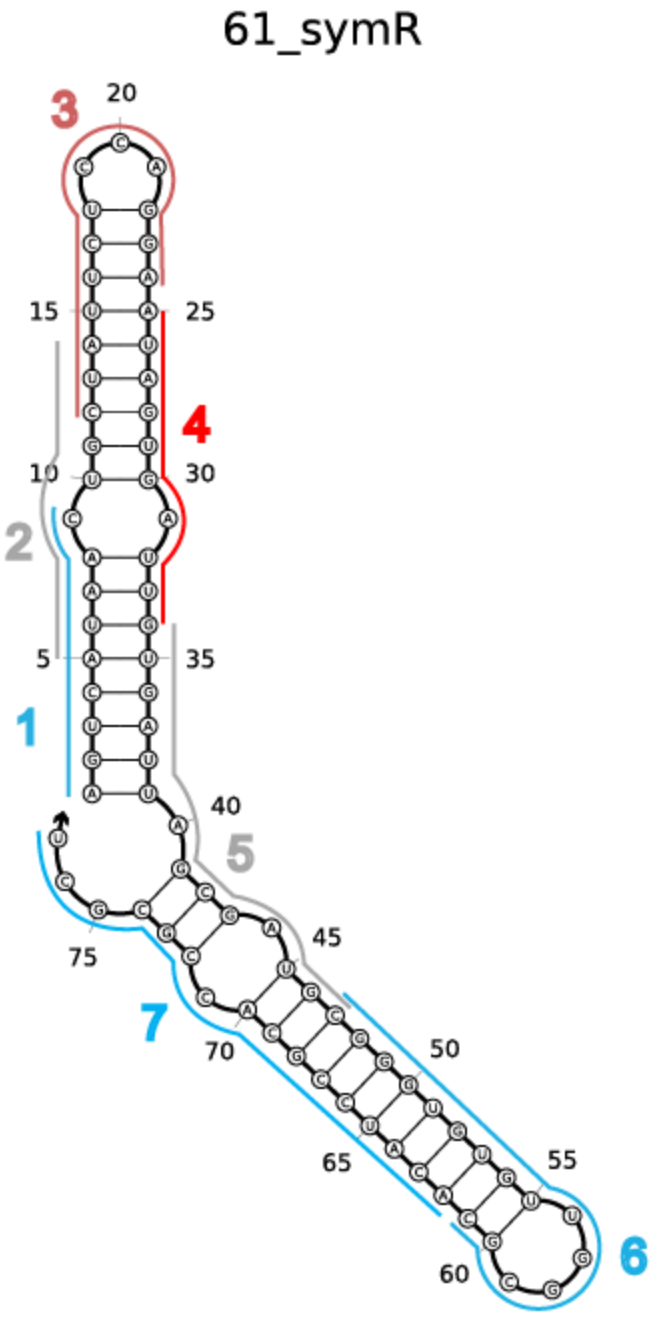

57_sraL
1
2
3
4
5
6
7
8
9
10
11
10
20
30
40
50
60
70
80
90
100
110
120
130
140

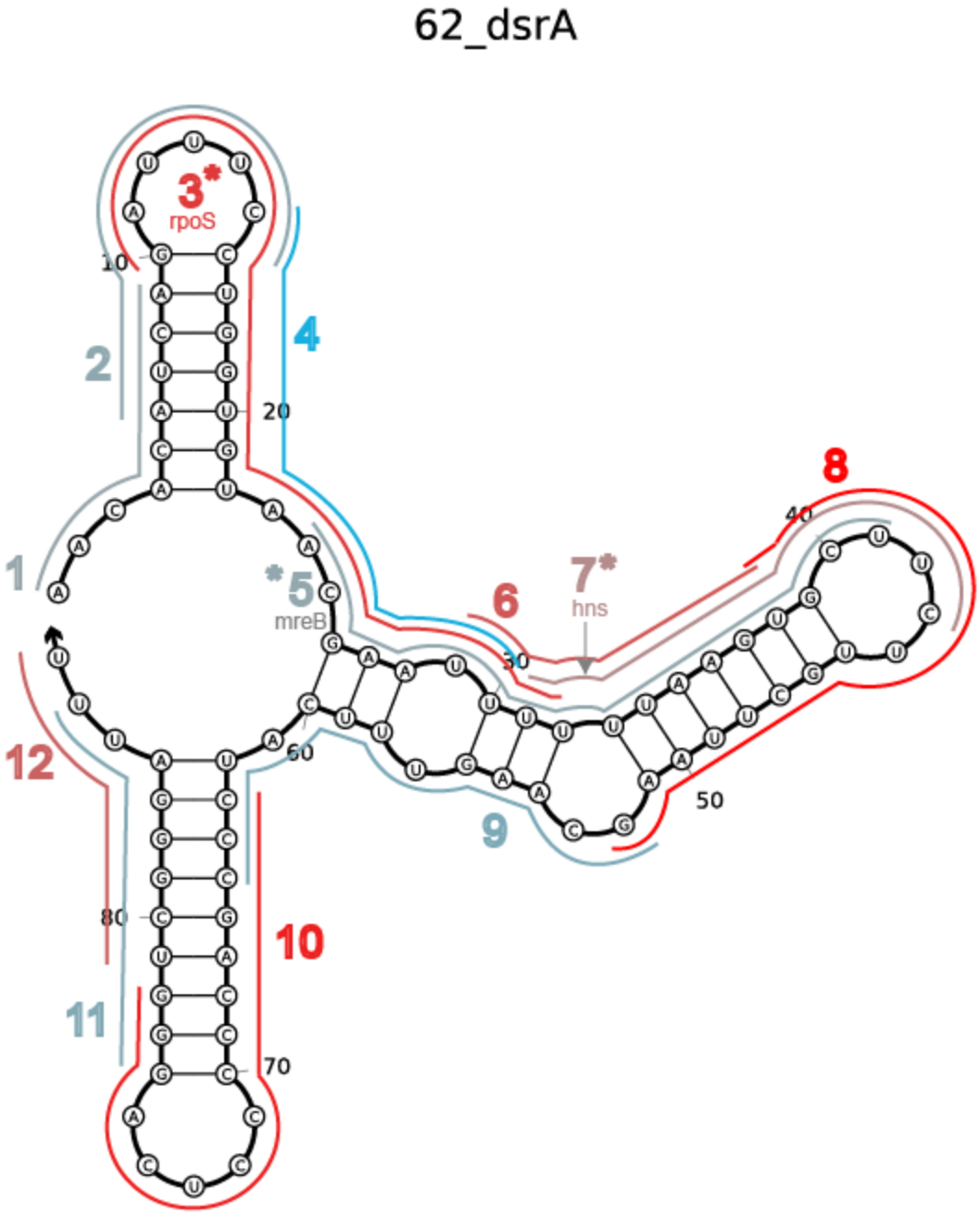

58_ryjA
1
2
3
4
5
6
7
8
9
10
11
12
13
10
20
30
40
50
60
70
80
90
100
110
120
130
140

59_nc5

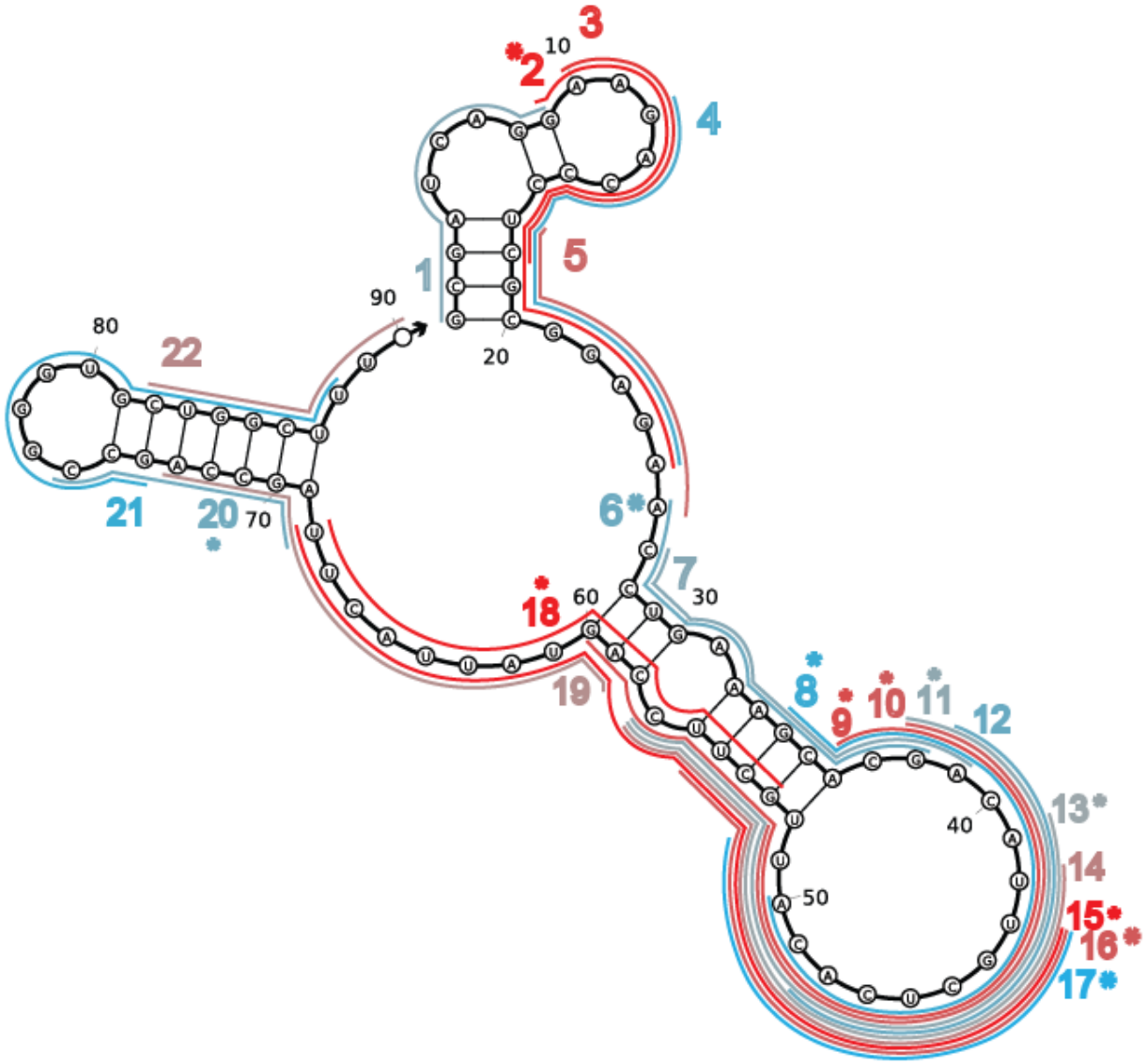


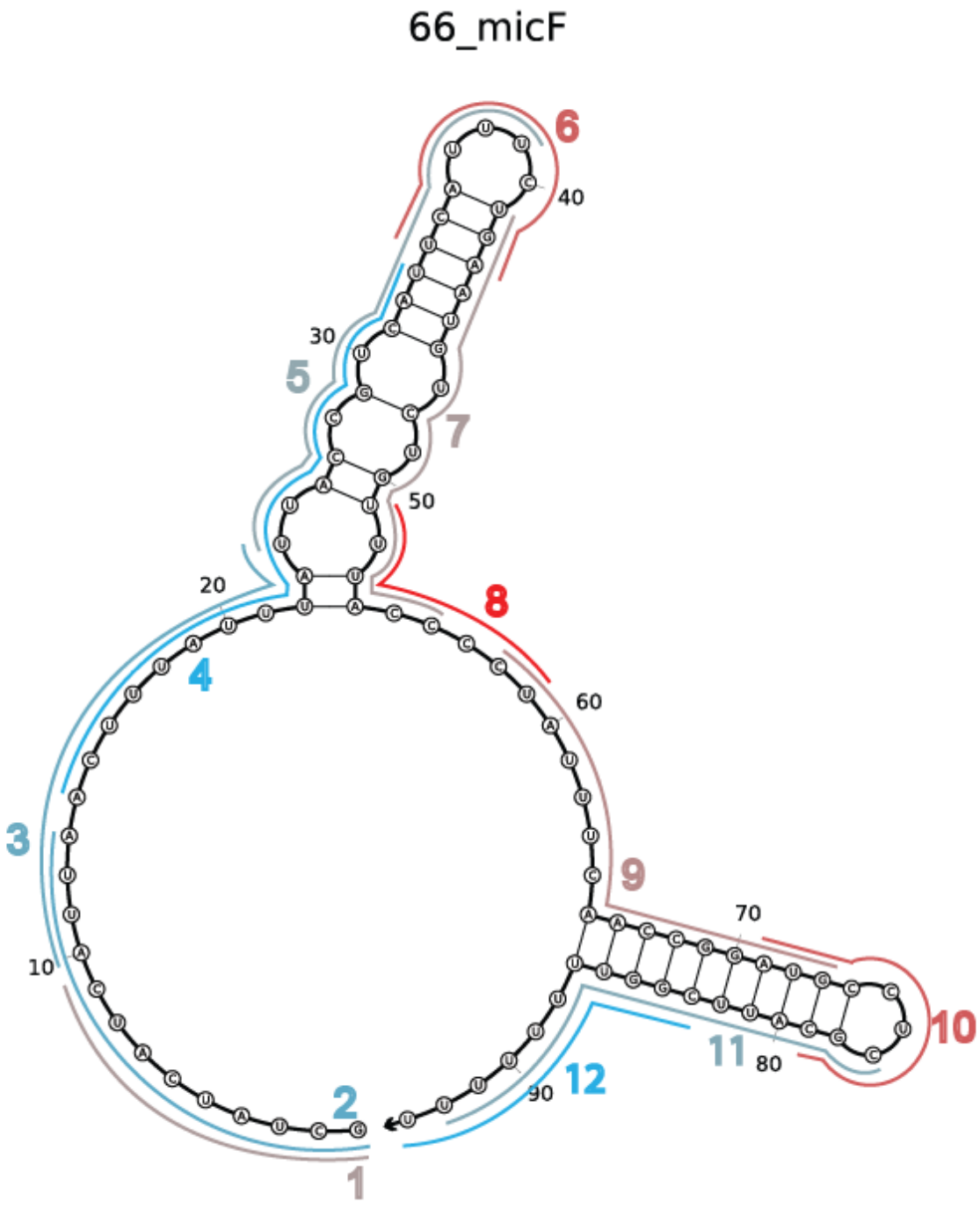

60_ryjB
20
30
10
40
50
60
70
80
90
1
2
3
4
5
6
7
8


61_symR
3
20
15
25
4
10
30
2
5
35
1
40
5
45
75
7
70
50
65
55
60
6

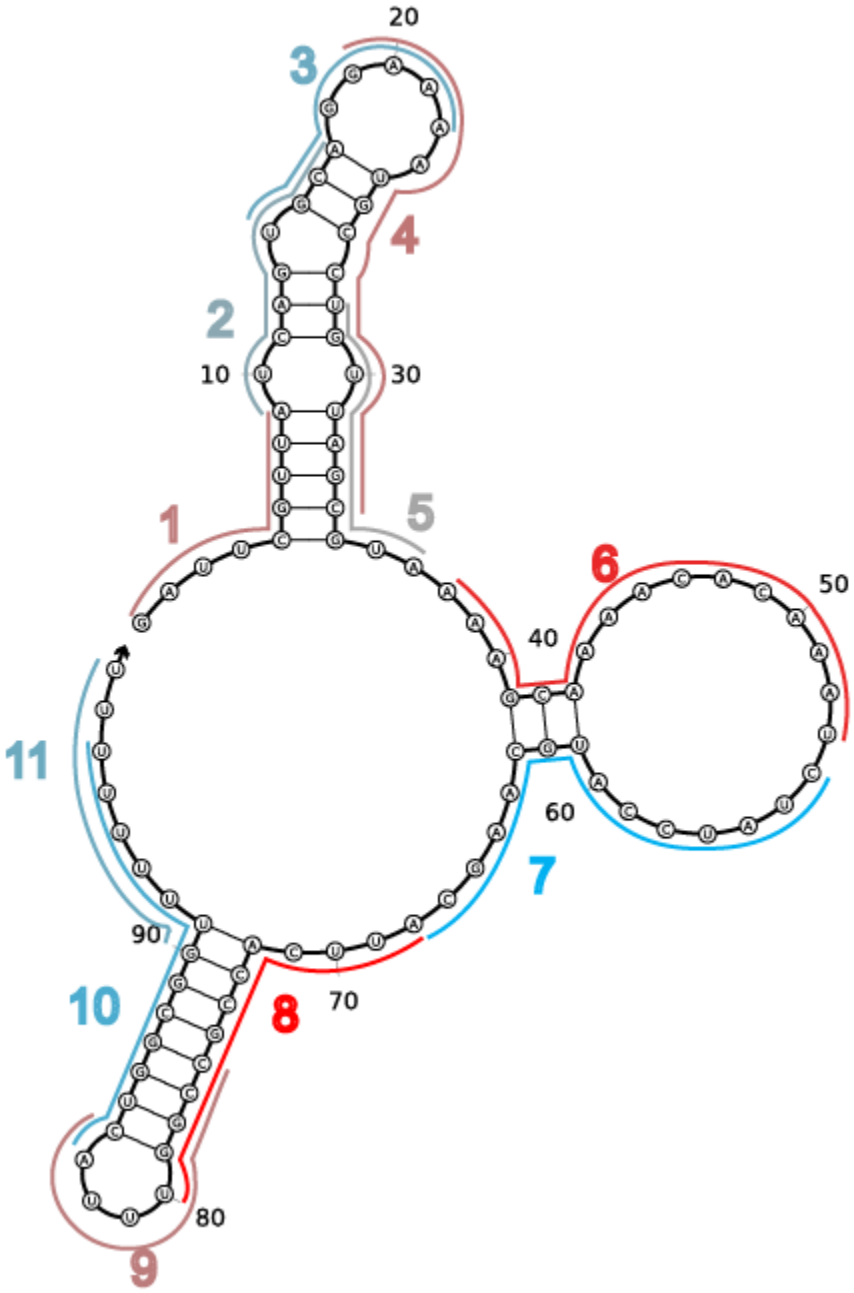

62_dsrA
3*
rpoS
10
2
4
20
1
*5
mreB
6
7*
hns
8
40
30
12
60
50
9
80
10
11
70

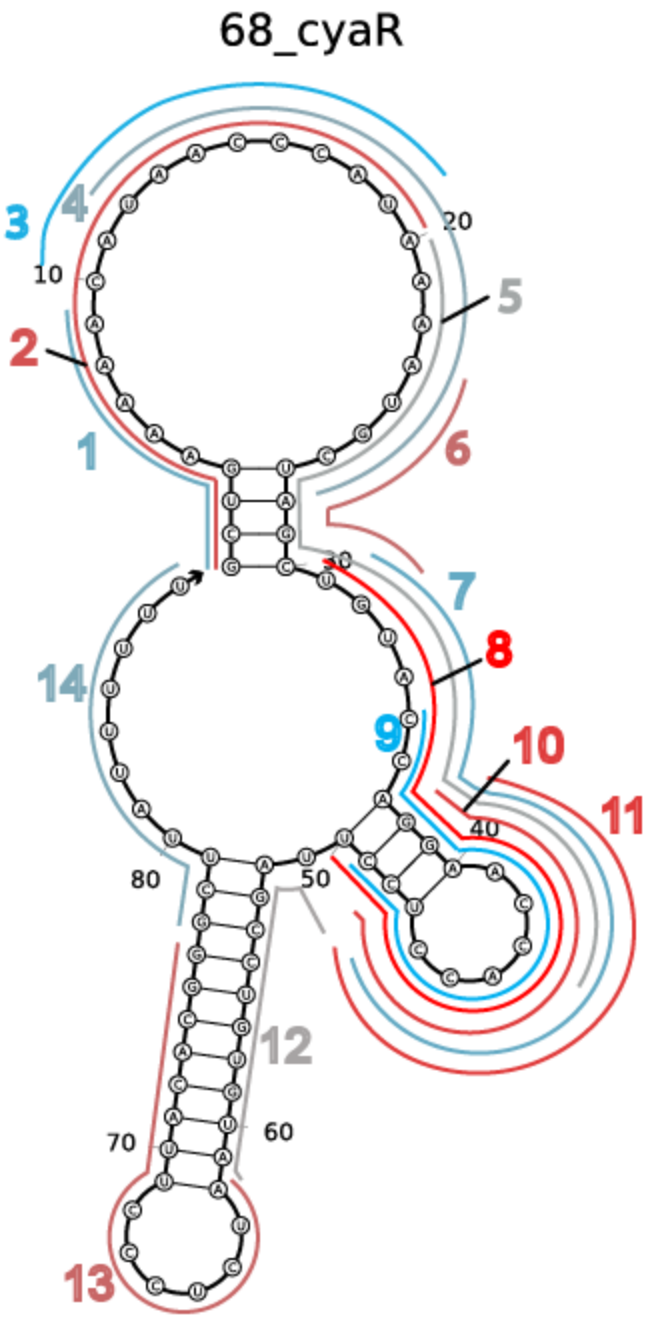
65_ryhB

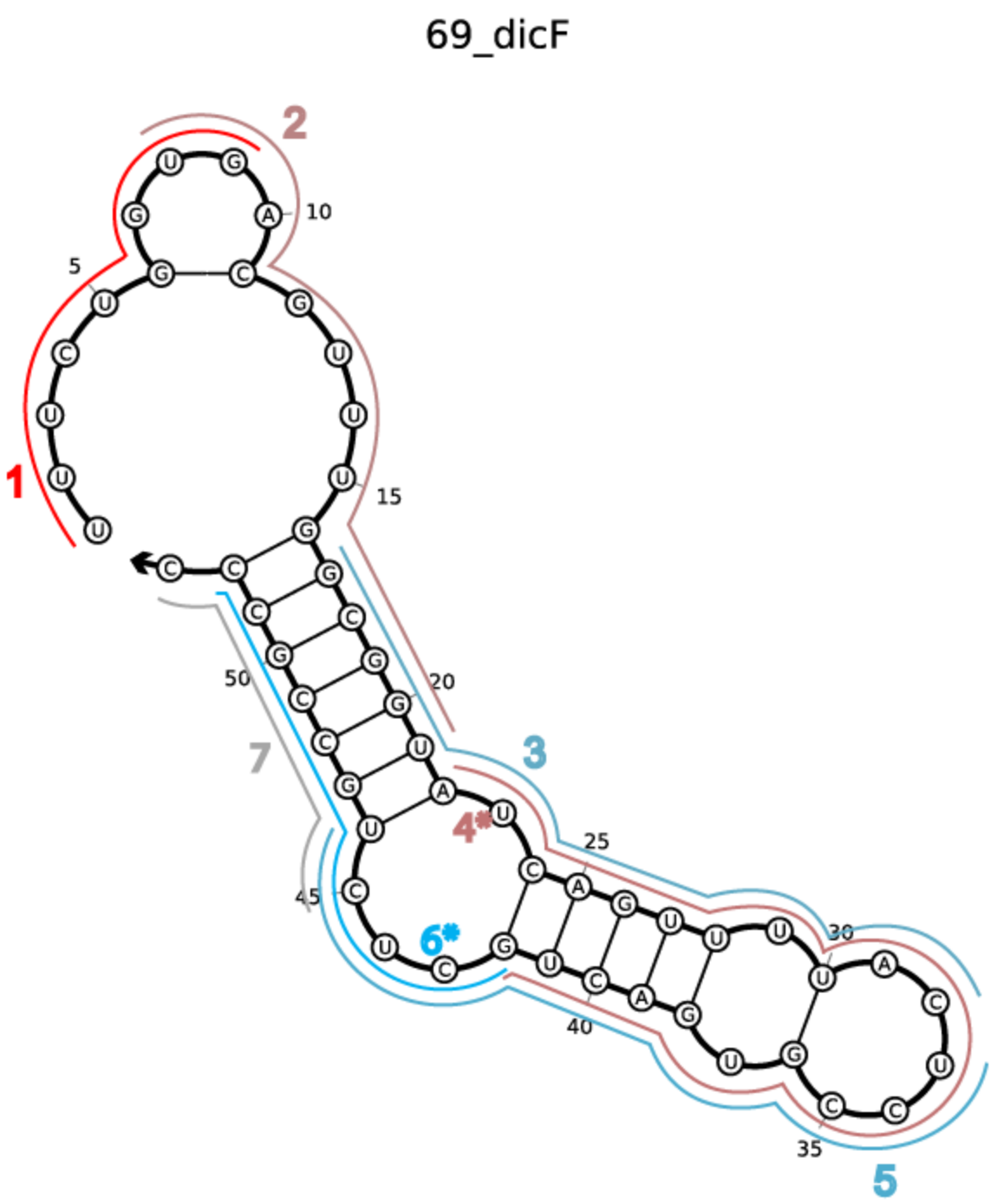

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
10
20
30
40
50
60
70
80
90


66_micF
6
40
30
5
7
50
20
8
4
60
3
9
70
10
11
80
10
12
90
2
1

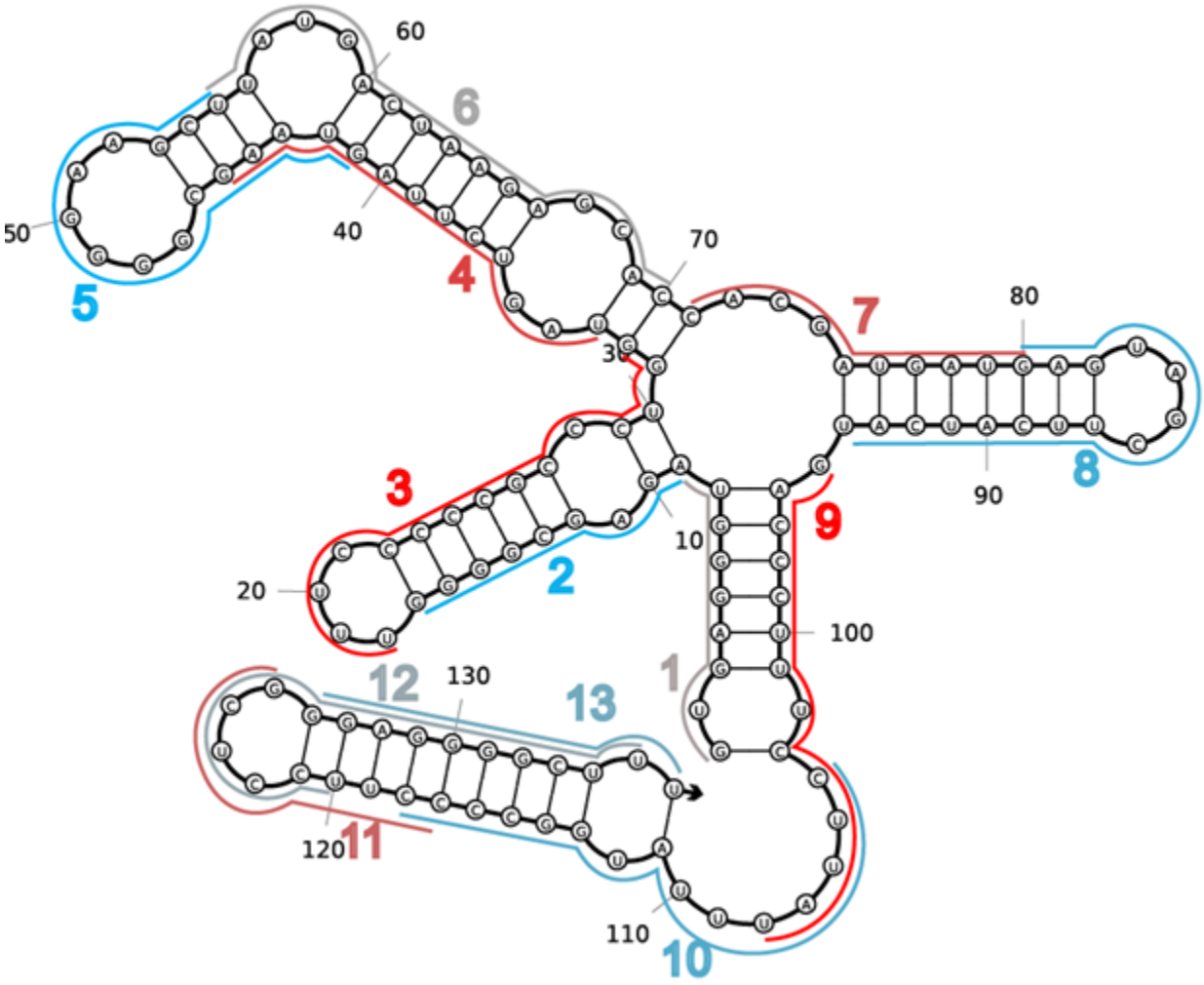
67_mgrR



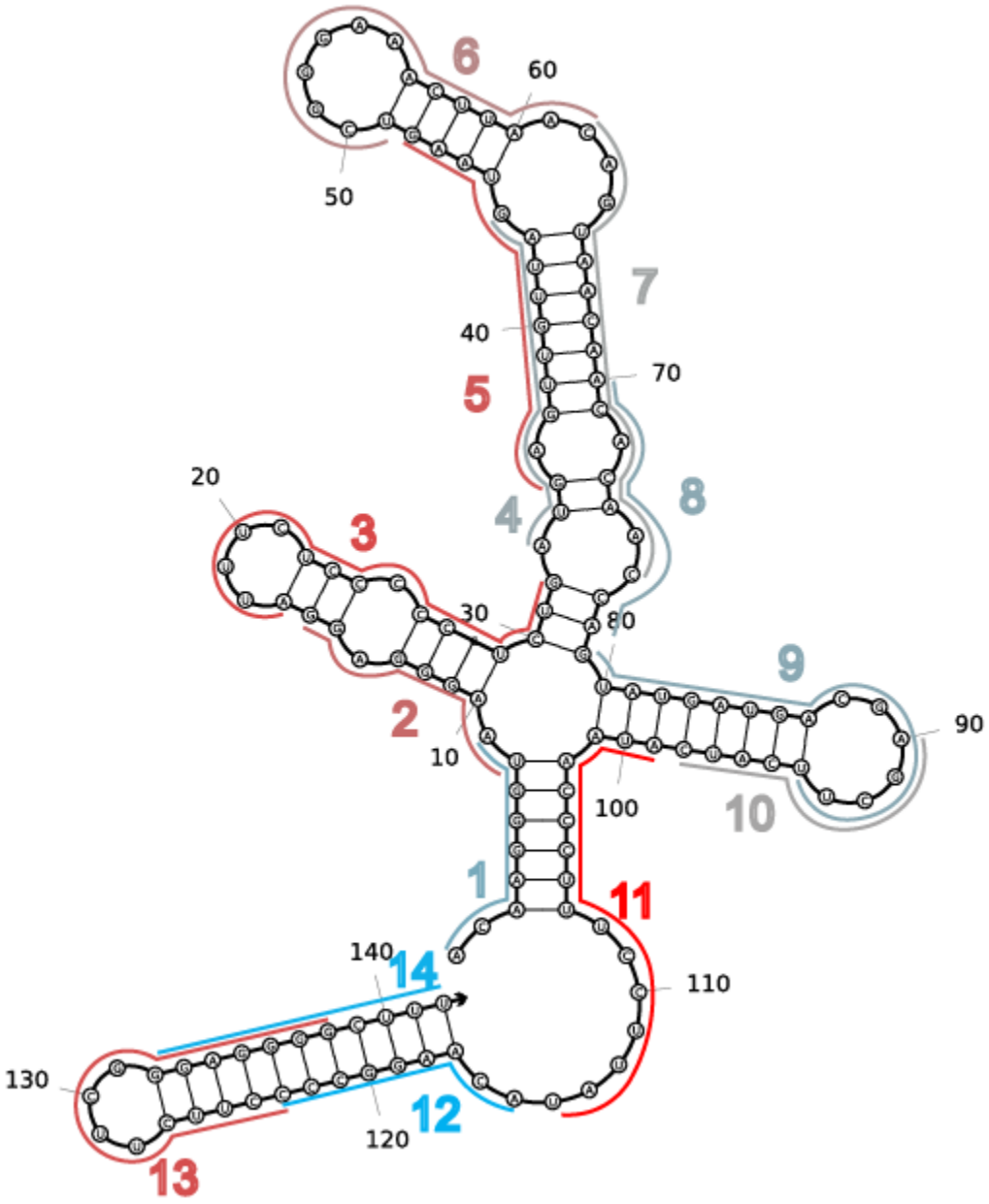

68_cyaR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
10
20
30
40
50
60
70
80


69_dicF
1
2
3
4*
5
6*
7
5
10
15
20
25
30
35
40
45
50

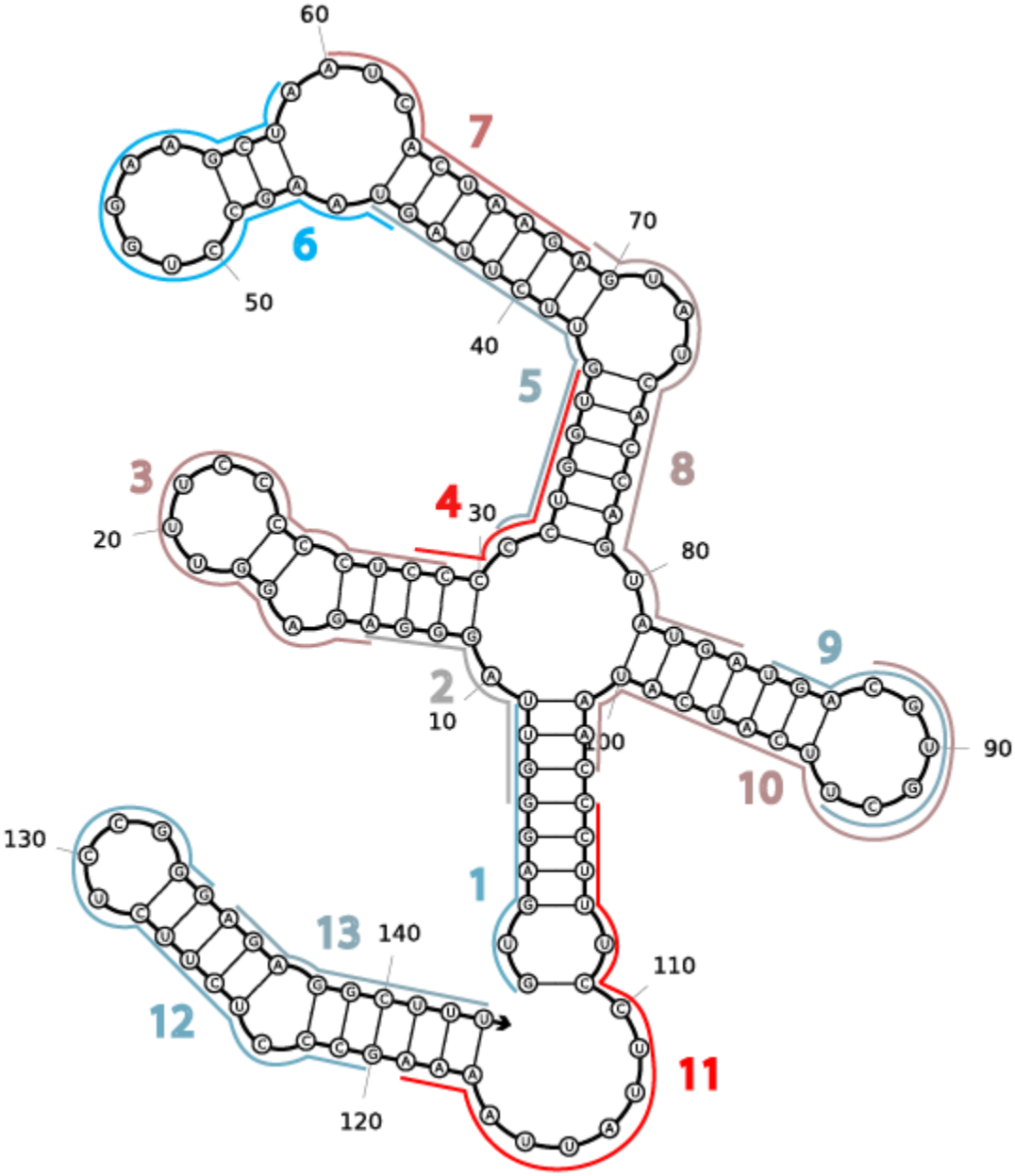
70_sibB



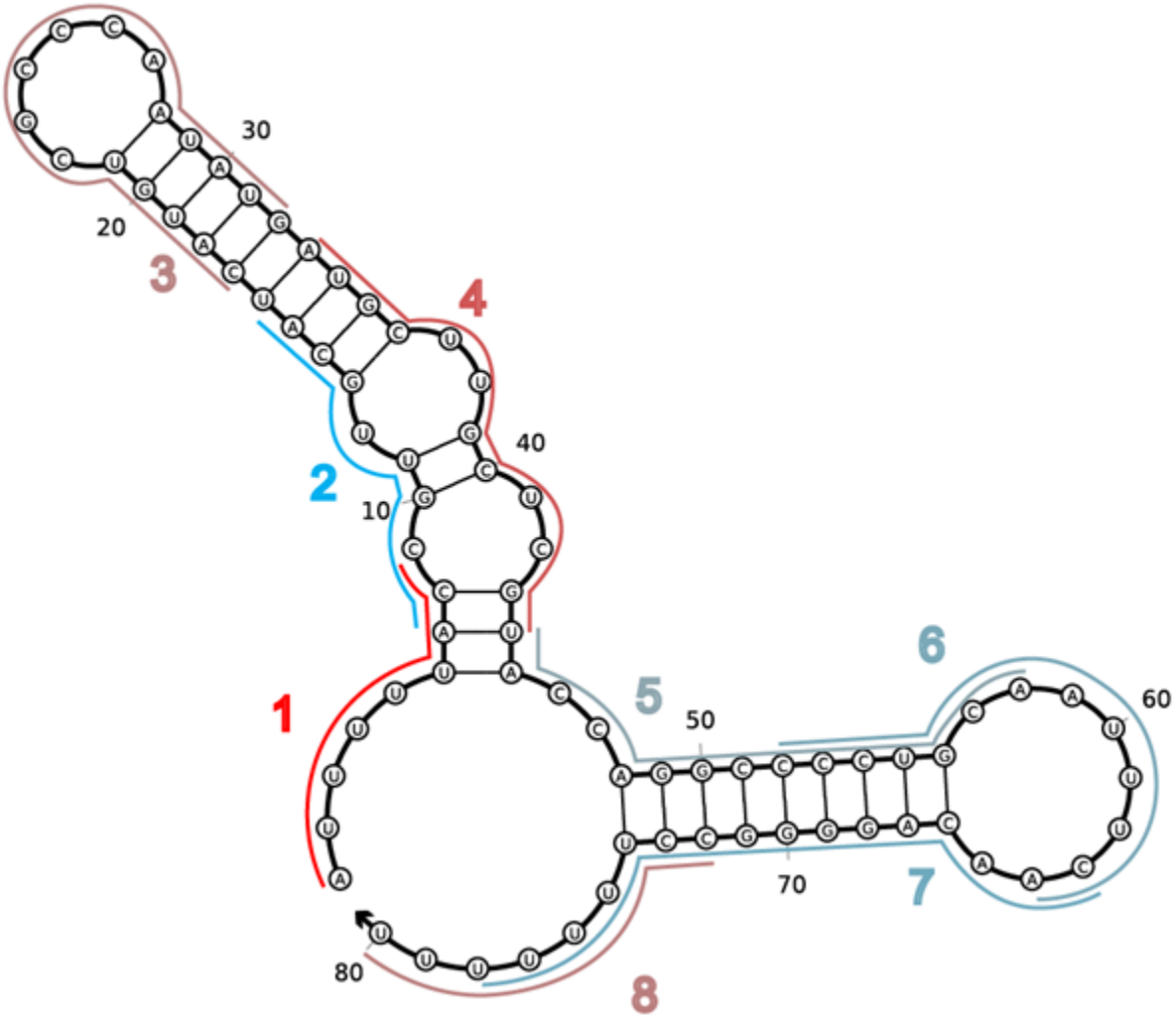
71_sibE


6
60
50
7
40
70
5
20
4
8
3
30
80
9
2
10
90
100
10
1
11
140
14
110
130
120
12
13

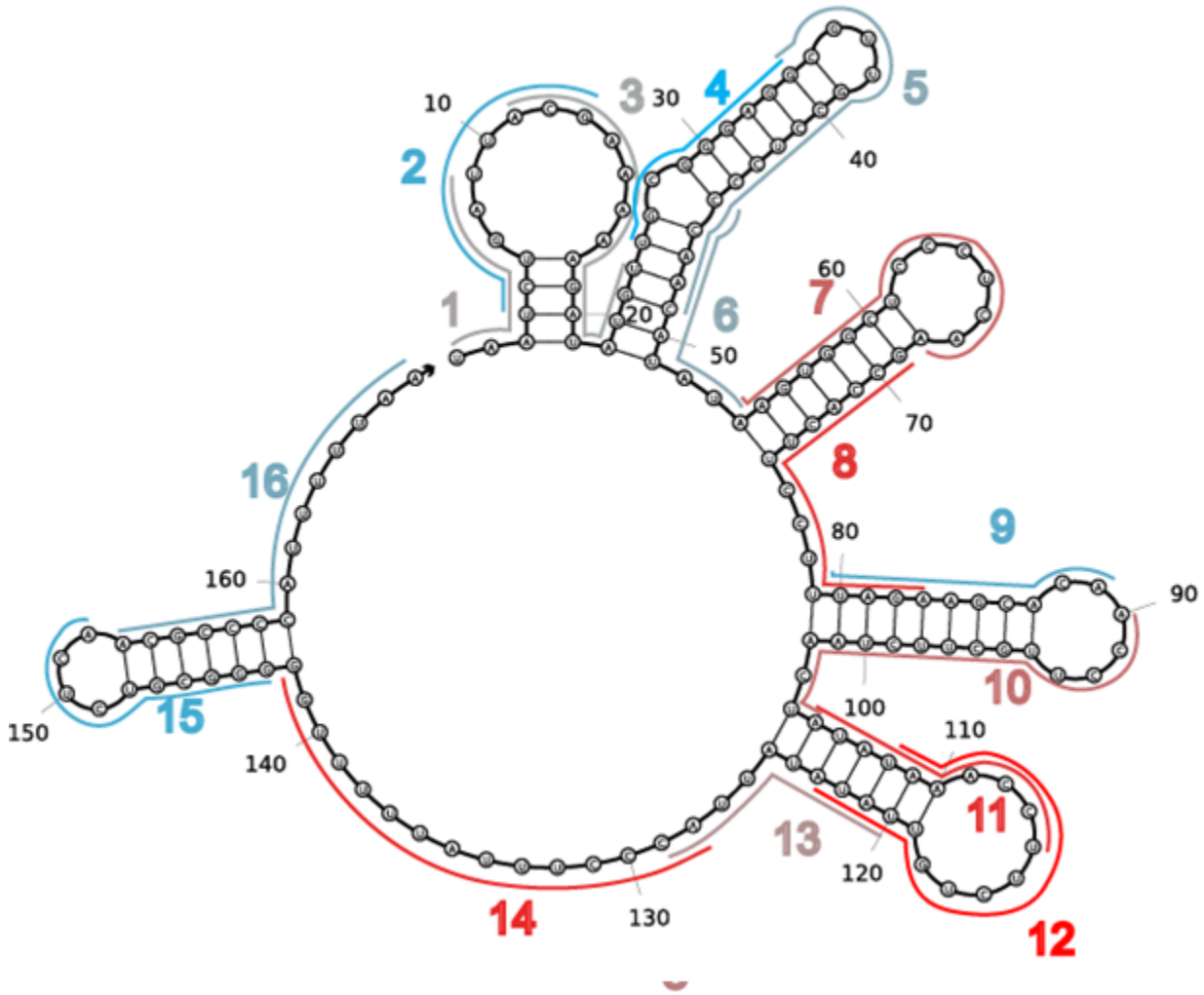
72_sibA



73_micL



74_ipeX


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160

Supplementary tables for Chapter Four

Table C.1. List of sRNAs in this study and relevant information on sRNA-mRNA binding sites, stress-related responses and Hfq-dependence

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
tpke11	TCGCCCTATAAACGG GTAATTATACTGACAC GGGCGAAGGGGAATT TCCTCTCCGCCCGTGC ATTCATCTAGGGGCA ATTTAAAAAAGA	89	Yes	Zhang, et al. (2003)	No			
sokC	GTTCAGCATATAGGA GGCCTCGGGTTGATG GTAAAATATCACTCG GGGCTTTTCT	55	Unknown		Unknown			
SroA	GTTCTCAACGGGGTG CCACGCGTACGCGTG CGCTGAGAAAATACC CGTCGAACCTGATCC GGATAACGCCGGCGA AGGGATTTGAGGCTC CTT	93	Unknown		Unknown			
SgrS	GATGAAGCAAGGGGG TGCCCCATGCGTCAGT TTTATCAGCACTATTT TACCGCGACAGCGAA GTTGTGCTGGTTGCGT TGGTTAAGCGTCCCAC AACGATTAACCATGC TTGAAGGACTGATGC	227	Yes	Ishikawa, et al. (2012)	Yes	ptsG	[157,187]: [157,172] [173,187]	17,22
						manX	[159,172]	18

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
	AGTGGGATGACCGCA ATTCTGAAAGTTGACT TGCCTGCATCATGTGT GACTGAG**TATTGGTG TAAAATCACCCGCCA GCAGATTATACCTGC TGGTTTTTTTT**					manY	[168,179]	20
						yigL	[168,187]	21
tp2	ACTAATTCTTTCGTTG CTCCAGACGACGCAG AGAACGCTCACGGCG GCTCTCTTCACGACTT CTGTCGAGCAAAATTT CTTCGATAAAGGCCA GATGGCGATGCGATG CTTCGCGCGCTTCTTC CGGCTTACCGGCCAT AATCGCTTCAAATATG CGGGTG	161	Unknown		Unknown			
tff	CGGACTTCCGATCCAT TTCGTATACACAGACT GGACGGAAGCGACAA TCTCACTTTGTGTAAC AACACACACGTATCG GCACATATTCCGGGG TGCCCTTTGGGGTCGG TAATATGGGATACGT GGAGGCATAACC	136	Unknown		Unknown			
sraA	CATTCAACGCCGAGA ATAGAGGAAAAATTA AAGGGGAGATAAAAT CCCCCCTTTTTG	57	Unknown		Unknown			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
ffs (4.5S)	GGGGGCTCTGTTGGTT CTCCCGCAACGCTACT CTGTTTACCAGGTCAG GTCCGGAAGGAAGCA GCCAAGGCAGATGAC GCGTGTGCCGGGATG TAGCTGGCAGGGCCC CCACCC	114	No	Zhang, et al. (2003); Pandey, et al. (2014)	No			
nc2	GGAAAAAATCCTCGG CTAATTCGAAAGCGC GCACGGACAGTCCCC TCGCCCCCTCGGGGA GAGGGTTAGGGTGAG GGGAACAGGCCCGCA CAAGCAAACTTATCA GCAATCTCAGGCCGG ATATTCATTCGGCCTT TTACAAAAA	145	Unknown		Unknown			
sroB (ChiX)	ACACCGTCGCTTAAA GTGACGGC**ATAATAA TAA**AAAAATGAAATT CC**TCTTT**GACGGGCC AATAGCGATATTGGC CA**TTTTT**	82	Yes	Moon and Gottesman (2011)	Yes	citA	[46,57]	5
						chiP	[81,92]	N/A

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
sroC	ACTAATTACAAGAACCA GGGGCGGAAATTCCAGC CCTCTCGATTGTTACGTA GCACGGACAGACTATAC GCCTGATGGTCGTTCCCC ATCGGGCCTGAAAACCG CAATACGCTGGGTAACA ATCTTCGAGGGTAGCAG TTAA**CGCTGCTACCC**TC TTTTTTCT	163	Yes	Papenfort and Vanderpool (2015)	Yes			
RybA (mntS)	TCATCCCTCAAGGATCG ACGGGATTAGCAAGTCA GGAGGTCTTATGAATGA GTTCAAGAGGTGTATGC GCGTGTTTAGTCATTCTC CCT	89	No	Gerstle, et al. (2012)	No			
RybB	GCCACTGCTTTTCTTT G**ATGTCCCCATTTTG TGGAGCCCA**TCAACC CCGCCATTTCGGTTCA AGGTTGATGGGTTTTT TGT	81	Yes	Zhang, et al. (2003)	Yes	ompC	[4,33] … [49,60]	7 - 9, 12 (partially)
						tsx	[1,16]	5
						ompW	[1,34]: [1,16][17,34]	5 - 9
						ompA	[1,13]	N/A
						ycfL	[1,25]	5 and 6
						ygiM	[1,22]	5 and 6
						fiu	[1,14]	5
						ompF	[1,22]	5 and 6
						fadL	[1,19]	5 and 6
						nmpC	[1,23]	5 and 6
						hinT	[1,16]	5

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
						rluD	[1,16]	5
						rbsB	[1,10]	N/A
						yfeK	[1,20]	5
						fimA	[5,18]	6
						lamB	[3,24]	N/A
						rraB	[1,23]	5 and 6
						ydeN	[6,19]	N/A
						fumC	[5,19]	N/A
						asr	[1,25]	5 and 6
						rbsK	[1,17]	5
						yhjJ	[1,19]	5 and 6
						sdhC	[3,15]	6
psrD	TAGGCATATTTTTTTC CATCAGATATAGCGT ATTGATGATAGCCATT TTAAACTATGCGCTTC GTTTTGCAGGTTGATG TTTGTTATCAGCACTG AACGAAAATAAAGCA GTAACCCGCAATGTG TGCGAATTATTGGCA AAAGGCAACCACAGG CTGCCTTTTTCTTT	169	No	Pandey, et al. (2014)	No			
rdlA	GTTCTGGTTCAAGATTA GCCCCCGTTCTGTTGTCA GGTTGTACCTCTCAACGT GCGGGGGTTTTCTC	67	No	Pandey, et al. (2014)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
rdlB	GTCTGGTTTCAAGATT AGCCCCCGTTCTGTTG TCAGGTTTTACCTCTC AACGTGCGGGGGTTT TCT	66	No	Bak, et al. (2015)	No			
rdlC	GTCTGGTTTCAAGATT AGCCCCCGTTTTGTTG TCAGGTTTTACCTCTC AACGTGCGGGGGTTT TCTCT	68	No	Bak, et al. (2015)	No			
McaS	TGAAATCTGTCACTGA AGAAAATTGGCAACT AAAGGTTAAAACCGT TATAACACAGTCACC GGCGCAGAGGAGACA ATGCCGGATTTAAGA CGCGGATGCACTGCT GTGTGTACTGTAGAGT CTGGCGGATGTCGAC AGACTCTATTTTTTTA TGCAG	158	Yes	Jorgensen, et al. (2013)	No	csgD	[98,113]	16
						flhD	[127,135]	20
						flhD	[69,78]	13
FnrS	GCAGGTGAATGCAAC GTCAAGCGATGGGCG TTGCGCTCCATATTGT CTTACTTCCTTTTTTG AATTACTGCATAGCA CAATTGATTCGTACGA CGCCGACTTTGATGA GTCGGCTTTTTTTT	122	Yes	Tree, et al. (2014)	No	metE	[37,67]	5-6 (contained)
						sodB	[40,74]	5-7 (contained)
						sodA	[11,47]	3-4 (contained)
						gpmA	[38,57]	5 (contained)
						folX	[1,6]	N/A
						folX	[36,53]	N/A

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
						folE	[1,12]	1 (contained)
						maeA	[31,65]	5 (contained)
						iscR	[44,54]	N/A
						iscR	[27,42]	N/A
						iscR	[2,26]	2 (contained)
						marA	[1,62]	1-5 (contained)
MicC	GTTATATGCCTTTATT GTCACAGATTTTATTT TCTGTTGGGCCATTGC ATTGCCACTGATTTTC CAACATATAAAAAGA C**AAGCCCGAACAGT CGTCCGGGCTTTTTT T**	109	Yes	Tree, et al. (2014)	Yes	ompC	[1,30]	1-3 (contained)
RydC	CTTCCGATGTAGACCC G**TATTCTTCGCCTGT ACCACGGGTCGGTT TTAGTACAGGCGTTT TCTT**	64	Yes	Zhang, et al. (2003)	Yes			
sokB	GCTAGGTTCATTCGTT GGCCTCGGTTGATAG AAATATCGGTCGGGG CCTTCGTCTT	56	Unknown		Unknown			
rydB	ATTATTCTTATCGCCC CTTCAAGAGCTAAGC CACTGAGAGTGCCGG AGATAAGCGCCGGAT GGGGTAG	68	No	Pandey, et al. (2014)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
RprA	ACGGTTATAAATCAA CATATTGATTTATAAG CATGGAAATCCCCTG AGTGAAACAACGAAT TGCTGTGTGTAG**TCTT TGCCCATCTCCCACG ATGGGCTTTTTTT**	105	Yes	Tree, et al. (2014)	Yes	csgD	[79,97]	11
						csgD	[60,74]	N/A
						csgD	[28,45]	5
						rpoS	[33,62]	7 (contained)
						ydaM	[45,75]	8 (contained)
sroD	TTACGTGACGAAGCG CGCGGCAAAGTGGAC AATAAAGCCTGAGCG TTAAGTCAGTCGTCAG ACGCCGGTTAATCCG GCGTTTTTTT	86	Yes		No			
RyeA (SraC)	AAAGTCAGCGAAGGA AATGCTTCTGGCTTTT AACAGATAAAAAGAG ACCGAACACGATTCC TGTATTCGGTCCAGGG AAATGGCTCTTGGGA GAGAGCCGTGCGCTA AAAGTTGGCATTAAT GCAGGCTTAGTTGCCT TGCCCTTTAAGAATAG ATGACGACGCCAGGT TTTCCAGTTTGCGTGC AAAATGGTCAATAAA AAGCGTGGTGGTCAT CAGCTGAAATGTTAA AAACCGCCCGTTCTG GTGA	249	Yes	Pandey, et al. (2014)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
RyeB (sdsr)	GCTGATGACCACCAC GCTTTTTATTGACCAT TTTGCACGCAAACTG GAAAACCTGGCGTCG TCATCTATTCTTAAAG GGCAAGGCAACTAA**G CCTGCATTAATGCCA ACTTTTAGCGCACG**	121	Yes	Zhang, et al. (2003)	Yes			
RyeF	TGATTTTTACCGTTGCAT CATGTCGCCC**AATATGA TGCTTGCTCGTACCAG GCCCC**TGCAATTTCAAC AGGGGCCTTTTTTTATCC CTGAACAGTATAAAAAA CGAACGATAACCGTGAT CTGTTGAGCGGGTGACA GTGCGCATAGCGTTGTG CTAAAAATATTGTATAT ATTCACATTAATTATGG GATTTAAATTACTAAAA CTGATAAATATATATTCT AAATAGCAACTGGGTTA TTCCTTAGCAATTAATGA TTACATTGTAATAAATC ATATTCTTTATCGATTGT TTCAGGCAGTGTTGTGC CTAATTATGCAAGCGGT TAATTCGTTGTATATTTA ATTATACAATGATTTCG GTGTCCAGTAATTTAATT AGAGGAATCT	390	Yes	Zhang, et al. (2003)	Yes			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
isrB	GACAATAACACCTGT ATAACAAATGGTCGG AGTGCCGCGATGAAA CTGCGCAAAATCCTG AAAAGTATGTTCAAT AACTATTGCAAGACG TTCAAAGACGTACCG CCAGGCAATATGTTCC GATAACAAAAAACCT GCTCCGGCAGGTTTTT TTGTGTCC	160	No	Pandey, et al. (2014)	No			
rseX	TTTTTATTATTCTGTG TCATGATGCTTCCGTT ATTAGCCTTTTATCGT CTTGTTTATA**TTTTTT GGGCCGGCATGATG CCGGCTTTTTTTT**	91	Yes	Kim, et al. (2015)	Yes	ompC	[30,55]	5 (contained)
						ompA	[37,50]	5
isrC	ACGATCAATATCTATT TTATCGATCGTTTATA TCGATCGATAAGCTA ATAATAACCTTTGTCA GTAACATGCACAGAT ACGTACAGAAAGACA TTCAGGGAACAACAG AACCACAATTCAGAA ACTCCCACAGCCGGA CCTCCGGCACTGTAAC CCTTTACCTGCCGGTA TCCACGTTTGTGGGTA CCGGCTTTTTTATTCA CC	204	No	Pandey, et al. (2014)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
tpke70	AAAGCCATAAAAACCA TGAGGTTATTATGGCC GATTTGAGGAGGGAAA GAGTAAGAGCAGTTTG TTAAATGTACAACGAC GATTCTCCCACCGGGC GCGTTTTAAAGCGACG GTGGATCCAGAGGTAC TGCTCCGGTGCGCGCA TGATCTCTTTCTCGATA ATCTTGTTCATATAGGC AGCGGCTTGATTTTCAT CTGTCGGGTAGCCTTCC ATCTCTGGGGTGATGA ACAAACGATATCCGCT GTAATCCGCTTTTCTTA CCATCGTTACGGTCAA CATGGCTGCGCCAGAG AGACGGGAGAGAACAT AGGTGCCATTGGTTGT GGCGACATTTTCCACC GCAAAGAACGGCGCGA AGGAGCTGCCTTTACG ACCATAATCCTGATCG GGAGCAAACCATACCG CTTCACCTTTCTTCAGT GCACCGACAATGCC	436	Unknown		Unknown			
sroE	ATAACGTGATGGGAA GCGCCTCGCTTCCCGT GTATGATTGAACCCG CATGGCTCCCGAAAC ATTGAGGGAAGCGTT GAGGGTTCATTTTTAT	92	Unknown		Unknown			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
ryfA	GCGGCCCTTTCCGCCG TCTCGCAAACGGGCG CTGGCTTTAGGAAAG GATGTTCCGTGGCCGT AAATGCAGGTGTTTC ACAGCGCTTGCTATCG CGGCAATATCGCCAG TGGTGCTGTCGTGATG CGGTCTTCGCATGGAC CGCACAATGAAGATA CGGTGCTTTTGTATCG TACTTATTGTTTCTGG TGCGCTGTTAACCGA GGTAAATAATAACCG GAGTCTCTCCGGCGA CAATTTACTGGTGGTT AACAACCTTCAGAGC AGCAAGTAAGCCCGA ATGCCGCCCTTTGGGC GGCATATTTT	304	No	Pandey, et al. (2014)	No			
GlmY	AGTGGCTCATTCACCG ACTTATGTCAGCCCCT TCGGGACGTGCTACA TAAAATACGAATGAC GCACAACAAGGTGCC TGCCGTCCAACTTCTG ATATCAGCGTAGCTAT ATCAACCATCGGGCG AAACGTCGAGTTAGG CACCGCCTTATTCCAT AACAAAGCCGGGTAA TTCCCGGCTTTGTT	184	Yes	Göpel, et al. (2015)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
ryfB	CGTTATTGAAGATTTT GCTGTGCTTTACACCA TGCCACAGAATTCCCC CATTGAAACGAGTGG TGTCGTCAAAGCTCTG GTGTGGAGTGCAGCA TGCACCCTCAATAACT CGCACGTTCAGTTTTG GGGAGATGTAAGGGC TAATCTGAATGGCTGC ATTCCTTGTTTAAGGA AAAACGAATGACTGA TTGCCGATACCTGATT AAACGGGTCATCAAA ATCATCATTGCTGTTT TACAGCTGATCCTTCT GTTCTTATAACACAAG GAAACGTACTTAAGG TGCGTCCGGTGAACC AGTCGGACGCACCTTT AATAAC	319	No	Bak, et al. (2015)	No			
ryfC (ohsC)	GTTGAGGGTGCATGC TGCACAAAATTAAAG TTAAAAAGTAAAACC CCCGTTCCTTACCAGT TCGGGGGTTTTACTTT	77	Unknown		Unknown			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
ryfD	AATCAAGACGATCCG GTACGCGTGATTTTCT TTTCACATTAATCTGG TCAATAA**CCTTGAAT AATTGAGGGATGAC CTCA**TTTAATCTCCAG TAGCAACTTTGATCCG TTATGGGAGGAGTTA TGCGTCTGGATCGTCT TACT	143	Yes	Pandey, et al. (2014)	Yes			
SraD (MicA)	GAAAGACGCGCATTT GTTATCATCATCCCTG AATTCAGAGATGA**AA TTTT**GGCCACTCACG AGTGGCCTTTT	72	Yes	Tree, et al. (2014)	Yes	phoP	[6,31]	4(contained)
						lamB	[8,30]	4
						ompA	[8,25]	4-5(contained)
						ompW	[2,30]	3-4(contained)
						tsx	[1,23]	2
						ygiM	[6,30]	4(contained)
						yfeK	[11,28]	8
						ompX	[1,23]	2
						fimB	[4,14]	3
GcvB	ACTTCCTGAGCCGGA ACGAAAAGTTTTATC GGAATGCGTGTTCTG GTGAACTTTTGGCTTA CGGTTGTGATGTTGTG TTGTTGTGTTTGCAAT TGGTCTGCGATTCAGA CCATGGTAGCAAAGC	205	Yes	Tree, et al. (2014)	Yes	csgD	[69,86]	8 (69, 87)
						dppA	[60,94]	7-10 (contained)
						phoP	[148,174]	17-18 (contained)
						cycA	[124,161]	15-17 (contained)

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
	TACCTTTTTTCACTTC CTGTACATTTACCCTG TCTGTCCATAGTGA**TT AATGTAGCACCGCCT AATTGCGGTGCTTTT TTT**					sstT	[64,99]	8-11 (contained)
OmrA	CCCAGAGGTATTGATT GGTGAGATTATTCGGT ACGCTCTTCGTACCCT GTCTCTTGCACC**AACC TGCGCGGATGCGCA GGTTTTTTTT**	88	Yes	Tree, et al. (2014)	Yes	fepA	[16,48]	4-5 (contained)
						ompR	[1,19]	2
						ompT	[1,33]	1-3 (contained)
						flhD	[1,50]	1-5 (contained)
OmrB	CCCAGAGGTATTGAT AGGTGAAGTCAACTT CGGGTTGAGCACATG AATTACACCA**GCCTG CGCAGATGCGCAGG TT**	76	Yes	Tree, et al. (2014)	Yes	ompT	[1,32]	1-5 (contained)
						ompR	[1,19]	1-2 (contained),4 (contained)
						cirA	[2,15]	2
						csgD	[2,20]	3
						flhD	[1,49]	1-7 (contained)
InvR	TGCGGAATGCAGAAA GTTTTATGTAGGTTAA GGTGTGAAACGTCCG CACCAATAAAGCCCG GCGAGGTGATGCCAA CCTGGGCGTTCATGTT C	93	Unknown		No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
RygC	GTAAGGGTAAGGGAG GATTGCTCCTCCCCTG AGACTGACTGTTAAT AAGCGCTGAAACTTA TGAGTAACAGTACAA TCAGTATGATGACAA GTCGCATCATAACCCT TCTCCTTCAAGCCCTC GCTTCGGTGAGGGCTT T	140	Yes	Pandey, et al. (2014)	No			
sroG	GCTTATTCTCAGGGCG GGGCGAAATTCCCCA CCGGCGGTAAATCAA CTCAGTTGAAAGCCC GCGAGCGCTTTGGGT GCGAACTCAAAGGAC AGCAGATCCGGTGTA ATTCCGGGGCCGACG GTTAGAGTCCGGATG GGAGAGAGTAACG	149	Unknown		No			
RygD (sibD)	ACAAGGGTGAGGGAG GATTTCTCCCCCCTCT GATTGGCTGTTAATAA GCTGCGAAACTTACG AGTAACAACACAATC AGTATGATGACGAGC TTCATCATAACCCTTT CCTTCTGTAAGGCCCC CTTCTTCGGGAGGGG CTTTCC	145	No	Pandey, et al. (2014)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
psrN	GCAAAGGGGAGTAAC TTCATTGCCGGTCGAT CGTCATTACGATGTGT GAAAAAACACATCCG GTCACCGGGCAACCC GAAAGGAATACGCAG ACGTATTCCTTTTTTG TTGTAAGTGAGACCTT GCCGGAAGGCGAGGT CTATGCATAAAAAGC AGCGGCTGACGTCTTC CGACGTTGGCCGTTTT TT	188	Unknown		Unknown			
SraH (ArcZ)	GTGCGGCCTGAAAAA CAGTGCTGTGCCCTTG TAACTCATCATAATAA TTTACGGCGCAGCCA AGATTTCCCTGG**TGTT** GGCGCAGTATTCGCG C**AC**CCCGGTCTAGCC	108	Yes	Zhang, et al. (2003)	Yes	rpoS	[66,91]	9
						flhD	[1,17]	2
arrS	GTAATCCGATTTAAAT ATCGAGTCTCCTTGTT TCGACTTAAGCTGGC AATTGGATTGCCAGCT TTCTTT	69	Yes		No			
GadY	ACTGAGAGCACAAAG TTTCCCGTGCCAACAG GGAGTGTTATAACGG TTTATTAGTCTGGAGA CGGCAGACTATCCTCT TCCCGGTCCCCTATGC CGGGTTTTTTT	105	Yes	Kim, et al. (2015)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
rdlD	GTCTAGAGTCAAGAT TAGCCCCCGTGGTGTT GTCAGGTGCATACCT G**CAACGTGCGGGGG TTTT**	64	No	Pandey, et al. (2014)	Yes			
istR-1, istR-2	GTTGACATAATACAG TGTGCTTTGCGGTTAC CAGCCGCAGGCGACT GACGAAACCTCGCTC CGGCGGGGTTTTTT	75	No	Olejniczak (2011)	No			
GlmZ	GTAGATGCTCATTCCA TCTCTTATGTTCGCCT TAGTGCCTCATAAACT CCGGAATGACGCAGA GCCGTTTACGGTGCTT ATCGTCCACTGACAG ATGTCGCTTATGCCTC ATCAGACACCATGGA CACAACGTTGAGTGA AGCACC**CACTTGTTG TCATACAGACCTGTT TT**	172	Yes	Tree, et al. (2014)	Yes	glmS	[150,169]	16
Spot_ 42	GTAGGGTACAGAGGT AAGATGTTCTATCTTT CAGACCTTTTACTTCA CGTAATCGG**ATTT**GG CTGAAT**ATTTT**AGCC GCCCCAGTCAGTAAT GACTGGGGCG**TTTTT TA**	109	Yes	Kim, et al. (2015)	Yes	gltA	[4,13]	5
						caiA	[23,38]	N/A
						sucC	[23,34]	9
						sthA	[48,55]	13(48-56)
						srlA	[20,34]	7
						fucI	[4,57]	4-13 (contained)
						galK	[1,62]	1-14 (contained)

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
						nanC	[1,17]	2
						paaK	[1,9]	1
						ascF	[1,8]...[52,60]	1(1-9), 14
						sdhC	[4,15]	4
						fucP	[46,56]	12
						xylF	[1,33]	1-6 (contained)
						atoD	[4,11]	N/A
						gdhA	[40,56]	12-13 (contained)
						puuE	[29,57]	10 (contained), 12-13 (contained)
OxyS	GAAACGGAGCGGCACCTCTTTTAACCCTTGAAGTCACTGCCCGTTTCGAGAGTTTCTCAACTCG**AATAACTAA**AGCCAACGTGAACTTTTGCGGATCTCCAGGATCCGCT	110	Yes	Tree, et al. (2014)	Yes	flhD	[53,74]	6 (contained)
sroH	GAAAATAAGAACACATGTTCTCATCTTCCAGGATGCAGCAGACTGAAGAAATTCAGACATCCCGCAACCTGCGATTATCGCAAGGTCAAGGCAAAGTCCGGTAATGGCGTTCTGAATACCAGAGATAATTCTCTGGCGAAACCCACCTTAAGGTGGGTTTT	161	Unknown		N/A			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
SraL	ATCAACACCAACCGG AACCTCCACCACGTG CTCGAATGAGGTGTG TTGACGTCGGGGGAA ACCCTCCTGTGTACCA GCGGGATAGAGAGAA AGACAAAGACCGGAA AACAAACTAAAGCGC CCTTGTGGCGCTTTAG TTT	140	No		No			
ryjA	ATCAACACCAACCGG AACCTCCACCACGTG CTCGAATGAGGTGTG TTGACGTCGGGGGAA ACCCTCCTGTGTACCA GCGGGATAGAGAGAA AGACAAAGACCGGAA AACAAACTAAAGCGC CCTTGTGGCGCTTTAG TTTT	141	Yes	Wassarman, et al. (2001)	No			
nc5	ATCCGGCCTACGGAGGG TGCGGGAATTTGTAGGC CTGATAAGACGCGCAAG CGTCGCATCAGGCAGTC GGCACCATTGCCGGATG CGGCGTAAACGCCTTAT CC	104	Unknown		N/A			
ryjB	TCATCCGTCGTTGACT CCATGCCGATTCGGGT TAATCTGGTAGCGATC CCCGTCGATACTTTTG ACGAAGGCGGCAGGG ATCGCAGAAGG	90	Yes	Zhang, et al. (2003)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
symR	AGTCATAACTGCTATT CTCCAGGAATAGTGA TTGTGATTAGCGATGC GGGTGTGTTGGCGCA CATCCGCACCGCGCT	77	No	Kawano, et al. (2007)	No			
dsrA	AACACATCAGATTTCC TGGTGTAACGAA**TTT TTT**AAGTGCTTCTTGC TTAAGCAAGTTTCATC CCGACCCCCTCAGGG TCGGGATTT	87	Yes	Tree, et al. (2014)	Yes	mreB	[24,41]	5
						rpoS	[10,32]	3
						hns	[31,43]	7
ryhB	GCGATCAGGAAGACC CTCGCGGAGAACCTG AAAGCACGACATTGC TCACATTGCTTCCAG**T ATTA**CTTAGCCAGCC GGGTGCTGGCTTTT	90	Yes	Tree, et al. (2014)	Yes	iscS	[43,68]	15
						sdhD	[9,50]	2-8 (contained)
						sodB	[34,64]	8-14 (contained)
						nirB	[38,57]	11
						marA	[38,57]	11
						erpA	[36,60]	9
						sdhC	[43,51]	17
						sdhCDAB	[9,50]	2-8 (contained)
						shiA	[19,75]	6-19 (contained)
						cirA	[41,57]	13
						fur	[38,76]	10-20 (contained)
						msrB	[1,12]...[26, 38]	1 (contained), 6

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
MicF	GCTATCATCATTAACT TTA**TTTATT**ACCGTCA TTCATTTCTGAATGTC TGTTTACCCC**TATTTC AACCGGATGCCTCG CATTCGGTTTTTTTT**	93	Yes	Zhang, et al. (2003)	Yes	OmpFs	[1,33]:[1,13] [14,33]	2,4
						lrp	[1,13]	2
MgrR	GATTCGTTATCAGTGC AGGAAAATGCC**TGTT** AGCGTAAAAGCAAAA CACAAATCTATCCATG CAAGCATTC**ACCGCC GGTTTACTGGCGGTT TTTTTT**	98	Yes	Kim, et al. (2015)	Yes			
CyaR	GCTGAAAAACATAAC CCATAAAATGCTAGC TGTACCAGGAACCAC C**TCCTTAGCCTGTGT AATCTCCCTTACACG GGCTTATTTTTT**	87	Yes	Tree, et al. (2014)	Yes	luxS	[35,49]	9
						yqaE	[31,50]	8
						nadE	[35,48]	9(35-49)
						yobF	[1,43]	1-5 (contained)
						OmpX	[38,48]	10
dicF	TTTCTGGTGACGTTTG GCGGTATCAGTTTTAC TCCGTGACTGCTCTGC CGCCC	53	Yes	Zhang, et al. (2003)	No	ftsZ	[22,52]	4-6 (contained)
SibB	GTGAGGGTAGAGCGGGG TTTCCCCCGCCCTGGTAG TCTTAGTAAGCGGGGAA GCTTATGACTAAGAGCA CCACGATGATGAGTAGC TTCATCATGACCCTTTCC TTATTTATGGCCCCTTCC TCGGGAGGGGCTTT	136	Unknown		N/A			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
SibE	ACAAGGGTAAGGGAG GATTTCTCCCCCCTCT GATGAGTTGTTAGTA AGTCGGGAAACTTAA CAGTAACAACACAAC CAGTATGATGACGAG CTTCATCATAACCCTT TCCTTATACAAGGCCC CTTCTTCGGGAGGGG CTTT	142	Unknown		N/A			
SibA	GTGAGGGTTAGGGAG AGGTTTCCCCCTCCCC CTGGTGTTCTTAGTAA GCCTGGAAGCTAATC ACTAAGAGTATCACC AGTATGATGACGTGC TTCATCATAACCCTTT CCTTATTAAAAGCCCT CTTCTCCGGGAGAGG CTTT	143	Unknown		N/A			
MicL	ATTTTTACCGTTGCAT CATGTCGCCCAATATG ATGCTTGCTCGTACCA GGCCCCTGCAATTTCA ACAGGGGCCTTTTTTT	80	Yes	Guo, et al. (2014)	No			

sRNA	Sequence	Length	Hfq Dependent?	Reference for Hfq Dependency	Known Hfq Binding Site	Target mRNA	Interacting Nucleotides	Probe within Interacting Region
IpeX	TAATCTGATTACGAA AAAGATATGTTGCGG GAGGCGTTGCCTCCCC AACATATAAGTGGCT CCCTCAAGCCACTTCC TTTAGAAGCACAACC TTGCTTCTAACTATAT AAACCTTCTGTTATAT ATTACCCTTTATTTTT GGGGGCGTCTCAACG CCCCATTTTTAA	167	No	Catillo-Keller, et al. (2006)	No			

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Arunasri, K., M. Adil, P. A. A. Khan and S. Shivaji (2014). "Global Gene Expression Analysis of Long-Term Stationary Phase Effects in <italic>E. coli</italic> K12 MG1655." PLoS ONE **9**(5): e96701.

Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner and H. Mori (2006). "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection." Mol Syst Biol **2**: 2006 0008.

Babitzke, P. and T. Romeo (2007). "CsrB sRNA family: sequestration of RNA-binding regulatory proteins." Curr Opin Microbiol **10**(2): 156-163.

Backofen, R. (2014). Computational Prediction of RNA–RNA Interactions. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. J. Gorodkin and L. W. Ruzzo. Totowa, NJ, Humana Press**:** 417-435.

Bahar, I. and R. L. Jernigan (1998). "Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms1." Journal of Molecular Biology **281**(5): 871-884.

Baker, C. S., I. Morozov, K. Suzuki, T. Romeo and P. Babitzke (2002). "CsrA regulates glycogen biosynthesis by preventing translation of glgC in Escherichia coli." Mol Microbiol **44**(6): 1599-1610.

Beaudry, A. A. and G. F. Joyce (1990). "Minimum secondary structure requirements for catalytic activity of a self-splicing group I intron." Biochemistry **29**(27): 6534-6539.

Beisel, C. L. and G. Storz (2011). "The base pairing RNA Spot 42 participates in a multi-output feedforward loop to help enact catabolite repression in Escherichia coli." Molecular cell **41**(3): 286-297.

Bennett, C. F. and E. E. Swayze (2010). "RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform." Annu Rev Pharmacol Toxicol **50**: 259-293.

Benz-Moy, T. L. and D. Herschlag (2011). "Structure–Function Analysis from the Outside In: Long-Range Tertiary Contacts in RNA Exhibit Distinct Catalytic Roles." Biochemistry **50**(40): 8733-8755.

Bernhart, S. H., U. Muckstein and I. L. Hofacker (2011). "RNA Accessibility in cubic time." Algorithms Mol Biol **6**(1): 3.

Brion, P. and E. Westhof (1997). "Hierarchy and dynamics of RNA folding." Annu Rev Biophys Biomol Struct **26**: 113-137.

Buratti, E., A. F. Muro, M. Giombi, D. Gherbassi, A. Iaconcig and F. E. Baralle (2004). "RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon." Mol Cell Biol **24**(3): 1387-1400.

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

Cech, T. R., S. H. Damberger and R. R. Gutell (1994). "Representation of the secondary and tertiary structure of group I introns." Nat Struct Mol Biol **1**(5): 273-280.

Cech, T. R., A. J. Zaug and P. J. Grabowski (1981). "In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence." Cell **27**(3 Pt 2): 487-496.

Chae, T. U., W. J. Kim, S. Choi, S. J. Park and S. Y. Lee (2015). "Metabolic engineering of Escherichia coli for the production of 1,3-diaminopropane, a three carbon diamine." Sci Rep **5**: 13040.

Chan, J. H., S. Lim and W. S. Wong (2006). "Antisense oligonucleotides: from design to therapeutic application." Clin Exp Pharmacol Physiol **33**(5-6): 533-540.

Chaudhary, A. K., D. Na and E. Y. Lee (2015). "Rapid and high-throughput construction of microbial cell-factories with regulatory noncoding RNAs." Biotechnology Advances **33**(6, Part 1): 914-930.

Chitsaz, H., R. Salari, S. C. Sahinalp and R. Backofen (2009). "A partition function algorithm for interacting nucleic acid strands." Bioinformatics **25**(12): i365-i373.

Cho, S. H., K. Haning and L. M. Contreras (2015). "Strain engineering via regulatory noncoding RNAs: not a one-blueprint-fits-all." Current Opinion in Chemical Engineering **10**: 25-34.

Cho, S. H., R. Lei, T. D. Henninger and L. M. Contreras (2014). "Discovery of Ethanol-Responsive Small RNAs in Zymomonas mobilis." Applied and Environmental Microbiology **80**(14): 4189-4198.

Cho, S. H., R. Lei, T. D. Henninger and L. M. Contreras (2014). "Discovery of ethanol responsive small RNAs in Zymomonas mobilis." Applied and Environmental Microbiology.

Coleman, J., P. J. Green and M. Inouye (1984). "The use of RNAs complementary to specific mRNAs to regulate the expression of individual bacterial genes." Cell **37**(2): 429-436.

Contreras, L. M., T. Huang, C. L. Piazza, D. Smith, G. Qu, G. Gelderman, J. P. Potratz, R. Russell and M. Belfort (2013). "Group II intron-ribosome association protects intron RNA from degradation." RNA **19**(11): 1497-1509.

Contreras, L. M., T. Huang, C. L. Piazza, D. Smith, G. Qu, G. Gelderman, J. P. Potratz, R. Russell and M. Belfort (2013). "Group II intron–ribosome association protects intron RNA from degradation." RNA.

Cray, J. A., A. Stevenson, P. Ball, S. B. Bankar, E. C. A. Eleutherio, T. C. Ezeji, R. S. Singhal, J. M. Thevelein, D. J. Timson and J. E. Hallsworth (2015). "Chaotropicity: a key factor in product tolerance of biofuel-producing microorganisms." Current Opinion in Biotechnology **33**: 228-259.

Cruz, J. A. and E. Westhof "The Dynamic Landscapes of RNA Architecture." Cell **136**(4): 604-609.

Cui, X., M. Matsuura, Q. Wang, H. Ma and A. M. Lambowitz (2004). "A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing." J Mol Biol **340**(2): 211-231.

Das, R., L. W. Kwok, I. S. Millett, Y. Bai, T. T. Mills, J. Jacob, G. S. Maskel, S. Seifert, S. G. J. Mochrie, P. Thiyagarajan, S. Doniach, L. Pollack and D. Herschlag (2003). "The Fastest Global Events in RNA Folding: Electrostatic Relaxation and Tertiary Collapse of the Tetrahymena Ribozyme." Journal of Molecular Biology **332**(2): 311-319.

DiChiacchio, L., M. F. Sloma and D. H. Mathews (2016). "AccessFold: predicting RNA-RNA interactions with consideration for competing self-structure." Bioinformatics **32**(7): 1033-1039.

DiChiara, J. M., L. M. Contreras-Martinez, J. Livny, D. Smith, K. A. McDonough and M. Belfort (2010). "Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis." Nucleic acids research **38**(12): 4067-4078.

Dimastrogiovanni, D., K. S. Frohlich, K. J. Bandyra, H. A. Bruce, S. Hohensee, J. Vogel and B. F. Luisi (2014). "Recognition of the small regulatory RNA RydC by the bacterial Hfq protein." Elife **3**.

Ding, Y., C. Y. Chan and C. E. Lawrence (2004). "Sfold web server for statistical folding and rational design of nucleic acids." Nucleic Acids Res **32**(Web Server issue): W135-141.

Ding, Y., C. Y. Chan and C. E. Lawrence (2005). "RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble." RNA **11**(8): 1157-1166.

Ding, Y. and C. E. Lawrence (2001). "Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond." Nucleic Acids Research **29**(5): 1034-1046.

Ding, Y., Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua and S. M. Assmann (2014). "In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features." Nature **505**(7485): 696-700.

Dirks, R. M., J. S. Bois, J. M. Schaeffer, E. Winfree and N. A. Pierce (2007). "Thermodynamic Analysis of Interacting Nucleic Acid Strands." SIAM Review **49**(1): 65-88.

Doherty, E. A. and J. A. Doudna (1997). "The P4-P6 domain directs higher order folding of the Tetrahymena ribozyme core." Biochemistry **36**(11): 3159-3169.

Duan, S., D. H. Mathews and D. H. Turner (2006). "Interpreting oligonucleotide microarray data to determine RNA secondary structure: application to the 3' end of Bombyx mori R2 RNA." Biochemistry **45**(32): 9819-9832.

Edwards, A., A. Garst and R. Batey (2009). Determining Structures of RNA Aptamers and Riboswitches by X-Ray Crystallography. Nucleic Acid and Peptide Aptamers. G. Mayer, Humana Press. **535:** 135-163.

Emerick, V. L. and S. A. Woodson (1993). "Self-splicing of the Tetrahymena pre-rRNA is decreased by misfolding during transcription." Biochemistry **32**(50): 14062-14067.

Engler, C. and S. Marillonnet (2014). Golden Gate Cloning. DNA Cloning and Assembly Methods. S. Valla and R. Lale. Totowa, NJ, Humana Press**:** 119-131.

Faoro, C. and S. F. Ataide (2014). "Ribonomic approaches to study the RNA-binding proteome." FEBS Lett **588**(20): 3649-3664.

Franch, T., M. Petersen, E. G. Wagner, J. P. Jacobsen and K. Gerdes (1999). "Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure." J Mol Biol **294**(5): 1115-1125.

Frith, M. C., N. F. W. Saunders, B. Kobe and T. L. Bailey (2008). "Discovering Sequence Motifs with Arbitrary Insertions and Deletions." PLOS Computational Biology **4**(5): e1000071.

Frommer, J., B. Appel and S. Muller (2015). "Ribozymes that can be regulated by external stimuli." Curr Opin Biotechnol **31**: 35-41.

Georg, J. and W. R. Hess (2011). "cis-antisense RNA, another level of gene regulation in bacteria." Microbiol Mol Biol Rev **75**(2): 286-300.

Gerlach, W. and R. Giegerich (2006). "GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing." Bioinformatics **22**(6): 762-764.

Gibson, D. G. (2011). "Enzymatic assembly of overlapping DNA fragments." Methods Enzymol **498**: 349-361.

Golden, B. L., A. R. Gooding, E. R. Podell and T. R. Cech (1998). "A Preorganized Active Site in the Crystal Structure of the Tetrahymena Ribozyme." Science **282**(5387): 259-264.

Gorodkin, J. and W. L. Ruzzo (2014). RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods, Humana Press.

Gottesman, S. (2004). "The small RNA regulators of Escherichia coli: roles and mechanisms*." Annu Rev Microbiol **58**: 303-328.

Grohman, J. K., R. J. Gorelick, S. Kottegoda, N. L. Allbritton, A. Rein and K. M. Weeks (2014). "An Immature Retroviral RNA Genome Resembles a Kinetically Trapped Intermediate State." Journal of Virology **88**(11): 6061-6068.

Grohman, J. K., R. J. Gorelick, S. Kottegoda, N. L. Allbritton, A. Rein and K. M. Weeks (2014). "An immature retroviral RNA genome resembles a kinetically trapped intermediate state." Journal of Virology.

Gruber, A. R., R. Lorenz, S. H. Bernhart, R. Neuböck and I. L. Hofacker (2008). "The Vienna RNA Websuite." Nucleic Acids Research **36**(suppl 2): W70-W74.

Gu, W., Y. Xu, X. Xie, T. Wang, J.-H. Ko and T. Zhou (2014). "The role of RNA structure at 5□ untranslated region in microRNA-mediated gene regulation." RNA **20**(9): 1369-1375.

Haning, K., S. H. Cho and L. M. Contreras (2014). "Small RNAs in mycobacteria: an unfolding story." Frontiers in Cellular and Infection Microbiology **4**.

Holmqvist, E., C. Unoson, J. Reimegård and E. G. H. Wagner (2012). "A mixed double negative feedback loop between the sRNA MicF and the global regulator Lrp." Molecular Microbiology **84**(3): 414-427.

Holmqvist, E., P. R. Wright, L. Li, T. Bischler, L. Barquist, R. Reinhardt, R. Backofen and J. Vogel (2016). "Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo." EMBO J **35**(9): 991-1011.

Hoynes-O'Connor, A. and T. S. Moon (2016). "Development of Design Rules for Reliable Antisense RNA Behavior in E. coli." ACS Synthetic Biology.

Huang, F. W. D., J. Qin, C. M. Reidys and P. F. Stadler (2009). "Partition function and base pairing probabilities for RNA–RNA interaction prediction." Bioinformatics **25**(20): 2646-2654.

Ikawa, Y., T. Yoshimura, H. Hara, H. Shiraishi and T. Inoue (2002). "Two conserved structural components, A-rich bulge and P4 XJ6/7 base-triples, in activating the group I ribozymes." Genes to cells : devoted to molecular & cellular mechanisms **7**(12): 1205-1215.

Ikawa, Y., W. Yoshioka, Y. Ohki, H. Shiraishi and T. Inoue (2001). "Self-splicing of the Tetrahymena group I ribozyme without conserved base-triples." Genes to Cells **6**(5): 411-420.

Ingram, L. O. (1989). "Ethanol Tolerance in Bacteria." Critical Reviews in Biotechnology **9**(4): 305-319.

Isaacs, F. J., D. J. Dwyer, C. Ding, D. D. Pervouchine, C. R. Cantor and J. J. Collins (2004). "Engineered riboregulators enable post-transcriptional control of gene expression." Nat Biotech **22**(7): 841-847.

Ishikawa, H., H. Otaka, K. Maki, T. Morita and H. Aiba (2012). "The functional Hfq-binding module of bacterial sRNAs consists of a double or single hairpin preceded by a U-rich sequence and followed by a 3□ poly(U) tail." RNA **18**(5): 1062-1074.

Jaeger, L., F. Michel and E. Westhof (1997). The Structure of Group I Ribozymes. Catalytic RNA. F. Eckstein and D. M. J. Lilley. Berlin, Heidelberg, Springer Berlin Heidelberg**:** 33-51.

Johnson, D. S. (1974). "Approximation algorithms for combinatorial problems." Journal of computer and system sciences **9**(3): 256-278.

Kast, P. (1994). "pKSS--a second-generation general purpose cloning vector for efficient positive selection of recombinant clones." Gene **138**(1-2): 109-114.

Kast, P. and H. Hennecke (1991). "Amino acid substrate specificity of Escherichia coli phenylalanyl-tRNA synthetase altered by distinct mutations." J Mol Biol **222**(1): 99-124.

Kertesz, M., Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang and E. Segal (2010). "Genome-wide measurement of RNA secondary structure in yeast." Nature **467**(7311): 103-107.

Kertesz, M., Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang and E. Segal (2010). "Genome-wide measurement of RNA secondary structure in yeast." Nature **467**(7311): 103-107.

Kieft, J. S. and I. Tinoco (1997). "Solution structure of a metal-binding site in the major groove of RNA complexed with cobalt (III) hexammine." Structure (London, England : 1993) **5**(5): 713-721.

Koduvayur, S. P. and S. A. Woodson (2004). "Intracellular folding of the Tetrahymena group I intron depends on exon sequence and promoter choice." RNA **10**(10): 1526-1532.

Kozak, M. (2005). "Regulation of translation via mRNA structure in prokaryotes and eukaryotes." Gene **361**: 13-37.

Kruger, K., P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling and T. R. Cech (1982). "Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena." Cell **31**(1): 147-157.

Lai, D., J. R. Proctor and I. M. Meyer (2013). "On the importance of cotranscriptional RNA structure formation." RNA **19**(11): 1461-1473.

Lapouge, K., R. Perozzo, J. Iwaszkiewicz, C. Bertelli, V. Zoete, O. Michielin, L. Scapozza and D. Haas (2013). "RNA pentaloop structures as effective targets of regulators belonging to the RsmA/CsrA protein family." RNA Biol **10**(6): 1031-1041.

Leamy, K. A., S. M. Assmann, D. H. Mathews and P. C. Bevilacqua (2016). "Bridging the gap between in vitro and in vivo RNA folding." Quarterly Reviews of Biophysics **49**: e10 (26 pages).

Lease, R. A., D. Smith, K. McDonough and M. Belfort (2004). "The Small Noncoding DsrA RNA Is an Acid Resistance Regulator in Escherichia coli." Journal of Bacteriology **186**(18): 6179-6185.

Lehnert, V., L. Jaeger, F. Michele and E. Westhof (1996). "New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme." Chemistry & Biology **3**(12): 993-1009.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Li, L., D. Huang, M. K. Cheung, W. Nong, Q. Huang and H. S. Kwan (2013). "BSRD: a repository for bacterial small regulatory RNA." Nucleic Acids Res **41**(Database issue): D233-238.

Li, X., J. Song and C. Yi (2014). "Genome-wide Mapping of Cellular Protein–RNA Interactions Enabled by Chemical Crosslinking." Genomics, Proteomics & Bioinformatics **12**(2): 72-78.

Li, Y., H. Liu and W. Powell (2015). "The knowledge gradient policy using a sparse additive belief model." arXiv preprint arXiv:1503.05567.

Li, Y., H. Liu and W. Powell (2016). A Lasso-based Sparse Knowledge Gradient Policy for Sequential Optimal Learning. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics.

Li, Y., K. G. Reyes, J. Vazquez-Anderson, Y. Wang, L. M. Contreras and W. B. Powell (2015). "A Knowledge Gradient Policy for Sequencing Experiments to Identify the Structure of RNA Molecules Using a Sparse Additive Belief Model." arXiv preprint arXiv:1508.01551.

Lindell, M., P. Romby and E. G. H. Wagner (2002). "Lead(II) as a probe for investigating RNA structure in vivo." RNA **8**(4): 534-541.

Liu, C. C., L. Qi, J. B. Lucks, T. H. Segall-Shapiro, D. Wang, V. K. Mutalik and A. P. Arkin (2012). "An adaptor from translational to transcriptional control enables predictable assembly of complex regulation." Nat Meth **9**(11): 1088-1094.

Liu, J. M., J. Livny, M. S. Lawrence, M. D. Kimball, M. K. Waldor and A. Camilli (2009). "Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing." Nucleic Acids Res **37**(6): e46.

Lorenz, R., M. T. Wolfinger, A. Tanzer and I. L. Hofacker (2016). "Predicting RNA secondary structures from sequence and probing data." Methods **103**: 86-98.

Lu, Z. J. and D. H. Mathews (2008). "Efficient siRNA selection using hybridization thermodynamics." Nucleic Acids Res **36**(2): 640-647.

Lucks, J. B., L. Qi, V. K. Mutalik, D. Wang and A. P. Arkin (2011). "Versatile RNA-sensing transcriptional regulators for engineering genetic networks." Proceedings of the National Academy of Sciences **108**(21): 8617-8622.

Lutz, R. and H. Bujard (1997). "Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements." Nucleic Acids Res **25**(6): 1203-1210.

Mathews, D. H., M. E. Burkard, S. M. Freier, J. R. Wyatt and D. H. Turner (1999). "Predicting oligonucleotide affinity to nucleic acid targets." RNA **5**(11): 1458-1469.

Mathews, D. H., J. Sabina, M. Zuker and D. H. Turner (1999). "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." J Mol Biol **288**(5): 911-940.

Melamed, S., A. Peer, R. Faigenbaum-Romm, Y. E. Gatt, N. Reiss, A. Bar, Y. Altuvia, L. Argaman and H. Margalit (2016). "Global Mapping of Small RNA-Target Interactions in Bacteria." Mol Cell **63**(5): 884-897.

Memczak, S., M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble and N. Rajewsky (2013). "Circular RNAs are a large class of animal RNAs with regulatory potency." Nature **495**(7441): 333-338.

Mitchell, D., 3rd, I. Jarmoskaite, N. Seval, S. Seifert and R. Russell (2013). "The long-range P3 helix of the Tetrahymena ribozyme is disrupted during folding between the native and misfolded conformations." J Mol Biol **425**(15): 2670-2686.

Mitchell III, D. and R. Russell (2014). "Folding Pathways of the Tetrahymena Ribozyme." Journal of Molecular Biology **426**(12): 2300-2312.

Mitra, S., I. V. Shcherbakova, R. B. Altman, M. Brenowitz and A. Laederach (2008). "High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis." Nucleic Acids Res **36**(11): e63.

Muckstein, U., H. Tafer, J. Hackermuller, S. H. Bernhart, P. F. Stadler and I. L. Hofacker (2006). "Thermodynamics of RNA-RNA binding." Bioinformatics **22**(10): 1177-1182.

Naito, Y., H. Shiraishi and T. Inoue (1998). "P5abc of the Tetrahymena ribozyme consists of three functionally independent elements." RNA **4**(7): 837-846.

Nakashima, N. and K. Miyazaki (2014). "Bacterial Cellular Engineering by Genome Editing and Gene Silencing." International Journal of Molecular Sciences **15**(2): 2773-2793.

Nakashima, N. and T. Tamura (2009). "Conditional gene silencing of multiple genes with antisense RNAs and generation of a mutator strain of Escherichia coli." Nucleic Acids Research **37**(15): e103-e103.

Nguyen, T. C., X. Cao, P. Yu, S. Xiao, J. Lu, F. H. Biase, B. Sridhar, N. Huang, K. Zhang and S. Zhong (2016). "Mapping RNA-RNA interactome and RNA structure in vivo by MARIO." Nat Commun **7**.

Paige, J. S., K. Y. Wu and S. R. Jaffrey (2011). "RNA mimics of green fluorescent protein." Science **333**(6042): 642-646.

Pan, J. and S. A. Woodson (1998). "Folding intermediates of a self-splicing RNA: mispairing of the catalytic core." J Mol Biol **280**(4): 597-609.

Pan, J. and S. A. Woodson (1999). "The effect of long-range loop-loop interactions on folding of the Tetrahymena self-splicing RNA." Journal of Molecular Biology **294**(4): 955-965.

Peer, A. and H. Margalit (2011). "Accessibility and Evolutionary Conservation Mark Bacterial Small-RNA Target-Binding Regions." Journal of Bacteriology **193**(7): 1690-1701.

Pervouchine, D. D. (2004). "IRIS: intermolecular RNA interaction search." Genome Inform **15**(2): 92-101.

Ponchon, L. and F. Dardel (2007). "Recombinant RNA technology: the tRNA scaffold." Nat Meth **4**(7): 571-576.

Rangan, P., B. Masquida, E. Westhof and S. A. Woodson (2003). "Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme." Proceedings of the National Academy of Sciences **100**(4): 1574-1579.

Rehmsmeier, M., P. Steffen, M. Hochsmann and R. Giegerich (2004). "Fast and effective prediction of microRNA/target duplexes." RNA **10**(10): 1507-1517.

Reuter, J. and D. Mathews (2010). "RNAstructure: software for RNA secondary structure prediction and analysis." BMC Bioinformatics **11**(1): 129.

Rodrigo, G., T. E. Landrain and A. Jaramillo (2012). "De novo automated design of small RNA circuits for engineering synthetic riboregulation in living cells." Proceedings of the National Academy of Sciences **109**(38): 15271-15276.

Romby, P. and E. Charpentier (2010). "An overview of RNAs with regulatory functions in gram-positive bacteria." Cellular and Molecular Life Sciences **67**(2): 217-237.

Romeo, T., C. A. Vakulskas and P. Babitzke (2013). "Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems." Environ Microbiol **15**(2): 313-324.

Russell, R., R. Das, H. Suh, K. Travers, A. Laederach, M. Engelhardt and D. Herschlag (2006). "The paradoxical behavior of a highly structured misfolded intermediate in RNA folding." Journal of molecular biology **363**(2): 531-544.

Russell, R., R. Das, H. Suh, K. J. Travers, A. Laederach, M. A. Engelhardt and D. Herschlag (2006). "The paradoxical behavior of a highly structured misfolded intermediate in RNA folding." J Mol Biol **363**(2): 531-544.

Russell, R., X. Zhuang, H. P. Babcock, I. S. Millett, S. Doniach, S. Chu and D. Herschlag (2002). "Exploring the folding landscape of a structured RNA." Proceedings of the National Academy of Sciences **99**(1): 155-160.

Said, N., R. Rieder, R. Hurwitz, J. Deckert, H. Urlaub and J. Vogel (2009). "In vivo expression and purification of aptamer-tagged small RNA regulators." Nucleic Acids Res **37**(20): e133.

Schroeder, R., R. Grossberger, A. Pichler and C. Waldsich (2002). "RNA folding in vivo." Current Opinion in Structural Biology **12**(3): 296-300.

Scott, L. and M. Hennig (2008). RNA Structure Determination by NMR. Bioinformatics. J. Keith, Humana Press. **452:** 29-61.

Shao, Y., C. Y. Chan, A. Maliyekkel, C. E. Lawrence, I. B. Roninson and Y. Ding (2007). "Effect of target secondary structure on RNAi efficiency." RNA **13**(10): 1631-1640.

Shao, Y., Y. Wu, C. Y. Chan, K. McDonough and Y. Ding (2006). "Rational design and rapid screening of antisense oligonucleotides for prokaryotic gene modulation." Nucleic Acids Research **34**(19): 5660-5669.

Sharp, P. A. "The Centrality of RNA." Cell **136**(4): 577-580.

Shcherbakova, I. and M. Brenowitz (2008). "Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting." Nat. Protocols **3**(2): 288-302.

Shtatland, T., S. C. Gill, B. E. Javornik, H. E. Johansson, B. S. Singer, O. C. Uhlenbeck, D. A. Zichi and L. Gold (2000). "Interactions of Escherichia coli RNA with bacteriophage MS2 coat protein: genomic SELEX." Nucleic Acids Res **28**(21): E93.

Silverman, I. M., N. D. Berkowitz, S. J. Gosai and B. D. Gregory (2016). "Genome-Wide Approaches for RNA Structure Probing." Adv Exp Med Biol **907**: 29-59.

Smith, B. R. and R. Schleif (1978). "Nucleotide sequence of the L-arabinose regulatory region of Escherichia coli K12." J Biol Chem **253**(19): 6931-6933.

Sowa, S. W., J. Vazquez-Anderson, C. A. Clark, R. De La Pena, K. Dunn, E. K. Fung, M. J. Khoury and L. M. Contreras (2015). "Exploiting post-transcriptional regulation to probe RNA structures in vivo via fluorescence." Nucleic Acids Res **43**(2): e13.

Spitale, R. C., P. Crisalli, R. A. Flynn, E. A. Torre, E. T. Kool and H. Y. Chang (2013). "RNA SHAPE analysis in living cells." Nat Chem Biol **9**(1): 18-20.

Srisawat, C. and D. R. Engelke (2002). "RNA affinity tags for purification of RNAs and ribonucleoprotein complexes." Methods **26**(2): 156-161.

Strack, R. L., M. D. Disney and S. R. Jaffrey (2013). "A superfolding Spinach2 reveals the dynamic nature of trinucleotide repeat-containing RNA." Nat Meth **10**(12): 1219-1224.

Strobel, E. J., K. E. Watters, D. Loughrey and J. B. Lucks (2016). "RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs." Current Opinion in Biotechnology **39**: 182-191.

Strobel, S. A. and K. Shetty (1997). "Defining the chemical groups essential for Tetrahymena group I intron function by nucleotide analog interference□mapping." Proceedings of the National Academy of Sciences **94**(7): 2903-2908.

Tafer, H. (2014). Bioinformatics of siRNA Design. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. J. Gorodkin and L. W. Ruzzo. Totowa, NJ, Humana Press**:** 477-490.

Tafer, H. and I. L. Hofacker (2008). "RNAplex: a fast tool for RNA-RNA interaction search." Bioinformatics **24**(22): 2657-2663.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Tijerina, P., S. Mohr and R. Russell (2007). "DMS footprinting of structured RNAs and RNA-protein complexes." Nat. Protocols **2**(10): 2608-2623.

Tjaden, B., S. S. Goodwin, J. A. Opdyke, M. Guillier, D. X. Fu, S. Gottesman and G. Storz (2006). "Target prediction for small, noncoding RNAs in bacteria." Nucleic Acids Res **34**(9): 2791-2802.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nat Protoc **7**(3): 562-578.

Vakulskas, C. A., Y. Leng, H. Abe, T. Amaki, A. Okayama, P. Babitzke, K. Suzuki and T. Romeo (2016). "Antagonistic control of the turnover pathway for the global regulatory sRNA CsrB by the CsrA and CsrD proteins." Nucleic Acids Res.

Vazquez-Anderson, J. and L. M. Contreras (2013). "Regulatory RNAs: Charming gene management styles for synthetic biology applications." RNA Biology **10**(12): 1778-1797.

Vazquez-Anderson, J. and L. M. Contreras (2013). "Regulatory RNAs: charming gene management styles for synthetic biology applications." RNA Biol **10**(12): 1778-1797.

Vickers, T. A., J. R. Wyatt and S. M. Freier (2000). "Effects of RNA secondary structure on cellular antisense activity." Nucleic Acids Res **28**(6): 1340-1347.

Voter, A. F. (2007). INTRODUCTION TO THE KINETIC MONTE CARLO METHOD. Radiation Effects in Solids. K. E. Sickafus, E. A. Kotomin and B. P. Uberuaga. Dordrecht, Springer Netherlands**:** 1-23.

Waldsich, C., R. Grossberger and R. Schroeder (2002). "RNA chaperone StpA loosens interactions of the tertiary structure in the td group I intron in vivo." Genes & Development **16**(17): 2300-2312.

Wan, Y., M. Kertesz, R. C. Spitale, E. Segal and H. Y. Chang (2011). "Understanding the transcriptome through RNA structure." Nat Rev Genet **12**(9): 641-655.

Wan, Y., K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal and H. Y. Chang (2014). "Landscape and variation of RNA secondary structure across the human transcriptome." Nature **505**(7485): 706-709.

Wan, Y., H. Suh, R. Russell and D. Herschlag (2010). "Multiple Unfolding Events during Native Folding of the Tetrahymena Group I Ribozyme." Journal of Molecular Biology **400**(5): 1067-1077.

Wang, J., T. Liu, B. Zhao, Q. Lu, Z. Wang, Y. Cao and W. Li (2015). "sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria." Nucleic Acids Research.

Waring, R. B., J. A. Ray, S. W. Edwards, C. Scazzocchio and R. W. Davies (1985). "The tetrahymena rRNA intron self-splices in E. coli: In vivo evidence for the importance of key base-paired regions of RNA for RNA enzyme function." Cell **40**(2): 371-380.

Warner, K. D., M. C. Chen, W. Song, R. L. Strack, A. Thorn, S. R. Jaffrey and A. R. Ferré-D'Amaré (2014). "Structural basis for activity of highly efficient RNA mimics of green fluorescent protein." Nat Struct Mol Biol **21**(8): 658-663.

Wassarman, K. M. (2002). "Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes." Cell **109**(2): 141-144.

Wells, S. E., J. M. Hughes, A. H. Igel and M. Ares, Jr. (2000). "Use of dimethyl sulfate to probe RNA structure in vivo." Methods Enzymol **318**: 479-493.

Wilson, D. N., S. Arenz and R. Beckmann (2016). "Translation regulation via nascent polypeptide-mediated ribosome stalling." Current Opinion in Structural Biology **37**: 123-133.

Woodson, S. A. (2005). "Metal ions and RNA folding: a highly charged topic with a dynamic future." Current opinion in chemical biology **9**(2): 104-109.

Woodson, S. A. (2005). "Structure and assembly of group I introns." Curr Opin Struct Biol **15**(3): 324-330.

Wuchty, S., W. Fontana, I. L. Hofacker and P. Schuster (1999). "Complete suboptimal folding of RNA and the stability of secondary structures." Biopolymers **49**(2): 145-165.

Wurst, R. M., J. N. Vournakis and A. M. Maxam (1978). "Structure mapping of 5'-32P-labeled RNA with S1 nuclease." Biochemistry **17**(21): 4493-4499.

Xia, T., J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox and D. H. Turner (1998). "Thermodynamic parameters for an expanded nearest-neighbor

model for formation of RNA duplexes with Watson-Crick base pairs." Biochemistry **37**(42): 14719-14735.

Xue, Y., B. Gracia, D. Herschlag, R. Russell and H. M. Al-Hashimi (2016). "Visualizing the formation of an RNA folding intermediate through a fast highly modular secondary structure switch." Nat Commun **7**: ncomms11768.

Yoo, S. M., D. Na and S. Y. Lee (2013). "Design and use of synthetic regulatory small RNAs to control gene expression in Escherichia coli." Nat Protoc **8**(9): 1694-1707.

Zadeh, J. N., C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks and N. A. Pierce (2011). "NUPACK: Analysis and design of nucleic acid systems." Journal of Computational Chemistry **32**(1): 170-173.

Zadeh, J. N., C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks and N. A. Pierce (2011). "NUPACK: Analysis and design of nucleic acid systems." J Comput Chem **32**(1): 170-173.

Zarrinkar, P. and J. Williamson (1996). "The P9.1-P9.2 peripheral extension helps guide folding of the Tetrahymena ribozyme." Nucleic acids research **24**(5): 854-858.

Zarrinkar, P. P. and J. R. Williamson (1994). "Kinetic intermediates in RNA folding." Science **265**(5174): 918-924.

Zemora, G. and C. Waldsich (2010). "RNA folding in living cells." RNA Biology **7**(6): 634-641.

Zhang, F., E. S. Ramsay and S. A. Woodson (1995). "In vivo facilitation of Tetrahymena group I intron splicing in Escherichia coli pre-ribosomal RNA." RNA **1**(3): 284-292.

Zhang, M., C. Eddy, K. Deanda, M. Finkelstein and S. Picataggio (1995). "Metabolic Engineering of a Pentose Metabolism Pathway in Ethanologenic Zymomonas mobilis." Science **267**(5195): 240-243.

Zhao, J. J. and G. Lemke (1998). "Rules for ribozymes." Mol Cell Neurosci **11**(1-2): 92-97.

Zou, S.-l., K. Zhang, L. You, X.-m. Zhao, X. Jing and M.-h. Zhang (2012). "Enhanced electrotransformation of the ethanologen Zymomonas mobilis ZM4 with plasmids." Engineering in Life Sciences **12**(2): 152-161.

Zwanzig, R. (2001). Nonequilibrium Statistical Mechanics, Oxford University Press.