THE FUNCTIONAL EFFECTS OF RNA STRUCTURE:
FROM RIBOSNITCHES TO TRANSLATIONAL CONTROL

Meredith Corley

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2017

Approved by:
Alain Laederach
Corbin Jones
Nathaniel Moorman
Praveen Sethupathy
William Valdar
ABSTRACT

Meredith Corley: The functional effects of RNA structure: from riboSNitches to translational control
(Under the direction of Alain Laederach)

Ribonucleic Acid (RNA) is a nucleotide polymer that, like Deoxyribonucleic acid (DNA), has an essential role in the cell. RNA molecules are structurally distinct from DNA in the diversity of structures they adopt, due to stable intramolecular interactions. There exist a few well-defined cases of functional RNA structures, but many classes of RNA, including mRNA, adopt more flexible structures that are poorly characterized and unlinked to biological function. Accurate structure determination is thus essential to the study of RNAs. Much work involving RNA structure relies on computational prediction of RNA secondary structures. Here I briefly summarize RNA structure prediction and how genetic variation can alter RNA structure. By benchmarking algorithms that detect single nucleotide variant-induced structural change, we show that considering the full set of structures an RNA may adopt is crucial for the most accurate predictions. Conversely, constraining computational prediction with experimental structure probing data has been shown to greatly improve single-structure predictions. Thus incorporating structure probing data like 2′-hydroxyl acylation analyzed by primer extension mutational profiling (SHAPE-MaP) is an alternative approach for modeling structural features of RNAs. In order to explore structure-function relationships in a model RNA we gathered SHAPE-MaP structural data on a set of mRNAs derived from the human gene SERPINA1. Here I discuss the effect structure may have on mRNAs, especially during protein translation. We show with SHAPE-MaP constrained structure prediction that RNA structure has a role in determining SERPINA1 protein translation efficiency and that this effect can be quantitatively modeled.
To my sister Olivia, who moved all the way out to North Carolina to keep me company during my PhD.
ACKNOWLEDGEMENTS

There are many people and programs that led me to this point. The Freshman Research Initiative at The University of Texas at Austin introduced me to research. Without this beginning, who knows what I would be doing now. John Stanton, Brad Hall, Scott Hunicke-Smith and Andy Ellington are all people that crucially encouraged me to continue research as an undergraduate. For my graduate years I would like to thank Corbin Jones and Shehzad Sheikh for my rotations with them. I am very grateful for the Curriculum in Bioinformatics and Computational Biology, which gave me the confidence to go into bioinformatics. For my current work, everyone in the Laederach lab has been instrumental: Amanda Solem, Lela Lackey, Katrina Kutchko, Chanin Tolson, Aaztli Coria, Matt Halvorsen, Cai Chen and of course Alain Laederach. Last but not least I would like to thank my committee members Praveen Sethupathy, Nathaniel Moorman, Corbin Jones and Will Valdar for their time and advice. Wish me luck.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 RNA chemical makeup</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Base pairing</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 DNA-RNA relationship</td>
<td>3</td>
</tr>
<tr>
<td>1.2 RNA structure</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1 Secondary structure prediction</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 Secondary structure probing</td>
<td>6</td>
</tr>
<tr>
<td>1.2.3 Secondary structure change</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Messenger RNA</td>
<td>9</td>
</tr>
<tr>
<td>1.3.1 Translation</td>
<td>10</td>
</tr>
<tr>
<td>1.3.2 Secondary structure in mRNAs</td>
<td>12</td>
</tr>
<tr>
<td>2 Detecting riboSNitches with RNA folding algorithms: A genome-wide benchmark</td>
<td>14</td>
</tr>
<tr>
<td>2.1 Overview</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Background</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Results and Discussion</td>
<td>16</td>
</tr>
<tr>
<td>2.3.1 Experimental benchmark criteria</td>
<td>16</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Algorithmic performance analysis</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Improving riboSNitch prediction performance</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Discussion</td>
</tr>
<tr>
<td>2.4</td>
<td>Materials and Methods</td>
</tr>
<tr>
<td>2.4.1</td>
<td>The benchmark dataset</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Benchmark design and distance metrics</td>
</tr>
<tr>
<td>2.4.3</td>
<td>ROC analysis</td>
</tr>
<tr>
<td>3</td>
<td>An RNA structure-mediated, post-transcriptional model of human α-1-antitrypsin expression</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview</td>
</tr>
<tr>
<td>3.2</td>
<td>Background</td>
</tr>
<tr>
<td>3.3</td>
<td>Results and Discussion</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Transcript complexity in SERPINA1</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Translation efficiency analysis</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Modeling translation efficiency</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Secondary structure of SERPINA1 transcripts</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Modeling translation efficiency with structure</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Modeling α-1-antitrypsin expression in tissue</td>
</tr>
<tr>
<td>3.3.7</td>
<td>Discussion</td>
</tr>
<tr>
<td>3.4</td>
<td>Materials and Methods</td>
</tr>
<tr>
<td>3.4.1</td>
<td>SERPINA1 annotation</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Heat map of tissue-specific isoform expression</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Cell line specific transcript expression</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Luciferase assays</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Upstream open reading frame (uORF) mutants</td>
</tr>
<tr>
<td>3.4.6</td>
<td>SHAPE-MaP sequencing and analysis</td>
</tr>
<tr>
<td>3.4.7</td>
<td>Secondary structure analysis</td>
</tr>
<tr>
<td>3.4.8</td>
<td>Models</td>
</tr>
</tbody>
</table>
LIST OF TABLES

2.1 riboSNitch and non-riboSNitch set sizes ........................................ 19
2.2 ROC analysis of riboSNitch prediction, specialized algorithms ............. 23
2.3 ROC analysis of riboSNitch prediction, general algorithms .................. 23
A.1 All ROC analysis AUC values ..................................................... 59
A.2 Best sensitivity and specificity values from ROC analysis ....................... 60
A.3 riboSNitch prediction with alternate distance metric .......................... 60
A.4 Parameters used for each algorithm ............................................. 61
B.1 Model performance and $\Delta G$ window size .................................. 65
# LIST OF FIGURES

1.1 The chemical composition of DNA and RNA ........................................... 2  
1.2 RNA structure example ................................................................. 5  
1.3 Representations of RNA structure ..................................................... 6  
1.4 Enzymatic structure probing with PARS ............................................. 7  
1.5 Chemical structure probing with SHAPE-MaP ...................................... 8  
1.6 A riboSNIch .......................................................... 10  
1.7 Transcription through translation .................................................... 11  
1.8 The Leaky Scanning mechanism of translation ..................................... 12  

2.1 The PARS riboSNIch dataset ......................................................... 18  
2.2 Predicting a riboSNIch ............................................................ 20  
2.3 RiboSNIch prediction ROC curves ............................................... 22  
2.4 n% tails in riboSNIch score distributions ..................................... 25  
2.5 Example improvements with the n% tails strategy ............................. 26  

3.1 The SERPINA1 transcripts ......................................................... 38  
3.2 Translation efficiency of SERPINA1 transcripts ............................. 40  
3.3 SERPINA1 SHAPE-MaP data .................................................. 43  
3.4 Modeling translation efficiency .................................................... 44  
3.5 Modeling translation efficiency in tissue ...................................... 46  

A.1 ROC performance with different input sequence lengths ................. 62  
A.2 ROC performance across categories with different input sequence lengths .... 63  
A.3 RiboSNIch distance score distributions .................................. 64  
B.1 Schematic of eleven SERPINA1 transcripts ................................ 65  
B.2 Translation efficiency of SERPINA1 uORF-mutants .......................... 66  
B.3 Correlation coefficients between SHAPE-MaP duplicates ................. 66
B.4 MFE structure around uORFs .................................................. 67
B.5 MFE structures for full SERPINA1 transcripts ........................... 68
B.6 Predicting translation efficiency in human tissues ...................... 68
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Area Under the (ROC) Curve</td>
</tr>
<tr>
<td>BPPM</td>
<td>Base Pairing Probability Matrix</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding Determinant Sequence</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic Obstructive Pulmonary Disease</td>
</tr>
<tr>
<td>ΔG</td>
<td>Gibb’s Free Energy Change</td>
</tr>
<tr>
<td>MFE</td>
<td>Minimum Free Energy</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger Ribonucleic Acid</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>PARS</td>
<td>Parallel Analysis of RNA Structure</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operator Characteristic</td>
</tr>
<tr>
<td>SHAPE</td>
<td>Selective 2’-Hydroxyl Acylation analyzed by Primer Extension</td>
</tr>
<tr>
<td>SHAPE-MaP</td>
<td>Selective 2’Hydroxyl Acylation analyzed by Primer Extension Mutational Profiling</td>
</tr>
<tr>
<td>SNV</td>
<td>Single Nucleotide Variant</td>
</tr>
<tr>
<td>TE</td>
<td>Translation Efficiency</td>
</tr>
<tr>
<td>uORF</td>
<td>Upstream Open Reading Frame</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
</tbody>
</table>
CHAPTER 1
Introduction

1.1 RNA chemical makeup

In this section I will introduce ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) by detailing their composition and chemistry. I will then describe how these molecules function biologically and their relationship with each other. I will subsequently focus on RNA and its distinct biology and function(s). Some scientists believe that all life originated solely with RNA and no DNA [81] and thus perhaps it is unfair to always introduce RNA in the company of DNA. RNA, however, is a much more colorful and diverse molecule and these attributes are amplified by comparison with the more inert DNA.

The essential molecule of living organisms is nucleic acid of two types: ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). Many biologically active molecules, including nucleic acids, form polymers—long molecules built from multiple subunits. The polymer subunit of RNA and DNA is very similar, consisting of a ribofuranose ring attached to a phosphate group, a hydroxyl group and a nitrogen base (Figure 1.1). The ribofuranose ring itself is composed of carbons (as well as an ether moiety that no one ever talks about) which are numbered from 1 through 5. The fifth carbon in the ribofuranose ring is attached to the phosphate group and so this part of the molecule is commonly referred to as the ”5-prime” end. The third carbon in the ribofuranose ring is attached to the hydroxyl group and is referred to as the ”3-prime” end. As trivial as it may seem, the concept of 5-prime (5’) and 3-prime (3’) is referred to constantly in molecular biology and biochemistry. To form a nucleic acid polymer, the phosphate group at the 5’ end of one nucleic acid reacts with the 3’ hydroxyl group of another nucleic acid to form a phospho-diester linkage. This 5’ to 3’ reaction is how nucleic acid polymers are synthesized in the cell, but this reaction can also be harnessed in vitro (in a tube) in a Polymerase Chain Reaction [110] and is essential to modern biological experiments. In the main
chemical difference between RNA and DNA, the second carbon in RNA is attached to an additional hydroxyl group, while this group is absent in DNA.

Figure 1.1: The chemical composition of DNA and RNA. An example DNA molecule is on the left; RNA on the right. The phosphate group, ribofuranose ring and bases are labelled and the 5’ and 3’ ends are indicated. Carbons in the ribofuranose rings are labelled 1-5.

1.1.1 Base pairing

In a polymer of nucleic acids, the ribofuranose rings and their phosphate groups connected with phospho-diesters links form the "backbone" of the molecule, with the nitrogen bases extending away from this backbone. The nitrogen bases are the exciting part of RNA and DNA, as nucleic acids can be attached to one of five different bases. The nitrogen bases in RNA can be either adenine (A), guanine (G), cytosine (C) or uracil (U). In DNA a base can be either adenine (A), guanine (G), cytosine (C) or thymine (T). Since the nitrogen bases are the only chemical moiety that varies between individual nucleic acids in a nucleic acid polymer, polymers can be uniquely identified by the particular sequence of bases they contain. By convention, the sequence of bases is written from the 5’ to the 3’ end of the molecule using the one-letter abbreviation of each base.
For example, the sequence of the DNA strand in Figure 1.1 is 5’-ATGC-3’
DNA usually exists as two complementary DNA molecules entwined together in a double helix structure. Base-on-base hydrogen bond interactions determine the complementarity of DNA molecules and these interactions follow well-known base-pairing rules. ‘A’ pairs with ‘T’ or ‘U’ and ‘C’ pairs with ‘G’. Thus the DNA sequence complementary to the sequence in Figure 1.1 would be 3’-TACG-5’. Crucially, new DNA molecules can be synthesized from a template DNA molecule based on these rules of complementarity. RNA molecules, on the other hand, normally exist as a single strand, but the base-pairing rules still stand, and in the cell RNA molecules are synthesized using DNA as a template.

1.1.2 DNA-RNA relationship

Beyond chemical makeup, DNA and RNA diverge even more, but their roles are very interrelated. Long molecules of DNA collectively make up an organism’s genome. In the case of humans, the genome is composed of 23 pairs of DNA molecules millions of nucleotides in length and totaling over 6 billion nucleotides. The particular sequence of nucleotides in the genome are important. Certain segments of DNA, in the range of thousands to hundreds of thousands of nucleotides, encode genes. A gene is any segment of DNA that is transcribed, which means that the cell uses this segment as a template to synthesize an RNA molecule. The entire set of RNAs that are transcribed from an organism’s genome is the ‘transcriptome’. While DNA is solely used for storing genetic information and normally remains in a fixed location in the cell, the RNA transcripts copied from genes occupy a vast array of different roles and often travel to many locations in the cell. At this point I will focus my attentions on RNA and its distinct structure and function.

1.2 RNA structure

In this section I will describe the importance of RNA structure and how we may detect and predict it. While DNA molecules always form a double-helix by base-pairing with a complementary molecule, RNA molecules usually base-pair with themselves, forming many possible different structures. An RNA molecule interacts with itself (and sometimes other RNAs) through hydrogen
bond interactions between adenine (A) and uracil (U) or between guanine (G) and cytosine (C). These interactions are referred to as ‘base pairs’, where the pairings described above are the canonical ‘Watson-Crick’ pairs. So-called ‘wobble’ base pairs are also possible between guanine and uracil and theoretically, base pairs may also form between any other two bases in non-canonical pairs [103, 116]. Given a sequence of RNA bases, the particular set of base pairs that the sequence adopts is referred to as its secondary structure, and the RNA molecule’s three-dimensional form is referred to as its tertiary structure (Figure 1.2). Sometimes RNA tertiary structures include interactions between secondary substructures, which can be very difficult to detect or predict. Since RNAs form base pairs on a much shorter timescale and interact more strongly than do tertiary contacts [103], the majority of RNA structure studies focus on RNA secondary structure. Because of these differences between secondary and tertiary interactions, it is assumed that RNA secondary structure can be computationally predicted in the absence of tertiary structure information [103]. Thus I will focus on the state of RNA secondary structure research. The structures of RNAs have definite effects on their role in the cell. Famous examples include ribozymes (RNA structural elements that catalyze reactions) [20, 40] and riboswitches (structural elements that change conformation to regulate transcription or translation) [60, 86, 160]. These RNAs represent only two classes of RNAs, but I will argue that every type of RNA is potentially affected by its secondary structure. The secondary of structure of most RNA molecules is unknown, but advances in experimental and computational techniques are expanding our understanding of RNAs and their structures.

1.2.1 Secondary structure prediction

RNA secondary structure prediction requires only primary base sequence and an empirical nearest-neighbor model, which sums the energetic payoffs for base pairs with the energetic costs of unstructured loops of bases [52, 176]. Normally a single RNA molecule is theoretically able to adopt multiple secondary structures, with the number of possible structures increasing exponentially with RNA sequence length [181]. Thus a brute force algorithm that calculates the free energy of every possible structure for a given RNA sequence is quickly overwhelmed. Instead, most RNA structure prediction algorithms focus on the minimum free-energy structure (MFE), which is the structure with the lowest folding free energy ($\Delta G$) that a given RNA sequence can form. Modern programs predict
MFE structures with a dynamic programming algorithm [113] that calculates the lowest free energy of ever larger subsequences of the full sequence, using stored values from smaller subsequences, until completing the full sequence. The algorithm then recursively backtracks to return the structure corresponding to the lowest free energy of each subsequence until returning the full structure. This method depends on the nearest-neighbor model’s main assumption: that each base’s interaction in a secondary structure depends only on the base it pairs with and adjacent base interactions. Of course this assumption is not always true and MFE structure prediction is incomplete: current algorithms correctly predict 60-70% of base pairs in known RNA structures [94].

Considering other ‘suboptimal’ structures a sequence may adopt can provide a more accurate picture of an RNA’s secondary structure. An RNA can be thought of as an ensemble or distribution of structures, where the frequency with which an RNA adopts a particular structure is equal to $e^{-\Delta G_{folding}/RT}$. The folding free energy of alternative structures are calculated by the same dynamic programming algorithm that calculates the MFE, except that the minimum energy is not taken at each pass. The sum of $e^{-\Delta G_{folding}/RT}$ over all structures is referred to as the partition function [105]. Recursive backtracking can count the frequency of each base pair, which once normalized by the partition function is stored as a probability in the base pairing probability matrix (BPPM). The BPPM contains the probability of every pairwise base pair in an RNA. An example BPPM is shown in Figure 1.3(top left) for a small sequence from the gene SERPINA1. Sampled suboptimal secondary
structures are shown as well, where the frequency of each base pair in these structures provides a rough estimate of the probabilities in the BPPM. The SERPINA1 sequence’s MFE structure is also shown in Figure 1.3(bottom right) and can be represented in dot-bracket format or in BPPM format constraining base pairing probabilities to be either 0 or 1.

Figure 1.3: Representations of RNA structure. The base pairing probability matrix (BPPM) shows the probability of any base i pairing with any base j in an RNA. Since the probability of i pairing with j is the same as j pairing with i, only half the matrix need be shown. The base pairing probabilities representing a single structure, like a Minimum Free Energy (MFE) structure (right side), are either 0 or 1. The base pairing probabilities for an ensemble of structures are equivalent to each base pair’s normalized frequency in the structures occurring in the ensemble (left side).

1.2.2 Secondary structure probing

There are many experimental methods for probing RNA structures. X-ray crystallography and NMR are frequently used [1, 66] but are extremely difficult and are feasible only with shorter RNAs. The most versatile RNA structure probing methods make up a class of chemical and enzymatic
probing experiments [38, 71, 95, 149] two of which I will describe here. Parallel analysis of RNA structure, or PARS, uses two enzymes that cleave RNA to measure the ’structuredness’ of RNA samples [71] (Figure 1.4). A given RNA sample is separately treated with each of the enzymes S1 and V1. Enzyme S1 preferentially cleaves regions of RNA molecules that are unpaired and thus single stranded, or ’unstructured’. Enzyme V1 preferentially cleaves regions of RNA molecules that are paired and thus double stranded, or ’structured’. Once treated, the remaining fragments of RNA are processed and sequenced. The sequenced ’read’ data from the samples is computationally aligned to known RNA reference sequence and the number of reads ending at each location in the reference is counted. The position where a read ends indicates a position in an RNA molecule at which one of the enzymes cleaved. The log differential of counts from the S1 and V1 samples indicates the relative structuredness of each nucleotide in the sample RNA molecules. PARS has been used to gather RNA structural information transcriptome wide in both yeast and human samples [71, 170].

![Figure 1.4: Enzymatic structure probing with PARS.](image)

All RNA structure probing techniques rely on reagents that react differently with structured (paired) nucleotides versus unstructured (unpaired) nucleotides. Another technique, 2’-hydroxyl acylation analyzed by primer extension mutational profiling (SHAPE-MaP) [149] uses a chemical reagent rather than the enzymes characteristic of PARS (Figure 1.5). The SHAPE reagent, 1-methyl-
7-nitroisatoic anhydride (1M7), covalently reacts with the 2'-hydroxyl intrinsic to RNA nucleotides, but does so preferentially. 1M7 is most likely to react with nucleotides whose bases are unpaired. As in PARS, after treatment with the reagent the sample RNA molecules are prepared for sequencing. During the reverse transcription step, where DNA is synthesized using the RNA as a template, the 2’ adduct created by reaction with 1M7 induces the addition of an incorrect nucleotide, or mutation. Such mutations are detected after sequencing, where the count of mutations at a given nucleotide is proportional to how receptive the nucleotide is to 1M7, which in turn is proportional to how structured the nucleotide tends to be. The result of a SHAPE-MaP experiment constitutes a structural profile, where every nucleotide in the input RNA has a SHAPE reactivity value.

![SHAPE-MaP diagram](image)

**Figure 1.5:** Chemical structure probing with SHAPE-MaP. In the SHAPE-MaP method, RNA is treated with the SHAPE reagent (1M7, blue triangle), which preferentially reacts with the 2’ hydroxyl of single stranded (unstructured) nucleotides in RNA, forming a bulky 2’ adduct. During reverse transcription of the RNAs, the 2’ adduct induces a mutation in the cDNA, which is detected after sequencing. The frequency of mutations at a given nucleotide is its SHAPE reactivity and is proportional to its structuredness.

While experimental methods like PARS and SHAPE-MaP do not explicitly detect the base pairs that RNAs adopt, they can be very useful for structural profile comparisons between RNA sequence variants or between RNAs under different conditions. Crucially, SHAPE-MaP data can be combined with secondary structure prediction to vastly improve the accuracy of predicted base pairs [34, 149]. To incorporate SHAPE-MaP data, the SHAPE reactivity of each nucleotide is incorporated as a pseudo free energy in nearest-neighbor summations (Equation 1.1). $\Delta G_{SHAPE}(i)$ is the free energy of folding for nucleotide $i$, incorporating its SHAPE reactivity. $m$ and $b$ are empirically determined slope and intercept values, respectively. Commonly used values are 2.6 and -0.8 kcal/mol, respectively [34, 92]. Predicting MFE structures with the aid of SHAPE data increases prediction accuracies to over 90% [34, 149].
\[ \Delta G_{SHAPE}(i) = mln[SHAPEreactivity(i) + 1] + b \] (1.1)

1.2.3 Secondary structure change

The main question of genetics is how genetic (i.e. DNA) differences result in phenotypic (physical) differences between individuals. Any two human genomes differ by .1%, meaning that .1% of their DNA nucleotides are different, 3 million nucleotides in total. Most of these differences occur as single nucleotide variants (SNVs), many of which change the sequence of RNAs transcribed from the genome, potentially affecting their secondary structures. While the difference of a single nucleotide in most cases will not result in significant structure change [28, 151], in some cases a very large structural change results [63, 131]. An SNV that causes such a change is called a riboSNitch; one such example is shown in Figure 1.6. The RNA transcribed from the human \( FTL \) gene contains a functional sequence element called the iron response element (IRE) which binds to the IRE-binding protein. The wild-type \( FTL \) RNA adopts a preferred secondary structure containing the IRE. A SNV that occurs proximal to the IRE results in a much different preferred structure, which constrains the IRE sequence such that the IRE-binding protein is unable to bind. Individuals who have the SNV in question suffer from Hyperferritinemia-cataract syndrome [101]. Thus structural differences can significantly impact biological functions and it is very pertinent to genetics to learn how to best detect SNVs that are riboSNitches. Our assessment of current algorithms for detecting riboSNitches is described in Chapter Two.

1.3 Messenger RNA

In this section I will describe one widely researched class of RNAs known as messenger RNAs (mRNAs) and the role that secondary structure may play in their function. Once transcribed from genes, mRNAs are themselves used as a template to make proteins in a process called 'translation'. Proteins, like RNA and DNA, are polymers and perform crucial roles in the cell. Much of a cell’s energy is directed toward protein production [16] and the order of events from gene to RNA transcript
Figure 1.6: An example riboSNitch. The wild-type \( FTL \) mRNA forms a preferred secondary structure (left) in its 5’ untranslated region containing the iron response element (IRE) motif, which binds the IRE-binding protein. Binding of this protein regulates \( FTL \) translation. A single nucleotide variant in the \( FTL \)’s untranslated region that changes the ‘U’ at position 22 to a ‘G’ significantly changes the preferred secondary structure around the IRE element such that the IRE-binding protein is unable to bind (right). Figure adapted from [101]

to protein translation is known as the ‘central-dogma’ of biology. Messenger RNAs were once recognized only as intermediates between gene and protein, but we now know that mRNAs are subject to complex and varied interactions during transcription [45, 174], post-transcriptionally [2, 70] and during translation [27, 165]. These interactions depend on the makeup of the cell and the sequence and structure of the mRNA.

1.3.1 Translation

The functional path of a typical mRNA is shown in Figure 1.7. Genes that transcribe mRNAs are called protein coding genes, and usually contain several sections: 5’ and 3’ untranslated regions (UTRs), introns and exons. Once transcribed, the intronic sequence in the ‘pre-mRNA’ must be ‘spliced’ out of the transcript in a process called splicing. The result is a mature mRNA. Some genes produce multiple mRNA variants by splicing the pre-mRNAs in alternative ways (leaving in intronic sequence, excising certain exonic sequence in addition to introns, etc.) Mature mRNA is scanned by a large catalytic molecule, the ribosome, which uses the mRNA as a template for a peptide (protein). The sequence in the mRNA that is translated is called the coding sequence, characterized by a start codon and a stop codon which indicate to the ribosome to begin and end translation, respectively.
Figure 1.7: Transcription through translation. A) Genes are transcribed into RNA, some of which is messenger RNA (mRNA). Genes that transcribe mRNAs are called protein coding genes, and usually contain several sections: 5' and 3' untranslated regions (UTRs) (tan), introns (black) and exons (pink). B) Once transcribed, the intronic sequence in the 'pre-mRNA' must be 'spliced' out of the transcript in a process called splicing. The result is a mature mRNA. C) mRNA is scanned by the ribosome, which uses the mRNA as a template for a peptide (protein). The sequence in the mRNA that is translated is called the coding sequence. The 5' and 3'UTR regions of the mRNA are not translated by the ribosome.

The 5' and 3' UTR regions that flank the coding sequence are not translated by the ribosome, but do serve regulatory roles post-transcriptionally, including during translation [9, 67, 76].

Translational control is especially important to understand because it ultimately determines the unique protein output that characterizes different cells, organisms, tissues and individuals. Normally the ribosome is recruited at the very 5' end of the 5'UTR, where it begins to scan the mRNA sequence until encountering a start codon. This is the scanning mechanism of translation [77]. Some 5'UTRs contain short regions of coding sequence before the main protein coding sequence, referred to as upstream open reading frames (uORFs). uORFs have been shown to decrease the translation efficiency of a given mRNA because translation of a uORF is thought to preclude the ribosome from translating the main coding sequence [17, 67, 76]. A modified scanning mechanism of translation is thus necessary to account for the effect of uORFs (Figure 1.8).
Figure 1.8: The Leaky Scanning mechanism of translation. Before initiating translation, the 40S subunit of the ribosome (blue) scans along the 5’UTR (tan) of an mRNA. Some 5’UTRs contain short coding sequences, or upstream open reading frames (uORFs), before the main protein coding sequence (pink). Once encountering the start codon of a uORF, the 40S subunit may recruit the 60S subunit to translate the uORF (top), or it may leak through the uORF sequence and continue scanning (bottom).

1.3.2 Secondary structure in mRNAs

While mRNAs are often pictured as unstructured strands it is important to remember that mRNAs, like all RNA molecules, adopt secondary structures. Therefore, regions in mRNAs that are important for post-transcriptional regulation are particularly interesting for how their secondary structures may partake in this regulation. There are several examples of mRNA secondary structures affecting mRNA regulation. Some regulatory proteins that bind to RNA recognize specific structures [69]. One example is the anomalous protein-binding riboswitch in the 3’UTR of the VEGF mRNA, which regulates VEGF translation [132]. Another such example is the previously described IRE-binding protein that binds to the structured iron response element in 5’UTR of FTL mRNA and regulates FTL translation [101, 141]. Secondary structures in the 5’UTR of mRNAs have especially large potential for affecting translation because of the initiating role the 5’UTR plays in translation. Strong secondary structures in the 5’UTR have been shown to inhibit translation [6], presumably because they impede the ribosome from scanning. The stability of secondary structures around the start codon is also thought to affect translation, where more stable secondary structures are correlated with lower
translation efficiencies [61, 78, 140]. The exact mechanism by which secondary structure would affect translation in eukaryotes is unclear, but by combining mRNA structural data with translation efficiency measurements we can develop a model that quantitatively describes their interaction. In Chapter Three I describe a quantitative model that closely predicts the translation efficiencies of mRNAs with varying 5' UTR sequence, uORFs and secondary structures.
CHAPTER 2
Detecting riboSNitches with RNA folding algorithms: A genome-wide benchmark

2.1 Overview

Ribonucleic acid (RNA) secondary structure prediction continues to be a significant challenge, in particular when attempting to model sequences with less rigidly defined structures, such as messenger and non-coding RNAs. Crucial to interpreting RNA structures as they pertain to individual phenotypes is the ability to detect RNAs with large structural disparities caused by a single nucleotide variant (SNV), or riboSNitches. A recently published human genome-wide parallel analysis of RNA structure (PARS) study identified a large number of riboSNitches as well as non-riboSNitches, providing an unprecedented set of RNA sequences against which to benchmark structure prediction algorithms. Here we evaluate eleven different RNA folding algorithms’ riboSNitch prediction performance on this data. We find that recent algorithms designed specifically to predict the effects of SNVs on RNA structure, in particular remuRNA, RNAsnp and SNPfold, perform best on the most rigorously validated subsets of the benchmark data. In addition, our benchmark indicates that general structure prediction algorithms (e.g. RNAfold and RNAstructure) have overall better performance if base pairing probabilities are considered rather than minimum free energy calculations. Although overall aggregate algorithmic performance on the full set of riboSNitches is relatively low, significant improvement is possible if the highest confidence predictions are evaluated independently.

2.2 Background

Accurate RNA structure prediction remains a contemporary challenge in the field of bioinformatics [146]. The most common approach for predicting RNA structure is minimizing a free energy function derived from thermodynamic parameters for base pairing and stacking energies [91, 104, 180]. Extensive benchmarking of such algorithms has contributed to significant advances in our ability to correctly predict the secondary structure of RNA [11, 54, 173]. Most improvements in RNA structure prediction have focused on highly structured transcripts, i.e. RNAs that have evolved to adopt a narrow range of well-defined conformations often conferring a specific activity such as self-splicing [55, 68, 74, 107].

Messenger RNAs (mRNAs) and non-coding (ncRNAs) RNAs are not evolved to adopt rigidly defined structures, in general adopting an ensemble of diverse conformations. Minimum free energy (MFE) structure prediction strategies are therefore not well suited for these types of RNAs [63, 101]. Accurate prediction of the accessibility of specific sequence motifs in transcripts plays a decisive role in understanding post-transcriptional regulation, as transcript secondary structure can impact the binding of RNA binding proteins, ribosomes and miRNAs [41, 69, 114, 117, 119, 155]. However, given that these RNAs adopt a wide range of structures, traditional structural benchmarking is complicated by the fact that experimental techniques to determine an ensemble of structures do not exist for large RNAs. An alternative strategy is to benchmark folding algorithms’ performance in predicting the perturbation on the structural ensemble by particular mutations [134]. A comprehensive and consistent RNA structure dataset on a large number of mutations in mRNA transcripts was not available until very recently [170].

The advent of transcriptome wide RNA structure probing, and in particular the development of PARS (parallel analysis of RNA structure), provides us with the most comprehensive mRNA and ncRNA benchmark dataset available to date [71, 169]. PARS gathers RNA sequencing reads from transcripts processed by one of two nucleases with diametric affinities for structured versus unstructured regions of RNA. The information from the two nucleases is combined to produce scores reflecting the degree of base pairing at single nucleotide resolution [71]. While other important studies have probed RNA structure at large scale [37, 137, 157, 163], the recent PARS dataset is the first to have detected riboSNitches genome-wide [170]. The comparative structural analysis of a
human family trio’s (mother, father, child) transcriptome structure by PARS has identified almost 2000 riboSNitches [170] in the human transcriptome. A riboSNitch is an element of RNA that changes structure if a specific SNV (Single Nucleotide Variant) is present [63, 89, 101, 134, 170]. Although the majority of riboSNitches have no known phenotypic consequence, specific examples of changes in transcript structure near regulatory regions in mRNAs are associated with human disease [63, 89, 101].

Accurately predicting the extent to which an SNV or mutation disrupts RNA structure is important for the interpretation of personal genomes, since the structural consequences of sequence variants on an individual’s transcripts can impact overall phenotypic characteristics [5, 108]. Even though the vast majority of riboSNitches will likely have limited phenotypic consequences, a structural prediction interpreted in the context of known functional motifs in a transcript can predict function [63, 89, 170]. A series of algorithms have recently been proposed to tackle this challenge [26, 101, 138, 139]. Traditional MFE class algorithms can also be used to predict riboSNitches, although previous benchmarks on in vitro transcribed structured RNAs suggests they overestimate the potential structural disruption of an SNV [129, 131, 134]. The most recent algorithms for predicting the structural disruption of an SNV have therefore focused on analyzing changes in base pairing probability matrix (BPPM) computed from partition function analysis of the Boltzmann suboptimal ensemble [105, 126, 167]. The benchmark carried out below uses the PARS dataset to identify the best algorithmic practices for riboSNitch detection. Furthermore, the performance trends of all prediction algorithms on subsets of differentially validated riboSNitches reveal the relative importance of thermodynamically controlled base pairing in mRNA structure change. Our analysis illustrates the significant remaining computational challenge of riboSNitch prediction and the importance of biological context when making these predictions.

2.3 Results and Discussion

2.3.1 Experimental benchmark criteria

The experimental data for the benchmark is based on a PARS analysis of three related individuals (mother, father and child) from the 1000 Genomes Project [169]. PARS measures the differential
reactivity of each nucleotide in a folded RNA to the V1 and S1 RNases which selectively cleave double and single stranded regions, respectively [111]. Thus the PARS score for each nucleotide is correlated with the extent of base pairing [71]. By comparing the PARS scores at loci in individuals with different alleles (Figure 2.1A) it is possible to detect riboSNitches. Given that these experiments are carried out in a genome-wide manner, thousands of putative riboSNitches were identified [131]. The precise number of riboSNitches identified in such a genome-wide screen depends on the threshold in PARS scores used to call a structural difference. A careful analysis of PARS score differences identified 1907 loci in the human genome as riboSNitches and 10,326 loci where no significant change was observed (non-riboSNitches) [170]. These data form the basis of our benchmark study.

A riboSNitch in the context of the PARS dataset was identified by pairwise comparison of PARS score profiles for transcripts from two individuals in the family trio. Three individual-to-individual comparisons are therefore possible as illustrated in Figure 2.1B. We used this redundancy to identify subsets of riboSNitches with differing levels of experimental confidence. In the most consistent case, the "symmetric" riboSNitch, PARS score profiles are significantly different in all cases comparing individuals with different genotypes. In certain cases however, not every genotypic difference results in a significant PARS profile difference; we consider these cases "asymmetric" riboSNitches.

An alternative approach for validating a riboSNitch using the PARS data involves allele-specific mapping at heterozygous loci in the child dataset. This type of validation was performed at loci where the parents are homozygous different with a necessarily heterozygous child, as diagramed in Figure 2.1C. Successful validation also requires that the significant structure change be 3’ of the SNV. This is a consequence of PARS library preparation, in which only fragments 5’ of the endonuclease cut site are sequenced; only sequencing reads that include the SNV can be mapped in an allele-specific manner. Allele-specific analysis of the PARS data validated 115 riboSNitches [170] and we refer to these riboSNitches as "validated". The most rigorous form of validation is independent chemical and enzymatic structure probing on in vitro transcribed constructs for each allele, as illustrated on Figure 2.1D (45-47). This involves a separate experiment, and 11 riboSNitches were further validated in this way [170]. This set of riboSNitches is referred to as "probed" for the purposes of this benchmark. The categorization proposed here yields subsets of riboSNitches for use in our benchmark with differing levels of experimental validation. The entire set of riboSNitches we refer to as the "all" dataset or category. Additional filtering was applied to the "all" set of 1907 reported [170] riboSNitches to
ensure a completely non-redundant dataset. A riboSNitch in the PARS dataset is identified as an SNV in a specific transcript isoform. In this study we use sequences centered on these SNVs to test folding algorithms. However, many transcripts have isoforms with the same window (+/- 50 nt) of sequence around a particular SNV. As a result we benchmarked on the 1058 unique sequence subset of the 1907 reported riboSNitches. The numbers of unique riboSNitches and corresponding non-riboSNitches for each level of experimental validation are reported in Table 2.1. Non-riboSNitch set sizes were matched to each riboSNitch category size in order to reduce computational costs.
as well as to match non-riboSNitches and riboSNitches in terms of their experimental validation (methods).

### 2.3.2 Algorithmic performance analysis

Given the unprecedented number of riboSNitches discovered in the human genome, our benchmark has the potential to broadly evaluate the performance of prediction algorithms. A general strategy for riboSNitch prediction and our approach to benchmarking is summarized in Figure 2.2. A structure prediction is made on RNA sequences containing both alleles (Figure 2.2A) for subsequent comparison. To benchmark the algorithms, predictions are made for sequences identified as riboSNitches and those where no experimental structure change is observed. For example, in Figure 2.2A, a T/A SNV in the 3’ UTR of SUB1 (activated RNA polymerase II transcriptional co-activator p15 or SUB1 homolog) is a riboSNitch since PARS scores differ at this locus, indicating a structure difference between the two alleles. The T/C SNV in the 3’ UTR of PARP1 (poly [ADP-ribose] polymerase 1) does not alter structure in the PARS data and is therefore a non-riboSNitch. The structures of each variant are compared (Figure 2.2B), and in this example, RNAfold [91] correctly predicts that the base pairing probabilities computed for each allele are very different for the riboSNitch, and nearly identical for the non-riboSNitch (Figure 2.2B, left vs. right panel).

<table>
<thead>
<tr>
<th></th>
<th>Probed</th>
<th>Validated</th>
<th>Symm</th>
<th>Asymm</th>
<th>All</th>
<th>25% tails²</th>
<th>5% tails³</th>
</tr>
</thead>
<tbody>
<tr>
<td>riboSNitch</td>
<td>11</td>
<td>63</td>
<td>223</td>
<td>835</td>
<td>1058</td>
<td>260-288</td>
<td>47-59</td>
</tr>
<tr>
<td>Non-riboSNitch</td>
<td>11</td>
<td>63</td>
<td>223</td>
<td>835</td>
<td>1058</td>
<td>240-268</td>
<td>46-79</td>
</tr>
</tbody>
</table>

In choosing RNA structure prediction algorithms to test in the benchmark we opted for a variety of algorithms. We differentiate prediction algorithms into two broad classes: specialized (i.e. algorithms specifically engineered to predict the effects of mutations on RNA structure) and general (i.e. algorithms that are developed to predict RNA structure). Specialized algorithms generally report a score (or confidence) for a predicted difference between mutant RNAs. For non-specialized

²Set sizes from the MFE algorithms were not included in this range.
³Set sizes from the MFE algorithms were not included in this range.
Figure 2.2: RiboSNitch prediction and benchmarking strategy. A.) To benchmark, all the prediction algorithms estimate structure features for both alleles of an RNA. This is done for both the riboSNitch sets and non-riboSNitch sets. The riboSNitch in this example is the sequence flanking an SNV in the 3’ UTR of SUB1 (activated RNA polymerase II transcriptional co-activator p15 or SUB1 homolog) that yields a differential PARS score between alleles. The non-riboSNitch is the sequence flanking an SNV in the 3’ UTR of PARP1 (poly [ADP-ribose] polymerase 1) where no significant PARS score differences were measured. Alleles are color-coded with green and red for the T and A riboSNitch alleles and green and yellow for the T and C non-riboSNitch alleles. B.) RNAfold [91] predicted minimum free energy structures and base pairing probability matrices for the riboSNitch and non-riboSNitch alleles. In this example, the prediction is correct in that the riboSNitch T (green) and A (red) alleles show a large difference in predicted base pairing probabilities while the non-riboSNitch T (green) and C (yellow) alleles do not. C.) The difference between alleles’ structures are measured to produce a distance score. Whether the riboSNitch is considered a True Positive (TP) or False Negative (FN) and the non-riboSNitch a True Negative (TN) or False Positive (FP) depends upon the score threshold used. The distance scores between alleles are used as thresholds to perform Receiver Operator Curve (ROC) analysis [152] and evaluate predictive performance.

algorithms a scoring metric to compare base pairing probabilities or MFE predictions is required. We used RNAdistance and a custom implementation of RNApdist, algorithms implemented in the ViennaRNA package, to carry out these comparisons and quantify structural distances [10, 59, 91]. This allows us to define a variable threshold over predictions on riboSNitches and non-riboSNitches
for each algorithm (Figure 2.2C) and perform receiver operator curve (ROC) analysis to benchmark performance [152]. By performing ROC analysis for each algorithm on the subsets of riboSNitches with differing levels of experimental evidence (Figure 2.1), we are also able to indirectly evaluate the relative importance of the different types of experimental validation.

Representative ROC curves for the five different levels of experimental riboSNitch validation are shown in Figures 2.3A-E, respectively. Large differences in performance are observed between algorithms (especially with the “probed” riboSNitches, Figure 2.3A); the specialized algorithms in most cases outperform the general class of algorithms (Tables 2.2 and 2.3). This is unsurprising given that the general algorithms were not specifically designed for riboSNitch prediction; however, some of the general algorithms still show comparable performance to the specialized (Table 2.3, bolded values). The best prediction performance is observed on the most highly validated subset of riboSNitches, i.e. those validated independently by in vitro chemical and enzymatic probing. The prediction performance of the specialized algorithms on the in vitro validated (or “probed”) data is on par with, and in some cases higher than, what was observed in a previous benchmark on structured RNAs performed in 2012 [134]. Recent algorithmic developments, and in particular the analysis of local structure change performed in RNAsnp, appear to improve riboSNitch predictions for mRNA and ncRNAs [139].

Performance benchmarks for all algorithms tested are summarized in Tables 2.2 and 2.3 along with 95% confidence intervals. While the intervals are often wide due to small sample sizes, a few consistent patterns are apparent throughout. The majority of algorithms perform best on the most experimentally validated dataset. This is true of both generalized and specialized prediction algorithms. In addition, aggregate prediction accuracy of all algorithms decreases with lower levels of experimental validation. None of the algorithms have good performance on the ”asymmetric” and ”all” datasets (Figure 2.3E), yielding area under the curve (AUC) values slightly greater than 0.5 (Tables 2.2 and 2.3). This is expected to some extent since the experimental false discovery rate (FDR) increases with lower levels of experimental confidence. However, it could also result from a population of RNAs in the data that are riboSNitches only in vivo and not purely driven by thermodynamic changes in base pairing probabilities. Since the PARS experiment was carried out on in vivo transcribed RNAs, it has the potential to detect such riboSNitches [170]. RNA structure in the cell may differ from that in the tube for a number of reasons. Co-transcriptional folding
may encourage certain structures over others [82] and small RNA (miRNA and siRNA) and protein binding to transcripts may stabilize or induce certain structural elements [177]. Folding is known to depend on solvent conditions and could thus be influenced by conditions in the cell like salt concentration, pH and molecular crowding. Post-transcriptional RNA sequence modifications could allow for additional structure disruptions that are unique to the cell environment and detectable by PARS. The thermodynamic prediction algorithms used here likely fail to predict large structure changes for riboSNitches that result from these environmental contributors.

The variation of structure predictions with sequence length is an important factor to consider in interpreting these results. We found that AUC values do vary with sequence length, but are robust
Table 2.2: ROC results for general RNA folding algorithms. Results are reported as the area under the curve (AUC) for each ROC curve with 95% confidence intervals underneath. Top performers are in bold.

<table>
<thead>
<tr>
<th>Software</th>
<th>Prediction program</th>
<th>Probed</th>
<th>Validated</th>
<th>Symm</th>
<th>Asymm</th>
<th>All</th>
<th>25% tails</th>
<th>5% tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>remuRNA [139]</td>
<td>McCaskill-remuRNA</td>
<td><strong>0.736</strong></td>
<td>0.537</td>
<td>0.543</td>
<td>0.524</td>
<td>0.537</td>
<td>0.574</td>
<td>0.567</td>
</tr>
<tr>
<td>03Nov2012</td>
<td>RNAfold 1.4</td>
<td>(0.514-0.957)</td>
<td>(0.456-0.658)</td>
<td>(0.49-0.597)</td>
<td>(0.497-0.552)</td>
<td>(0.517-0.51)</td>
<td>(0.513-0.562)</td>
<td>(0.525-0.622)</td>
</tr>
<tr>
<td>1.0</td>
<td>RNAfold 1.1</td>
<td><strong>0.777</strong></td>
<td>0.583</td>
<td>0.568</td>
<td>0.529</td>
<td>0.533</td>
<td>0.583</td>
<td>0.658</td>
</tr>
<tr>
<td>1.1</td>
<td>RNAfold 1.1</td>
<td>(0.573-0.987)</td>
<td>(0.483-0.684)</td>
<td>(0.515-0.621)</td>
<td>(0.501-0.557)</td>
<td>(0.508-0.558)</td>
<td>(0.535-0.631)</td>
<td>(0.556-0.761)</td>
</tr>
<tr>
<td>SNPfold [63]</td>
<td>RNAfold 2.1.1</td>
<td><strong>0.703</strong></td>
<td>0.581</td>
<td>0.571</td>
<td>0.520</td>
<td>0.528</td>
<td>0.591</td>
<td>0.736</td>
</tr>
<tr>
<td>1.01</td>
<td>RNAfold 2.1.1</td>
<td>(0.466-0.939)</td>
<td>(0.48-0.681)</td>
<td>(0.493-0.548)</td>
<td>(0.487-0.546)</td>
<td>(0.504-0.553)</td>
<td>(0.543-0.639)</td>
<td>(0.638-0.835)</td>
</tr>
</tbody>
</table>

Table 2.3: ROC results for general RNA folding algorithms. Results are reported as the area under the curve (AUC) for each ROC curve with 95% confidence intervals underneath. Top performers are in bold.

<table>
<thead>
<tr>
<th>Software</th>
<th>Structure type</th>
<th>Probed</th>
<th>Validated</th>
<th>Symm</th>
<th>Asymm</th>
<th>All</th>
<th>25% tails</th>
<th>5% tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentroidFold [64]</td>
<td>BPPM</td>
<td>0.579</td>
<td>0.561</td>
<td>0.569</td>
<td>0.529</td>
<td>0.532</td>
<td>0.596</td>
<td>0.637</td>
</tr>
<tr>
<td>0.0.9</td>
<td>MFE</td>
<td>(0.324-0.833)</td>
<td>(0.466-0.662)</td>
<td>(0.516-0.622)</td>
<td>(0.502-0.557)</td>
<td>(0.507-0.556)</td>
<td>(0.534-0.645)</td>
<td>(0.528-0.746)</td>
</tr>
<tr>
<td>CONTRAfold [39]</td>
<td>BPPM</td>
<td>0.463</td>
<td>0.562</td>
<td>0.567</td>
<td>0.528</td>
<td>0.535</td>
<td>0.562</td>
<td>0.613</td>
</tr>
<tr>
<td>2.02</td>
<td>MFE</td>
<td>(0.257-0.768)</td>
<td>(0.474-0.674)</td>
<td>(0.523-0.627)</td>
<td>(0.497-0.552)</td>
<td>(0.51-0.558)</td>
<td>(0.509-0.57)</td>
<td>(0.489-0.531)</td>
</tr>
<tr>
<td>MC-Fold [116]</td>
<td>BPPM</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>17Mar2008</td>
<td>MFE</td>
<td>0.424</td>
<td>0.460</td>
<td>0.478</td>
<td>0.497</td>
<td>0.493</td>
<td>0.493</td>
<td>0.493</td>
</tr>
<tr>
<td>RNAfold [91]</td>
<td>BPPM</td>
<td><strong>0.686</strong></td>
<td>0.597</td>
<td>0.581</td>
<td>0.525</td>
<td>0.549</td>
<td>0.562</td>
<td>0.613</td>
</tr>
<tr>
<td>2.1.1</td>
<td>MFE</td>
<td>(0.441-0.931)</td>
<td>(0.498-0.697)</td>
<td>(0.528-0.634)</td>
<td>(0.497-0.553)</td>
<td>(0.509-0.559)</td>
<td>(0.540-0.637)</td>
<td>(0.606-0.808)</td>
</tr>
<tr>
<td>RNAmutants [168]</td>
<td>BPPM</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2.0</td>
<td>MFE</td>
<td>0.517</td>
<td>0.474</td>
<td>0.504</td>
<td>0.510</td>
<td>0.509</td>
<td>0.501</td>
<td>0.493</td>
</tr>
<tr>
<td>RNAstructure [133]</td>
<td>BPPM</td>
<td><strong>0.612</strong></td>
<td>0.578</td>
<td>0.567</td>
<td>0.527</td>
<td>0.536</td>
<td>0.553</td>
<td>0.622</td>
</tr>
<tr>
<td>5.6</td>
<td>MFE</td>
<td>(0.358-0.865)</td>
<td>(0.478-0.678)</td>
<td>(0.513-0.62)</td>
<td>(0.499-0.554)</td>
<td>(0.51-0.56)</td>
<td>(0.504-0.602)</td>
<td>(0.514-0.731)</td>
</tr>
<tr>
<td>UNAFold [100]</td>
<td>BPPM</td>
<td><strong>0.471</strong></td>
<td>0.537</td>
<td>0.548</td>
<td>0.524</td>
<td>0.526</td>
<td>0.528</td>
<td>0.578</td>
</tr>
<tr>
<td>3.8</td>
<td>MFE</td>
<td>(0.21-0.732)</td>
<td>(0.435-0.639)</td>
<td>(0.494-0.601)</td>
<td>(0.496-0.551)</td>
<td>(0.502-0.551)</td>
<td>(0.478-0.577)</td>
<td>(0.467-0.688)</td>
</tr>
</tbody>
</table>

within certain ranges of lengths (Supplementary Figure A.1). Even though AUC values can have high variance, especially in the smaller categories of riboSNitches, the general trend of AUC values decreasing from highly validated riboSNitch categories to the lowest is still evident across a range of sequence lengths, tested from 21 to 201 nucleotides (Supplementary Figure A.2). Furthermore, the 95% confidence intervals on algorithms’ AUC values (Tables 2.2 and 2.3) capture most of the variation in performance observed with different sequence lengths (Supplementary Figures A.1-A.2).
2.3.3 Improving riboSNitch prediction performance

Both riboSNitch prediction and experimental validation require defining a threshold for what is and is not a significant change in structure. Experimentally, the threshold is based on whether the PARS signal is measurably different between genotypes [170]. Our benchmark shows that higher levels of experimental validation correlate with improved prediction accuracy for most algorithms (Tables 2.2 and 2.3), indicating that the FDR in the data decreases with increasing validation. To address this possible higher FDR in the category of all riboSNitches, we tested ROC analysis on subsets of riboSNitches with the lowest p-values, as determined by their PARS scores. Subsets used were the best 25% and the best 5% out of all riboSNitches. The resulting ROC curves show little to no improvement in their AUC values (Supplementary Table A.1). This argues against false positives as the main cause of poor predictions in the "all" category. It suggests instead that a prohibitive number of the experimentally predicted riboSNitches are in fact environmental riboSNitches that do not show large structure changes driven purely by base pairing thermodynamics.

All algorithms in this benchmark predict the extent of structural change caused by an SNV. They differ in the metric used to quantify the structural change and/or the specific free-energy functions used for the prediction. The specialized algorithms report a score for each structure comparison, while for the generalized algorithms we evaluate the change in either the predicted minimum free energy (MFE) structures or base pairing probability matrices (BPPM) (Figure 2.2B). These scores are a proxy for the extent of structural disruption by the SNV. Though all the algorithms predict very similar score distributions for riboSNitches and non-riboSNitches, in many cases the non-riboSNitch score distribution is more skewed toward low distance scores, or strong non-riboSNitches (Supplementary Figure A.3). Consequently, the largest differences between riboSNitch and non-riboSNitch distance score distributions lie at the extremes. We therefore evaluated performance of the algorithms with just their extreme-valued predictions on the "all" dataset. We propose that such a strategy will mitigate some of the experimental thresholding issues that make genome-wide prediction of riboSNitches difficult. In practical terms, the algorithm is allowed to "opt-out" of making a prediction if the riboSNitch predicted score is not at one of the extremes of its score distribution.

To evaluate performance on these extreme-valued predictions we combined riboSNitch and non-riboSNitch distance scores from the "all" category for each algorithm in order to pick the values that
mark the tails occupying n% of the distribution (Figure 2.4A-B). Both 25% and 5% tails were tested, where the range of set sizes from scores selected in this way are listed in the last two columns of Table 2.1. Scores between the marker values were removed from the riboSNitch and non-riboSNitch scores to make filtered score sets for ROC analysis (Figure 2.4C). As can be seen in the ROC curves in Figure 2.5 5 and Tables 2.2 and 2.2, this strategy generally improves AUCs and in some cases results in prediction performance equivalent to that obtained on the "probed" riboSNitch subset. RNAfold and SNPfold in particular benefit from this method, with 95% confidence intervals on their 5% tail AUC values completely outside of the confidence intervals on their "all" category AUC values. Interestingly, this approach does not improve the performance of MFE-based predictions as much as BPPM-based predictions for any of the algorithms. The score distributions taken from MFE structure distances are not diverse, i.e. many of the scores are zeros, rendering this sort of filtering ineffective. However, this ”opt-out” method is still highly effective with BPPM-based algorithms and represents a simple strategy to improve predictions on genome-wide datasets and to identify the highest-confidence predictions.

Figure 2.4: Schematic on isolating the n% tails of riboSNitch and non-riboSNitch score distributions. A) riboSNitch and non-riboSNitch distance scores from a particular algorithm are temporarily combined into one larger set. B) Scores that delineate the middle 100-n% of the combined distribution are determined, shown here as \( \theta_L \) and \( \theta_H \). C) Scores that are lower than \( \theta_L \) and higher than \( \theta_H \) (shaded grey) are selected from the original riboSNitch and non-riboSNitch score sets for ROC analysis.
Figure 2.5: Example improvements with the n% tails strategy. A) ROC curves for RNAstructure on all riboSNitches (light gray), 25% tails (medium gray) and 5% tails (black). B) ROC curves for SNPfold on all riboSNitches (light gray), 25% tails (medium gray) and 5% tails (black).

2.3.4 Discussion

RNA folding is a complex process that is driven by both thermodynamic and kinetic factors [23, 130, 144]. In the cell, exogenous factors including protein chaperones and small molecules interact with the RNA, further altering the folding behavior [145, 150, 158, 175, 177]. In addition, RNA necessarily folds co-transcriptionally, adding another layer of kinetic complexity to the problem of accurate structure prediction [36, 85, 164, 178]. The genome-wide discovery of riboSNitches using PARS enzymatic structure probing provides an unprecedented experimental analysis of RNA structure and can significantly contribute to our understanding of the relative importance of these various factors in defining conformation [131].

Aside from CONTRAfold [39] the algorithms used in this benchmark model RNA structure relying on parameters derived from thermodynamic analysis of base pairing and base-stacking [104, 161, 176]. Although these energy functions are still being refined, they do capture a majority of the thermodynamically important features that are known to drive RNA folding [104, 143, 176]. The primary focus of this benchmark is to determine the best algorithms for riboSNitch prediction, though the consistent differences in performance of most algorithms on the subsets of differentially validated data reveal important lessons for RNA structure prediction in general.

Our benchmark shows that a majority of the algorithms best predict "probed" category riboSNitches (Figure 2.3A and Tables 2.2 and 2.3). This category of validation involves independent in vitro transcription and folding of the RNA prior to chemical probing. Thus all sequences in the
"probed" dataset are riboSNitches detected both in vitro and in vivo. The small size of the "probed" category makes it difficult to demonstrate significant improvement compared to the other categories with a given algorithm. Nevertheless, the highest scoring algorithms have "probed" category AUC values that are significantly greater than the AUC values derived from other categories. The remuRNA "probed" AUC value is greater than the "asymmetric" and "all" AUC values (p-val=.039 and .047, respectively) and the RNAsnp "probed" AUC is significantly greater than the "symmetric", "asymmetric” and "all” AUC values (p-val=.032, .014 and .015, respectively). The aggregate performance of algorithms on allele-specific validated riboSNitches ("validated” category) is lower than the "probed” set but still consistently better than predictions on all riboSNitches (Figure 2.3A-B and Tables 2.2 and 2.3). "Symmetric” riboSNitches are better predicted than "asymmetric” (Figure 2.3C-D and Tables 2.2 and 2.3). These results immediately suggest that higher levels of experimental validation reduce the FDR in the PARS data, since the algorithmic prediction accuracy is likely constant. However, it may also suggest that the increasingly validated categories correspond to RNAs whose folding is governed more by thermodynamic changes in base pairing over other cellular factors. The presence of a class of thermodynamic, versus environmental, riboSNitches is more consistent with our results. Environmental riboSNitches would be poorly predicted by RNA folding algorithms, and in this benchmark, they likely blur the difference in score distributions between riboSNitches and non-riboSNitches. The "asymmetric” dataset, which yields the worst folding prediction performance on average, is likely enriched in RNAs whose folding is determined more by cellular environment and genetic background than by sequence-based folding dynamics, as individual variation in environment can cause RNAs to fold differently. We also cannot exclude the possibility that some of the false-positive and negative predictions are due to post-transcriptional editing [31].

RNAsnp [138] is the highest performing algorithm on the "probed” dataset (Table 2.2). This algorithm analyzes local structure changes between mutant structures, while other algorithms measure structure change over the entire length of the sequence. This result is, perhaps, unsurprising given that the riboSNitches used in this benchmark had been detected by comparing PARS scores in a small region (5 nt) around each SNV [170]. While RNAsnp surpasses some of the other algorithms only marginally, considering its comparable performance, it is clear that considerations of local structure change can be successful in riboSNitch prediction.
Despite fair performance of some of the algorithms on the "probed" category, performance in other categories remains poor. Two strategies were attempted to improve the performance on the "all" riboSNitch category, as good performance on the most comprehensive datasets should be the ultimate goal of prediction algorithms. The first, making predictions only on riboSNitches with PARS score p-values in the bottom 25% and 5% of all riboSNitches, resulted in few improvements (Supplementary Table A.1). This suggests that a higher experimental FDR is not solely responsible for the poor performance on the "all" dataset. However, we cannot exclude that the p-value from structural differences in the PARS data may be a poor indicator of confidence in a riboSNitch. It is important to note that the lack of improvement in AUC values with these subsets of the "all" category also verifies that smaller sample sizes do not superficially improve results. Thus the better performance in the more validated categories is not simply due to their smaller sample size.

The second strategy, ROC analysis on the 5% (and 25%) tail values of riboSNitch and non-riboSNitch score distributions, resulted in marked improvements. In some cases, as in the case of SNFpold and RNAfold, performance with the 5% tails is on par with that of the chemically probed category and is significantly greater than the corresponding "all" category performance (p-val=5.28e-05 and 7.53e-04 respectively). The algorithms that did not improve were those that only predict MFE structures. Closer examination of MFE distance scores for "all" category riboSNitches and non-riboSNitches showed sparse score distributions (i.e. a lot of zeros). Based on these results we suggest that users can make the most of genome-wide predictions by using just the top and bottom 5% of BPPM-based distance scores to indicate riboSNitches and non-riboSNitches, respectively.

It is interesting to note that changing the metric used to measure structure differences can result in better performance. All the specialized algorithms essentially apply new metrics to existing RNA folding algorithms, so by benchmarking the specialized algorithms we are testing a number of different metrics. For instance, SNPfold takes its structure predictions from RNAfold, but measures structure differences with a Pearson correlation coefficient between base pairing probabilities instead of with one of the RNAfold distance functions. Remarkably, SNPfold performs better than RNAfold (BPPM) on average with its simpler distance metric. Some of the other BPPM folding algorithms also exhibit better scores when switched to using a correlation coefficient, namely CentroidFold and CONTRAfold, while RNAstructure and UNAFold exhibit slightly lower scores (Supplementary Table A.3). Different folding parameter options may affect results as well. In this study we used
default parameters for many of the algorithms, but users may boost performance with specifically tailored options. For example, RNAsnp is based on the assertion that measuring local structure changes is essential for riboSNitch prediction. To this end, RNAsnp provides an option to define the minimum size of local intervals in which to measure structure change. In this benchmark we used a minimum interval length of 10, but users may choose to change this based on their sequence length or the expected scope of structural change.

Choosing BPPM predictions over MFE structure prediction results in large improvements in performance among the general algorithms. In particular, RNAfold's "probed", "symmetric", "all", 25% tails and 5% tails categories were all significantly greater with BPPM AUC values than MFE (p-val=.044, 8.9e-03, .047, 6.2e-03 and 9.0e-05, respectively). RNAstructure BPPM AUC values were significantly greater than MFE values in the "probed" and 5% tails categories as well (p-val=.014 and .024, respectively). Furthermore, MFE-based predictions on structure disruption have been shown to underperform BPPM-based predictions previously [134]. Based on inspection of MFE-based distance score distributions, many of the non-riboSNitch sets yielded distance scores just as high as their matched riboSNitch sets, indicating a tendency to over-predict structure changes. However, MFE predictions on both the riboSNitch and non-riboSNitch sets often contain an abundance of zero-valued structure distance scores, representing a tendency to under-predict as well. Predictions based on MFE structures appear to be too reductive, that is, MFE predictions do not capture enough information about RNA structural ensembles to be useful for structure comparisons. An exception among thermodynamic algorithms, CentroidFold does not in fact return true MFE structures, instead choosing the structure that optimally agrees with each sequence's predicted base paring probability matrix. Interestingly, the predictions from CentroidFold have the highest AUC values out of all the algorithms' MFE results. Improved performance from representative dot bracket structures chosen in this way further underscores the importance of considering the entire Boltzmann ensemble in RNA structure prediction.

Recommendations

Based on this benchmark we propose the following best practices for riboSNitch prediction and experimental validation.

- For single riboSNitch prediction we recommend using one of the BPPM-based specialized
algorithms—remuRNA, RNAsnp or SNPfold.

- For genome-wide prediction, performance will be greatly improved if only the top and bottom 5% of SNPfold and/or RNAfold (BPPM) predictions on the particular dataset are used to indicate riboSNitches and non-riboSNitches.
- Experimentally, if a structural disruption is consistently observed in multiple individuals and by allele-specific mapping it is likely that thermodynamic changes are driving the observed structure change.

2.4 Materials and Methods

2.4.1 The benchmark dataset

The benchmark dataset The PARS dataset tested a total of 12,233 specific SNV-transcript pairs, with 1,907 of these determined to be riboSNitches. For consistency in benchmarking algorithms we considered the 50 bases 5’ and 3’ around each SNV as standard input sequence for folding prediction, or 101 bases total. SNV-transcript pairs that contain less than 50 nucleotides between the SNV and the transcription start site were excluded. In many cases one SNV tested with several different transcripts has the same surrounding sequence in each isoform. These SNVs were condensed into one entry to ensure a set of non-redundant sequences. This curated set of SNV-transcript pairs contains 1058 riboSNitches and 5469 non-riboSNitches. RiboSNitches were organized into “symmetric” and “asymmetric” categories based on whether or not pairwise comparisons between mother, father or child consistently indicated a riboSNitch in the presence of different genotypes. RiboSNitches that were further validated with allele-specific mapping were also added to the ”validated” category, and riboSNitches that were validated with chemical probing were added to the ”probed” category. For sequences corresponding to multiple SNV-transcript pairs, the presence of one pair qualifying as a riboSNitch was enough to consider the sequence a riboSNitch in this benchmark. Likewise, an SNV-transcript pair categorized as a ”symmetric”, or ”validated” or ”probed” riboSNitch was sufficient to place the sequence into those categories. Sequences for riboSNitches and the matched
1058 non-riboSNitches are provided as text files. RiboSNitch sequences are organized according to category.

2.4.2 Benchmark design and distance metrics

Structure prediction programs were tested on the sequences containing each allele for every riboSNitch and non-riboSNitch. Non-riboSNitch sets were matched in size to each riboSNitch set to reduce computational costs. As a strategy for matching non-riboSNitches and riboSNitches in terms of their experimental validation, a non-riboSNitch set was matched to a riboSNitch set of size n by selecting the top n non-riboSNitches according to their FDR-adjusted p-values from PARS comparisons. Since each RNA has p-values on three potential comparisons–mother vs. father, mother vs. child and father vs. child–the p-value used here is the average of the comparisons. The Unix commands used for each algorithm are listed in Supplementary Table A.4. The ”specialized” algorithms directly score the distance between sequence pairs. SNPfold scores with a Pearson correlation coefficient, RNAsnp returns a p-value on Euclidean distance, remuRNA measures the relative entropy between two RNAs and RNAmute measures the edit distance between MFE structures. For algorithms that do not intrinsically compare the structures of sequences between two RNA variants, predictions on dot bracket structures or base pairing probability matrices (BPPM) were compared for each allele. All the general algorithms except CONTRAfold and CentroidFold return a minimum free energy (MFE) structure as their dot bracket structure, so dot bracket structures are referred to as MFE structures in this benchmark. MFE structures were compared with the RNAdistance function from ViennaRNA 2.1.1 and BPPMs were compared with RNApdist. The RNApdist function used here is a modified version of the RNApdist function implemented by ViennaRNA [12, 91]. Essentially, base pairing probability differences are summed without performing an alignment of the base pairing probability matrices. The distance between base pair probability matrices of sequences 1 and 2 is given by

\[ \sum_{i=1}^{n} 1 - \left( \sqrt{p_{1}^{1}p_{2}^{1}_i} + \sqrt{p_{1}^{1}p_{2}^{2}_i} + \sqrt{p_{1}^{1}p_{2}^{3}_i} \right) \]

and

\[ p_{i}^{j} = \sum_{n \geq j > i} p_{i,j} \]
\[ p_i^j = \sum_{1 \leq j < i} p_{i,j} \]

\[ p_i^o = 1 - p_i^j - p_i \]

\( p_{i,j} \) is the probability of base \( i \) being paired with base \( j \). \( p_i^j, p_i^o \) and \( p_i \) are the probabilities of base \( i \) being upstream paired, downstream paired and unpaired, respectively, for base pairing probability matrices 1 and 2. Note that this modification on RNApdist assumes that the sequences being compared have the same length. The ViennaRNA implementation of the RNApdist function was used for benchmarking RNAfold (which is the main folding algorithm in the ViennaRNA package.)

### 2.4.3 ROC analysis

ROC analysis The distance scores predicted for a riboSNitch set and matched non-riboSNitch set were compared with Receiver Operating Characteristic analysis. ROC curves were constructed and their areas measured using the R package pROC version 1.7.3 [135]. ROC curves were constructed in this way for every RNA folding algorithm across all riboSNitch categories.

25% and 5% subsets of all riboSNitches were selected based on riboSNitches with the lowest average FDR-adjusted p-values. Non-riboSNitch sets were matched to these subsets using the method applied to the other riboSNitch categories described above. ROC analysis was performed on these subsets as before. Results shown in Supplementary Table A.1.

To test the n% tails of a riboSNitch category—25% and 5% tails tested here—all of a program’s scores from riboSNitches and non-riboSNitches were combined to determine the threshold values that mark the middle 100-n% of the distribution. Scores above and below these threshold values were then selected from riboSNitch and non-riboSNitch score sets separately. ROC analysis was then performed on these selected values as described previously. Score distributions and 5% tails for every algorithm are illustrated in Supplementary Figures A.3A and A.3B. Distributions were graphed with the density function in R 3.1.1.

The 95% confidence interval was calculated for each AUC value using the DeLong method in the pROC package [35, 135]. Essentially, 95% confidence intervals are calculated as \( \text{AUC} \pm 1.96s \), where \( s \) is the standard deviation of the given AUC. Any comparisons between AUC values generating
a p-value were completed with the pROC package roc.test function using a one-tailed test with DeLong’s test for two ROC curves [35, 135].

An ROC curve’s “best” point was considered to be the point closest to the top left corner of the graph. The threshold yielding the best point, as well as the specificity and sensitivity values of the point were listed for every ROC curve (Supplementary Table A.2). These were determined with the pROC R package coords function [135]. Thresholds in the “probed” category could be reasonably used as cutoff scores for riboSNitch detection.
CHAPTER 3

An RNA structure-mediated, post-transcriptional model of human α-1-antitrypsin expression

3.1 Overview

α-1-antitrypsin deficiency is a major genetic cause of Chronic Obstructive Pulmonary Disorder (COPD), which affects over 65 million individuals worldwide. The α-1-antitrypsin gene, SERPINA1, expresses an exceptional number of mRNA isoforms generated entirely by alternative splicing in the 5’ untranslated region (5’UTR). Although all SERPINA1 mRNAs encode exactly the same protein, the individual expression levels of the mRNAs vary substantially in different human tissues. We hypothesize that these transcripts behave unequally in a post-transcriptional regulatory program via their distinct 5’UTRs and that this regulation ultimately determines α-1-antitrypsin expression. Using whole-transcript SHAPE chemical probing, we show that splicing dictates the local 5’UTR secondary structure of the SERPINA1 transcripts. Splicing in the 5’UTR also determines the inclusion of long upstream open reading frames (uORFs). We demonstrate that disrupting the uORFs results in markedly increased translation efficiencies in luciferase reporter assays. This suggests that α-1-antitrypsin protein expression levels are controlled at the post-transcriptional level. A leaky scanning model of translation that uses the Kozak translation initiation sequence alone does not account for our quantitative data nor does the model adequately predict α-1-antitrypsin expression in human primary tissue. However, when we incorporate the experimentally derived structure data, the model accurately predicts the translation efficiencies in our reporter assays, and greatly improves translation efficiency predictions in primary human tissues. Our results reveal that RNA structure governs a complex post-transcriptional regulatory program of α-1-antitrypsin expression. Crucially, these

1This work has been submitted as a manuscript to an academic journal and is in revision. The authors are as follows: Corley, M., Solem, A., Phillips, G., Lackey, L., Ziehr, B., Vincent, H., Mustoe, A., Weeks, K., M., Moorman, N., J., Laederach, A.
findings describe a new mechanism by which genetic alterations in non-coding regions of the gene may result in α-1-antitrypsin deficiency.

### 3.2 Background

Human α-1-antitrypsin is of particular clinical interest because deficiencies in this protein are associated with Chronic Obstructive Pulmonary Disease (COPD), liver disease and asthma [19, 32, 33, 42, 98, 125, 128]. Smoking is the major environmental factor that contributes to COPD risk, although the inconsistency of COPD rates among smokers points to additional genetic factors that modulate risk [22, 88, 106, 124]. Multiple genetic variants in the gene encoding α-1-antitrypsin, SERPINA1, cause α-1-antitrypsin deficiency [127, 128], which can result in COPD, liver failure and inflammatory conditions like panniculitis, vasculitis and glomerulonephritis [43, 51, 87] α-1-antitrypsin is a protease inhibitor that specifically targets neutrophil elastase, which is present at chronic low levels in the lungs [29]. Deficiency of α-1-antitrypsin thus results in higher levels of neutrophil elastase, which in turn degrades elastin (especially in the lungs), resulting in COPD [72, 142]. Thus the role of SERPINA1 in COPD etiology is well described at the protein level; however, little is known about SERPINA1 at the transcript level and whether alteration of potential post-transcriptional controls can contribute to α-1-antitrypsin deficiency and ultimately COPD. Genome-wide association study (GWAS) identified COPD-associated variants that map to the SERPINA1 untranslated regions (UTRs), introns, and promoter region [22, 109]. Furthermore, genetic variants shown to alter SERPINA1 splicing patterns were identified in the SERPINA1 introns of two different patients with COPD [84, 147]. The presence of disease-associated variants in non-coding regions suggests that post-transcriptional regulation of SERPINA1 mRNA is an important component of disease risk. Nevertheless, variants in non-coding regions of SERPINA1 comprise only a small fraction of its disease-associated variants discovered to date, but this could be due to the tendency of variant discovery studies to focus exclusively on coding exons [47, 57, 179].

Several features of SERPINA1 emphasize the importance of its transcripts and their regulation. The SERPINA1 gene is exceptionally complex; eleven different splicing isoforms occur in human tissues [80]. While alternative splicing occurs in 95% of human multiexon genes [15, 115], the
eleven SERPINA1 transcripts are extreme, placing SERPINA1 in the top 0.5% of human genes in terms of transcriptional complexity [80]. A particularly salient feature of this SERPINA1 alternative splicing is that all variants differ only within their 5’UTRs [8, 115]. Therefore, all SERPINA1 mRNA isoforms code for the same α-1-antitrypsin protein; however, their differing 5’UTRs likely determine transcript-specific differences in post-transcriptional processes such as mRNA translation efficiency, subcellular localization, and stability [122, 148, 166]. Importantly, the SERPINA1 transcript isoforms are differentially expressed across tissue types [102], suggesting that post-transcriptional regulatory mechanisms adjust α-1-antitrypsin production based on the transcripts available to each tissue. The presence of up to three upstream open reading frames (uORFs) in the SERPINA1 5’UTRs [58, 121] suggest a potentially important yet unstudied mechanism for the translation efficiency regulation of these transcripts. In addition to the sequence-based differences between SERPINA1 transcripts, RNA secondary structure differences in the 5’UTR could also determine their regulation [61, 78, 123]. We propose here that non-coding features of SERPINA1 transcripts make up a post-transcriptional regulatory program that ultimately determines α-1-antitrypsin expression. We describe here a complex interplay between alternative splicing and translation efficiency mediated by uORFs and RNA structure, which together control tissue specific expression of α-1-antitrypsin in humans. Our quantitative and predictive model reveals an important and overlooked aspect of α-1-antitrypsin deficiency and suggests novel, RNA based targets for therapeutic consideration.

3.3 Results and Discussion

3.3.1 Transcript complexity in SERPINA1

Our motivation for studying SERPINA1 expression is the exceptional number of transcript isoforms produced by this clinically important gene. Two transcription start sites (TSS), six splicing donor (SD) and three acceptor (SA) sites yield a total of eleven transcript isoforms (Figure 3.1A and Supplementary Figure B.1) [172], which places SERPINA1 in the top 0.5% of transcriptionally complex human genes [80]. Interestingly, all of the alternative splicing occurs in the 5’UTR of SERPINA1 mRNA (Figure 3.1A). Thus, in healthy adults, α-1-antitrypsin exists as a single protein isoform that is produced from eleven different mRNAs. Multiple COPD and α-1-antitrypsin
deficiency-associated variants have been mapped to the SERPINA1 coding and non-coding sequences 3.1A [153]. In this study, we seek to determine whether the mRNAs are functionally different and how any differences relate to α-1-antitrypsin production or deficiency. Thus, we began this investigation by quantifying the expression of the various SERPINA1 transcripts in human tissues. Using data from the Illumina BodyMap 2.0 transcriptome-wide RNA-seq project, we quantified the relative amount of total SERPINA1 transcripts in 16 human tissues (Figure 3.1B); Figure 3.1C shows the relative amount of each SERPINA1 transcript in the form of a heat map. There are clear differences in the total amount of SERPINA1 present in each tissue. Liver noticeably yields the highest total SERPINA1 read counts (Figure 3.1B), which reflects the fact that α-1-antitrypsin is primarily expressed by hepatocytes and secreted into the bloodstream [18, 29, 90]. While the lungs are traditionally thought to acquire α-1-antitrypsin from the bloodstream [14, 29, 53], we found that lung tissue transcribes nontrivial amounts of SERPINA1 (Figure 3.1B), thus potentially producing its own α-1-antitrypsin. Although some SERPINA1 transcript isoforms are more prevalent than others, we detected all of the transcripts in the tissues, with some tissues like liver expressing every transcript (Figure 3.1C). To verify these findings with greater specificity, we designed SERPINA1 5’UTR-specific primers and amplified RNA extracted from liver and lung epithelial cells (HepG2 and A549 cell lines, respectively). All eleven transcripts were expressed in HepG2 cells, and all save one in A549 cells (Figure 3.1C). The varied expression of the SERPINA1 transcript isoforms across the tissues suggests that every SERPINA1 transcript has a distinct post-transcriptional function. Given that these transcripts vary only in their 5’UTR, we hypothesize that the splicing complexity in the 5’UTR of SERPINA1 plays an important role in its post-transcriptional regulation, especially, as detailed next, in SERPINA1 mRNA translation.

3.3.2 Translation efficiency analysis

The 5’UTR in mRNA regulates the translation of coding sequence and ultimately controls the expression of protein products [3]. To test the effect of different SERPINA1 5’UTRs on mRNA translation, we cloned upstream of firefly luciferase the 5’UTRs of six representative SERPINA1 transcripts. The constructs were transfected into HeLa cells and luciferase activity measured 24 hours post transfection. In order to determine the relative translation efficiency of each SERPINA1 5’UTR, we normalized luciferase activity to luciferase transcript abundance. This approach directly compares
Figure 3.1: The SERPINA1 gene produces 11 splice isoforms that all encode the same protein. (A) All exons in SERPINA1. Coding determinant sequence (CDS) exons are shown in red, untranslated regions (UTRs) in blues. Each exon is identified by a unique name. All alternative splicing in SERPINA1 occurs in the 5'UTR. Each splice donor and acceptor is identified by a unique name (SD1a, SA2, etc.) SERPINA1 also has two possible transcription start sites, 'TSS1' and 'TSS2.' Disease-associated variants, as catalogued by the Human Gene Mutation Database, are indicated with black marks and the common α-1-antitrypsin deficiency-associated Pi*S and Pi*Z alleles are labeled. Upstream open reading frames (uORFs) are indicated by red boxes and named. uORF δ/δ' spans a splice junction and is present only in isoforms with exon E1b.2. (B) The total amount of expressed SERPINA1 differs across 16 human tissue types. Total SERPINA1 transcript amounts were estimated from the Illumina BodyMap 2.0 project and are shown in log relative transcripts per million (TPM). (C) The SERPINA1 transcript isoforms are expressed, with different frequencies, across different tissues. Transcripts are specified with their NCBI names. The log(TPM) of each SERPINA1 transcript is shown for each tissue and for A549 and HepG2 cells. TPMs are relative to liver, which expresses the most SERPINA1 and is set to a total of $10^6$.

the amount of transcript to the amount of protein produced in each experiment, thus controlling for differences in transfection efficiency in a highly reproducible manner [79]. We further normalized translation efficiency measurements for each set of replicates to a control construct to correct for
technical variation between experiments; the translational efficiencies of the 5′UTRs are reported in Figure 3.2A. We found significant differences in translation efficiency between the different SERPINA1 5′UTRs (Figure 3.2A). Alternative splicing determines the inclusion (or exclusion) of up to three upstream open reading frames (uORFs) in the final SERPINA1 transcript isoform [58] (Figures 3.1A, 3.2B and Supplementary Figure B.1). Because uORFs can affect translation efficiency [17, 48], the uORFs in SERPINA1 may modulate translation of the different transcripts (an idea acknowledged decades ago [121] but untested until now). To evaluate the effect of uORFs on SERPINA1 translation, we mutated the start codon of a single uORF in each luciferase construct from ‘AUG’ to ‘AAG’ (Figure 3.2B). In total, the uORFs selected for mutation showed the effect of individually mutating every possible SERPINA1 uORF. Although it is possible that translation initiation at the mutated start codons could still occur [50, 156, 159], the estimated efficiency of the mutated start codons is very low—between 0-3% [120].

Mutating the uORF start codon(s) resulted in large increases in the translation efficiency of three of the six transcripts (Figure 3.2C), suggesting that the uORFs selected for mutation in these transcripts inhibit translation. The three transcripts with inhibitory uORFs are NM_000295.4, NM_001002236.2 and NM_001127705.1, and their mutated uORFs are uORFγ, uORFδ and uORFδ′, respectively (Figure 3.2B). It should be noted, however, that uORFγ is physically too close to the end of its respective 5′UTR to be translated [44], directly contradicting our results that it is functional in luciferase assays. A possible explanation for this result could be the use of an upstream ”late early” SV40 transcription start site [171] in our luciferase constructs that adds additional sequence to the 5′ end of each transcript, allowing uORFγ to be translated. If this is the case, uORFγ represents a dormant functional uORF in SERPINA1 transcript NM_000295.4. Exactly why uORFδ, uORFδ′ and the ”new” uORFγ are functional while the other SERPINA1 uORFs are not is of particular interest.

3.3.3 Modeling translation efficiency

The inhibitory uORFs identified above all have different sequences. However, closer inspection revealed that the three uORFs share the same initiation site, suggesting that features of this initiation site reduce translation efficiency in our assay. The shared initiation site among uORFγ, uORFδ and uORFδ′ encompasses most of the Kozak sequence, a well-characterized element that determines
Figure 3.2: Translation efficiency (TE) differs between SERPINA1 transcripts and is affected by uORFs. (A) The translation efficiencies of six SERPINA1 5’UTRs and their standard deviations, as measured by luciferase reporter assays. Transcripts are labeled by NCBI name (below). Measurements are relative to the luciferase assay control. The number of uORFs in each transcript is also indicated. (B) A schematic of the SERPINA1 luciferase constructs and pgl3 control, colored to match the results in Figure 3.2A and 3.2C. Constructs consist of 5’UTRs from the selected transcripts followed by luciferase CDS (CDS not to scale). 5’UTRs are segmented into their component exons (to scale). The uORFs in each transcript are indicated with Greek letters and shaded by Kozak sequence score as per the color scale. uORFs we selected for mutation are indicated with red arrows. *uORFγ has an incomplete Kozak sequence due to its proximity to the transcription start site, but its Kozak sequence strength was estimated by using an additional three bases upstream as part of its Kozak sequence. (C) TEs of the six SERPINA1 constructs with disrupted (mutated) uORFs and their standard deviations, as measured by luciferase reporter assays. TEs for each transcript are relative to the wild type (above). (D) The TEs of wild type and uORF mutant SERPINA1 constructs were predicted with a ‘leaky scanning’ model of translation. The model uses the Kozak sequence score(s) of the CDS and any uORFs. The plot shows fitted TE predictions from this model versus the measured TE of wild type and uORF mutant constructs ($r^2 = 0.400$, N=12).

translation initiation efficiency [75, 96, 112]. Thus, we retrieved the translation initiation strength of the Kozak sequence in each uORF, as measured previously for every possible Kozak sequence [112], and found that uORFγ, uORFδ and uORFδ’ have the strongest Kozak sequences of the SERPINA1 uORFs (greyscale in Figure 3.2C). Confident that uORFs and their Kozak sequences play an important role in regulating translation, we next used this information to model the differences in translation efficiency between the SERPINA1 transcripts. We first modeled translation efficiency with
a previously derived 'leaky scanning' model of translation [48], which we expanded to accommodate multiple uORFs (see Methods, equation 3.1). The model assumes the scanning mechanism of translation, whereby ribosomes migrate along the 5’UTR until encountering a start codon, and calculates the probability that ribosomes "leak through" any uORFs to translate the coding sequence [77]. The 'leaky scanning' model is based solely on the strength of the Kozak sequence of each open reading frame. However, the model poorly predicted our luciferase assay data ($r^2$ value of 0.4; Figure 3.2D), suggesting that other features in the SERPINA1 transcripts heavily influence their translation.

One additional factor that can tune translation efficiency is mRNA secondary structure. Evidence for the effect of secondary structure on translation has been conflicting [48, 61, 78, 112, 123], but such studies have typically relied on theoretical structure prediction, which falls far short of the accuracy achieved with direct chemical probing experiments [149].

### 3.3.4 Secondary structure of SERPINA1 transcripts

Recent advances in RNA structural mapping techniques, in particular Selective 2’ Hydroxyl Acylation by Primer Extension and Mutational Profiling (SHAPE-MaP) [149], have enabled accurate, high-throughput whole-transcript structural interrogation of RNA [131, 137, 154]. SHAPE-MaP interrogates the reactivity of each 2’ hydroxyl in an RNA, and the relative reactivities are used to estimate each base’s tendency to be structured (i.e. paired) or unstructured (i.e. unpaired). To measure structure differences between the SERPINA1 transcripts, we performed SHAPE-MaP separately on the six SERPINA1 transcript isoforms analyzed by luciferase assays. Performed in duplicate, the resulting data are highly correlated between replicates (Supplementary Figure B.3), with average correlation coefficients of 0.9 or more. Our experimental SHAPE-MaP data provide SHAPE reactivity profiles at nucleotide resolution for each of the six SERPINA1 transcripts. Regions with lower median SHAPE values (low SHAPE reactivities) consist of largely unreactive nucleotides (Figure 3.3A), whereas regions with higher median SHAPE values indicate the reverse (Figure 3.3B). The median-centered SHAPE reactivities of each transcript illustrate the relative reactivity of regions in the transcripts and indicate structured regions (Figure 3.3C). The high-reproducibility of SHAPE-MaP is immediately apparent in the median-centered SHAPE profiles: the reactivity patterns in the coding determinant sequence (CDS) across the six transcripts are nearly identical (Figure 3.3C).
This result is not surprising because all of the transcripts share the same CDS. Although long-range interactions in large RNAs do occur, local structure likely dominates the folding of mRNAs [49, 83], as suggested here by the similarity of the SHAPE data in the CDS of the six transcripts (Figure 3.3C). In addition, the SHAPE reactivities of residues in shared exons in the 5'UTR are generally comparable despite existing in unique contexts in the different transcripts. This observation suggests that folding in the 5'UTR is largely dominated by local structure. We combined our SHAPE-MaP data with RNA folding software [34, 91] to predict the minimum free energy structures of the various SERPINA1 transcripts (Figure 3.3D, Supplementary Figure B.5). We were especially interested in the structures of the uORFs and the beginning of the CDS and how this information could be used to model translation efficiency.

3.3.5 Modeling translation efficiency with structure

We next sought to gain a quantitative understanding of the contribution of structure to translation. The interplay of transcript structural elements with the translational machinery is not well understood, although studies in bacterial and mammalian systems suggest that structures near the start codon most likely affect mRNA translation [56, 61, 78]. We established above that the uORFs in SERPINA1 affect translation efficiency, but found that a model that accounts for only Kozak sequence strength did not quantitatively explain a large portion of the translation efficiency differences (Figure 3.2D). We hypothesized that, in addition to Kozak sequence strength, the model requires structural data encompassing the Kozak sequence to accurately capture the probability of the ribosome initiating at a given open reading frame. Our SHAPE-MaP data provided us with a high confidence structure of each transcript (Figure 3.3D and Supplementary Figure B.5), including the structures surrounding each Kozak sequence (Figures 3.4A, 3.4B and Supplementary Figure B.4). Predictions in prokaryotes suggest that translation initiation occurs in proportion to the exponent of the $\Delta G$ of folding of the local structure [140]. Thus, we modified the 'leaky scanning' model from equation 3.1 to include the $\Delta G$ of folding around the Kozak sequence (Methods, equation 3.2). The new 'structure leaky scanning' model dramatically improved the predictive power of the model to 94% (MLE ratio = .136, Figure 3.4C). The structural terms in the new model weigh each Kozak sequence by its accessibility in addition to its strength. From their location in uORF secondary structures, it is immediately clear that not all the uORF Kozak sequences are equally accessible (Figures 3.4A and 3.4B).
Figure 3.3: Secondary structure data collected on SERPINA1 transcripts with SHAPE-MaP. (A) The SHAPE reactivity of each base in a region of low median SHAPE values around the start codon of transcript NM_001002236.2. Each bar is shown with its standard error bars and colored by its SHAPE reactivity according to the color scale. Bases are numbered with their relative position within the transcript; the start codon is labeled with '+1'. (B) The SHAPE reactivity of each base in a region of high median SHAPE values in the coding sequence of transcript NM_001002236.2. Each bar is shown with its standard error bars and colored by its SHAPE reactivity according to the color scale. Bases are numbered with their relative position within the transcript. (C) The windowed, median-centered SHAPE profiles of six SERPINA1 transcripts ordered from longest to shortest. Regions below the median have low SHAPE reactivity and are more likely to be highly structured (base-paired). Regions above the median have high SHAPE reactivity and are less likely to be structured (base-paired). The locations of uORFs and coding sequences are indicated. uORFs are indicated with grey shaded regions and named with Greek letters. Vertical bars separate exons. All the transcripts share the same coding sequence, and their SHAPE profile patterns in this region are highly similar. (D) The minimum free energy (MFE) secondary structure of transcript NM_001002236.2, predicted by computational folding with SHAPE reactivity information. The transcript’s uORFs are indicated with grey shaded regions and labeled by name.

example, the Kozak sequence for uORFδ resides in an unstructured loop, while the Kozak sequence for uORFα is engaged in a stem-loop structure. It appears that uORFs γ, δ’ and δ are the only uORFs that have a Kozak sequence that is both strong and structurally accessible (Figures 3.2C, 3.4A-B and Supplementary Figure B.4), potentially explaining why only these uORFs inhibit SERPINA1 translation in our assays. Based on this data and our model, we propose that the Kozak sequence determines the likelihood of initiating translation, but the secondary structure determines whether
the Kozak sequence can actually be accessed. Thus, the translation efficiency of each SERPINA1 transcript is a combination of the initiation strength and structure of its CDS Kozak sequence, attenuated by the translation efficiency of any uORFs as established by the same parameters.

Figure 3.4: The incorporation of structural data greatly improves the 'leaky scanning' model of translation efficiency (TE) (A) The SHAPE-based predicted structures around the uORFs and coding sequence start in transcript NM_001002236.2. uORFs are labeled by name. Bases are colored according to their SHAPE reactivity, as measured by SHAPE-MaP. Bases with unknown SHAPE data are colored grey. Kozak sequences are boxed. The more rigid structures around the Kozak sequence in uORFs α and β may explain why these uORFs do not affect TE in the luciferase assays. (B) The SHAPE-based predicted structures around the uORF and coding sequence start in transcript NM_000295.4. Bases are colored according to their SHAPE reactivity, as measured by SHAPE-MaP. Bases with unknown SHAPE data are colored grey. Kozak sequences are boxed. (C) The ‘structure leaky scanning’ model modifies the ‘leaky scanning’ model to incorporate structural information around Kozak sequence(s) of the CDS and any uORFs. The plot shows fitted TE predictions from this model versus the TE of wild type and uORF mutant constructs as measured by luciferase assays ($r^2 = 0.936, N=12$).
3.3.6 Modeling α-1-antitrypsin expression in tissue

A goal of transcriptomics is to develop models that accurately describe the dynamics and effects of transcripts in living tissue. As we have seen from tissue-specific transcriptome data, SERPINA1 transcription is not limited to the liver, and different tissues express different combinations of the SERPINA1 transcript isoforms [102] (Figures 3.1B and 3.1C). Thus, optimized combinations of SERPINA1 transcripts could regulate the amount of α-1-antitrypsin protein produced in each tissue. Based on available protein quantification data (66), we calculated the overall SERPINA1 translation efficiency in each tissue as the ratio of α-1-antitrypsin to SERPINA1 transcript totals. If the translation efficiency of SERPINA1 mRNA is equal in every tissue, then we expect to observe that α-1-antitrypsin amounts and total SERPINA1 transcript amounts are correlated. However, we observed no such correlation (Figure 3.5A), indicating that different tissues have different net α-1-antitrypsin translation rates, potentially due to their unique combinations of SERPINA1 transcript isoforms. Assuming that the overall translation efficiency in a tissue is the average of the translational efficiencies of all its SERPINA1 transcripts weighted by abundance, we can use the models above to predict SERPINA1 translation efficiency in tissues (equation 3.3). While our luciferase assays show uORFγ to be functional, we have reason to believe this uORF escapes translation in its native setting, due to its extreme proximity to the 5’ end of transcript NM_000295.4. Thus we assessed our models in the case where uORFγ is functional (Supplementary Figure B.6) and in the case where it is non-functional and find that the latter better predicts translation efficiencies in our tissue data. The ‘leaky scanning’ model of translation (Figure 3.2D, equation 3.1) explains 59% of the variation in translation efficiency between tissues (Figure 3.5B), whereas the ‘structure leaky scanning’ model (Figure 3.4C, equation 3.2) explains 66% (Figure 3.5C). The addition of structural data to the model of translation thus improves predictions of translation efficiency in human tissues. It is interesting to note that both models considerably overestimated SERPINA1 translation efficiency of liver tissue. Liver tissue exports most of its α-1-antitrypsin into the bloodstream [18, 29, 90], which likely results in an underestimation of the amount of α-1-antitrypsin made in the liver and thus an underestimation of its observed translation efficiency (relative to predicted). This artifact indicates a need for cellular import/export dynamics to inform models of protein expression in tissues.
Figure 3.5: Predictions of SERPINA1 translation efficiency (TE) in 10 human tissues are improved with the ‘structure leaky scanning’ model. (A) Regression of total SERPINA1 transcript measurements with α-1-antitrypsin protein measurements shows no correlation between the two ($r^2 = 0.0$, N=10). Protein measurements are in normalized spectral counts [73]; transcript measurements are in Transcripts per Million (TPM). (B) The plot shows the ‘leaky scanning’ model’s predictions of TE versus measured TE in each tissue. This model explains 59% of the variation in TE between tissues (N=10). Each tissue is labeled and colored in the plot and in the human figure according to its percent error in the model’s predictions (equation 3.4). Errors were colored on a log scale; colors are shown on the scale. (C) The plot shows the ‘leaky scanning’ model’s predictions of TE versus measured TE in each tissue. This model explains 66% of the variation in TE between tissues (N=10). Each tissue is labeled and colored on the plot and in the human figure according to its percent error in the model’s predictions. Errors were colored on a log scale; colors are shown on the scale.

3.3.7 Discussion

The amount of protein produced from a gene is not a simple function of the amount of transcript the gene produces [24, 62, 97, 99]. The complex path between transcript expression and protein expression is too often a missing link in our understanding of cellular phenotype, indicating a need for integrative models that bridge this divide. SERPINA1 is exemplary of the effects of post-transcriptional regulation on protein output. While each of the SERPINA1 transcripts produces the same protein isoform, they do so with different efficiencies. Differences in uORF content and 5’UTR secondary structure combine to differentiate the translational efficiencies of SERPINA1 transcripts. Secondary structure plays a surprisingly important role in accounting for these differences, and in determining the repressive effect of individual SERPINA1 uORFs. When considering the role of secondary structure in a system, the elusive nature of RNA secondary structure demands more than a cursory computational prediction. Structural data accurate enough for successful biological models requires comprehensive chemical or enzymatic probing of the RNA molecules of interest [38, 149, 169]. In this study, we used SHAPE-MaP as our chemical probing method and successfully
applied the resulting structural data to improve mRNA translation efficiency predictions. Our results show that both uORFs and secondary structure are essential to modeling translation efficiency of SERPINA1 mRNA. Previously, no correlation was found between secondary structure and translation rate in experiments that measured the protein expression of constructs with varied uORF or CDS Kozak sequences [48, 112]. In these studies it is likely that the RNA structure predictions, which were purely computational, were inadequate for predicting structures around Kozak sequences. A biophysical model derived in one of the previous studies [48], referred to here as the 'leaky scanning’ translation model, performed poorly compared to our model incorporating structural data when we applied it to predicting SERPINA1 translation data (Figures 3.2D and 3.4C). Our data clearly demonstrates that the 'leaky scanning’ model alone is less accurate than a model that incorporates experimentally derived RNA secondary structure [34], probably because the 'leaky scanning’ model fails to distinguish the true effects of the various uORFs as determined by their structural contexts. Our results suggest a revised model of the leaky scanning mechanism, in which the ribosome is unable to initiate translation at a Kozak sequence if the surrounding secondary structure is stable, resulting in continued ribosome scanning.

Transcript-specific translation efficiencies may play an important role in tissue specific protein expression, especially in the case of α-1-antitrypsin, which shows a complex and varied expression pattern across human tissues. Our structural model greatly improved predictions of SERPINA1 translation efficiency among tissues (Figure 3.5). However, overall α-1-antitrypsin output in a tissue is not solely a consequence of translation efficiency. Transcripts travel through a coordinated post-transcriptional program, or 'regulon’ [70], with the possibility of diverging from their fellow isoforms at each step. This post-transcriptional program could be one reason why our model could not fully explain the high translational variation of α-1-antitrypsin amongst human tissues. Tissues could also have different overall rates of translation (e.g. in a fast-growing versus slow-growing tissue) or different rates of protein export that may have further confounded our predictions. These additional layers of regulation likely explain why our model of translation efficiency performs better in tissue culture cells than in tissues. However our model has still provided novel insights into the regulation of α-1-antitrypsin expression in tissues. First, liver tissue is always an extreme outlier in our models of SERPINA1 translation efficiency. Interestingly, predicting much higher translational efficiencies in liver tissue than observed based on measured levels of α-1-antitrypsin (Figure 3.5B-C)
is actually consistent with the understanding that liver exports most of its \( \alpha \)-1-antitrypsin into the bloodstream \([18, 29, 90]\). Conversely, our model predicts translation efficiency in lung tissue fairly accurately, suggesting that translation of \( SERPINA1 \) mRNA is a major source of its \( \alpha \)-1-antitrypsin in contrast to the paradigm that lung tissue derives its \( \alpha \)-1-antitrypsin from the bloodstream \([14, 30, 53]\). While contaminating blood cells could account for some portion of \( SERPINA1 \) transcripts in lung tissue, our detection of \( SERPINA1 \) transcripts in cultured lung cells (A549 cells, Figure 3.1C) and recent quantification of \( SERPINA1 \) transcripts in lung tissue \([102]\) further support the conclusion that cells in the lung itself express \( \alpha \)-1-antitrypsin. The close match between the lung’s protein-to-transcript ratio and our predictions suggests that most of lung’s \( \alpha \)-1-antitrypsin is self-produced. This surprising conclusion contradicts current models of the role of \( \alpha \)-1-antitrypsin in disease. The most common genetic variant in \( SERPINA1 \) associated with COPD and \( \alpha \)-1-antitrypsin deficiency, the Pi*Z allele, is thought to cause \( \alpha \)-1-antitrypsin to be poorly exported from the liver, leading to deficient \( \alpha \)-1-antitrypsin levels in the lungs and eventual neutrophilic overload \([14, 30, 53, 127]\). If lung tissue produces its own \( \alpha \)-1-antitrypsin, however, then this disease model is likely incomplete. Instead, disease-associated variants must also impact \( \alpha \)-1-antitrypsin levels in lung tissue, either by producing unviable \( \alpha \)-1-antitrypsin or greatly reducing its translation. For example, genetic variants could reduce \( \alpha \)-1-antitrypsin production if they shift the bulk of \( SERPINA1 \) transcription to isoforms with the lowest translation efficiencies. A potential therapy for such an imbalance would be a small RNA or ssDNA designed to target the uORFs in \( SERPINA1 \) that repress translation, i.e., uORF\( \gamma \), uORF\( \delta \) and uORF\( \delta ' \). A recent study quantified \( \alpha \)-1-antitrypsin and different \( SERPINA1 \) transcripts in the serum of \( \alpha \)-1-antitrypsin deficiency patients and healthy controls to determine if patients have different combinations of the transcripts \([102]\). Unfortunately the primer design in that study did not differentiate between the transcripts with the lowest and highest translational efficiencies, but the investigators did show a change in transcript proportions for at least one patient population \([102]\). Our findings here illustrate that we cannot underestimate the importance of the numerous \( SERPINA1 \) transcript isoforms in our understanding of disease and the impact of post-transcriptional regulation and secondary structure on phenotype in general.
3.4 Materials and Methods

3.4.1 SERPINA1 annotation

The known SERPINA1 transcript annotations were taken from RefSeq version hg38. In each transcript, uORFs are defined by a start and stop codon in the same frame within the 5’UTR. Distinct uORFs are named here with the Greek letters α, β, γ, δ and δ’.

3.4.2 Heat map of tissue-specific isoform expression

Paired-end RNA-seq reads from 16 different tissues were downloaded from the Illumina BodyMap 2.0 project (GEO accession number GSE30611). The tissues encompass 16 normal human tissue types: adrenal, adipocyte, brain, breast, colon, heart, kidney, leukocytes, liver, lung, lymph, ovary, prostate, skeletal muscle, testes and thyroid. Abundance estimates of the 11 known SERPINA1 transcripts were quantified with Sailfish version Beta 0.7.6 [118], using the full human transcriptome (RefSeq version hg38) as the reference. Estimates of total SERPINA1 expression in each tissue were calculated as the sum of transcript per million (TPM) estimates of each transcript. For better viewing in Figure 3.1B-C, total expression in liver was adjusted to $10^6$, and all other tissues’ TPM measurements were adjusted by the same factor.

3.4.3 Cell line specific transcript expression

A549 and HepG2 cells were provided by the Tissue Culture Facility at UNC-Chapel Hill. RNA was isolated using Trizol. Using Phase-lock heavy (Eppendorf) to remove the organic phase, the aqueous phase was then purified using a PureLink RNA mini kit (Life Technologies) and subjected to TurboDNase to digest DNA. The total RNA from each cell line was then reverse transcribed with Superscript III (Life Technologies and New England Biolabs Hot Start Q5 NEB) and amplified with 35 cycles in a reverse transcription polymerase chain reaction (RT-PCR). Because NM_000295.4 has a unique transcription start site, reverse transcription reactions with reverse primer GCCGCACGACAGAAGACGG were split into two different PCR reactions using forward primers TGGGCAGGAACTGGGCACCTG and ACAATGACTCCTTTCGGTAAGTGACGTGG to amplify
NM_000295.4 and all other transcripts, respectively. Following purification with a PureLink PCR cleanup kit (Life technologies), samples were assessed on an agarose gel. The double stranded DNA was then prepared using a Nextera DNA Library Prep Kit (Illumina). Following concentration determination via Qubit and library analysis using a Bioanalyzer, libraries were sequenced on a miSeq (Illumina). Isoform abundances in A549 and HepG2 cells were estimated with Sailfish version Beta 0.7.6 [118], mapping the sequenced reads to a reference that includes all known transcripts in RefGene hg38 excepting SERPINA1 isoform NM_000295.4. Abundance of transcript NM_000295.4 could not be compared directly to the other transcripts because it required a separate primer set. To estimate the relative abundance of NM_000295.4 in A549 and HepG2 cells, we used a dilution series amplified separately in 35 cycles of RT-PCR with two primer sets: ACTTAGCCCCTGTTTGCTCC (forward) and TGTCGATTCCTGTCGCAGG (reverse) for NM_000295.4 and ACCCTCAGAGTCCTCTGAGCTG (forward) and CTCTGTCTCTTCTGGCAGGC (reverse) for all other SERPINA1 transcripts. Both primer sets were designed to amplify 150 base pairs of sequence. Products from the dilution series of NM_000295.4 and other SERPINA1 transcripts from A549 and HepG2 cells were run on a 2% SEAkem GTG (Lonza) agarose gel and stained with 1X GelStar (Lonza). Product in each reaction was quantified with a gel imager. The quantifications of each dilution series were fit to logistic curves and inflection points were determined for the NM_000295.4 and other transcript curves. The ratio between the two inflection points was used as the ratio of NM_000295.4 transcript to all other SERPINA1 transcripts. TPM measurements for each transcript in A549 and HepG2 were adjusted based on their respective NM_000295.4:other ratios.

3.4.4 Luciferase assays

To assess the translation efficiency of SERPINA1 transcripts, we built six luciferase constructs containing 5’UTRs from selected SERPINA1 transcripts. Selected transcripts included NM_001002235.2, NM_000295.4, NM_001127700.1, NM_001127704.1, NM_001127705.1 and NM_00100236.2 (NCBI identifiers). The SERPINA1 5’UTRs were cloned via double digestion with SacII and NcoI into a modified pGL3 plasmid. The plasmid has an additional SacII site modified to minimize the amount of plasmid 5’UTR in the product. An SV40 site upstream of the SacII site drives transcription. For each of the SERPINA1 and control constructs, 0.5 µg of plasmid was
transfected into HeLa cells. Cells were harvested with Cell Culture Lysis Reagent (Promega # E153A) 24 hours post-transfection. Luciferase activity of the samples was measured by Luciferase Assay Substrate (Promega # E151C) and Luciferase Assay Buffer (Promega #E152B) with a luminometer (Molecular Devices). Luciferase activity measurements were taken in duplicate and averaged for each sample. The luciferase activity measurement for each sample was normalized to total sample protein concentration, as determined by Bradford Assay (N=4).

Luciferase measurements were further normalized to the abundance of luciferase RNA in each sample to obtain (luciferase activity)/(luciferase RNA). This normalization correlated the luciferase activity to luciferase RNA abundance, our proxy for translation efficiency. Without normalizing luciferase measurements by RNA abundance, changes in luciferase activity that resulted from variation in transfection efficiency or RNA stability could have confounded our interpretation. To quantify luciferase RNA abundance, after measuring luciferase activity, total RNA was extracted with TRIzol. Samples were depleted of DNA with Ambion Turbo DNA-free (AM1907) and reverse transcribed with High Capacity cDNA Reverse Transcription Kit (Applied Biosystems # 4368814). Luciferase and GAPDH cDNAs were quantified by real time PCR (qRT-PCR) on a BioRad CFX96 Real-Time System. Luciferase and GAPDH primers used were 5’-ACAAAGGCTATCAGGTGG- CT-3’ (forward), 5’-CGTGCTCCAAAACAACAACG-3’ (reverse) and 5’-CTGTTGCTGT- AGCCAAATTCGT-3’ (forward), 5’-ACCCACTCCTCCACCTTTGAC-3’ (reverse), respectively. Luciferase RNA abundance was determined by the ∆∆CT method (N=4). All luciferase/RNA measurements are reported relative to the control’s luciferase/RNA measurements to correct for systematic variations between experiments.

3.4.5 Upstream open reading frame (uORF) mutants

To disrupt uORFs in the original six SERPINA1 plasmid constructs, we designed primers to substitute the start codon of selected uORFs from AUG to AAG using the NEB Q5 site directed mutagenesis kit. uORFδ and uORFδ’ were mutated in the NM_001002236.2 and NM_001127705.1 luciferase plasmids using primers uORFT435A1F: CCAGGTACAAAGACTTTC/ uORFT435AR: CTCAGAAACCACAGCGTC. uORFβ was mutated in the NM_001127704.1 luciferase plasmid using primers uORFT285AF: ACTCAGT- AAAAGGTAGATCTTGCTAC/ uORFT285AR: CACCC-
CAAAATGCCTGATG. uORF$_\alpha$ was mutated the NM$_001002235.2$ and NM$_001127700.1$ luciferase plasmids using primers uORF32AF: GCCCAGGGCAAGCACTGCCTC/ uORF32AR: ACAGTGCCCGAGTTCTGCTCC. uORF$_\gamma$ was mutated in the NM$_000295.4$ luciferase plasmid using primers uORF4AF: CCGCGGACAAAGACTCTTTTC/ uORF4AR: CCTCGGCCTCTGCATA- AA. Mutant constructs were verified by sequencing. Luciferase assays were performed on the mutant constructs as above and results are reported in Supplementary Figure B.2.

### 3.4.6 SHAPE-MaP sequencing and analysis

First, the 5’UTRs and coding sequences of six selected SERPINA1 transcripts were cloned into pBLUNTII using overlap extension PCR and verified by sequencing. The selected transcripts are the same set analyzed by luciferase assays. Plasmids were named as follows - NM$_000295$: pAL0108, NM$_001002235$: pAL0096, NM$_001127700$: pAL0110 and pAL0111, NM$_001002236$: pAL0098, NM$_001127704$: pAL0100, NM$_001127705$: pAL0103 and pAL0105. Templates for transcription were amplified from 100 ng plasmid using Phusion high fidelity polymerase (NEB) and primers TAATACGACTCATATAGGGTGGGCGAGAACGTGGCACT (forward) and TTATTTTTGCTGGGATTCC- ACCAC (reverse) except for NM$_000295$:pAL0108, which required TAATACGACTCATATAGGGCAATGACTCCTTTCGGTAAGTGC as a forward primer. The T7 promoter was added by the forward primers. The PCR product was transcribed using a HiScribe™ T7 High Yield RNA Synthesis Kit (NEB) and the RNA was purified using an Ambion MEGAClear Transcription Clean-up kit (Thermo Fisher) or an RNEasy mini kit (Qiagen). Transcripts were verified using denaturing agarose gel electrophoresis with 2% SEAkem gold agarose and the Amresco Formaldehyde-Free RNA Gel Kit. 0.5 - 2 pmol of RNA was used for each modification reaction as described previously [149] with some modifications. Briefly, RNA was diluted in water, denatured at 95°C for 1 min and snap cooled on ice. After the addition of folding buffer (100 mM KCl, 10 mM MgCl$_2$, 100 mM HEPES pH 8.0 final concentration), the RNA was folded at 37°C for 10-15 minutes. Then 45 µl of folded RNA was either mixed with 5 µl DMSO (negative control) or 5 µl 100 mM 1-methyl-7-nitroisatoic anhydride (1M7) in DMSO (modified sample). After 5 min, reactions were desalted using G25 or G50 columns. A denatured control was performed in parallel in which the RNA was diluted into 50 mM HEPES pH 8.0, 4 mM EDTA and 50% formamide, then heated to 95°C and
treated with 5 µl 100 mM 1M7 in DMSO. After 1 minute, reactions were desalted using G25 or G50 columns. The RNA was reverse transcribed using SuperScript II (Life Technologies) and random nonamers as previously described [149] followed by clean-up with a G25 or G50 column. The second strand was synthesized using the NEBNext®mRNA Second Strand Synthesis Module (NEB). The double stranded DNA was then prepared using a Nextera or Nextera XT DNA Library Prep Kit (Illumina). Following concentration determination via Qubit and library analysis using a Bioanalyzer, libraries were run on a miSeq (Illumina) and resulting data were analyzed using the ShapeMapper pipeline [149] version 1.2. SHAPE-MaP sequencing data and processed SHAPE profiles are available in the NCBI Gene Expression Omnibus (GEO) accession number GSE81525. SHAPE data are also available in SNRNASM format at https://drive.google.com/open?id=1RpB9Jto1-UEmK-ocd9pGMOYrtet1ALuaA7XTyqI8ZA. Each transcript underwent SHAPE-MaP twice. SHAPE profiles averaged from the two experiments per transcript were used for later analyses requiring SHAPE data. The median SHAPE profiles in Figure 3c were generated for each transcript by calculating the median SHAPE value in windows of 20 bases (step size=1) and subtracting the global median.

3.4.7 Secondary structure analysis

Each transcript with SHAPE-MaP data was folded with RNAfold version 2.2.4 incorporating their respective SHAPE data with the –shape option and a max distance of 50 (–maxBPspan=50) to focus on local structures (Supplementary Figure B.5). 3’UTRs of the transcripts were excluded in structure prediction. Structures were also predicted for sequence constructs consisting of the 5’UTRs of the six transcripts followed by 200 bases of luciferase coding sequence to predict secondary structure in the 5’UTR of the luciferase constructs. SHAPE data for the 5’UTRs were still used in the structure prediction; SHAPE data for the luciferase coding sequence is unavailable. ∆G measurements were calculated around Kozak sequences in the predicted structures by removing base pairs that occur within +/- 15 bases around the ’A’ in the start codon. (Different window sizes were tested as well (Table B.1).) The free energy of the ”relaxed” structure was subtracted by the free energy of the original predicted structure to arrive at the ∆G of unfolding around the Kozak sequence. (The ∆G of unfolding = -∆G of folding.) The ∆G of unfolding was calculated
around the coding sequence and uORF Kozak sequences in the wild type transcripts and the luciferase constructs. ΔG values around the coding sequence differ between corresponding luciferase constructs and wild type transcripts because the luciferase constructs have different coding sequences. ΔG values around uORFs remain identical between corresponding luciferase constructs and wild type transcripts. SHAPE-MaP was not performed on uORF mutant SERPINA1 transcripts, but, because point mutations rarely cause perceptible changes in secondary structure [28, 151], we assumed that the structures of the wild type transcripts closely approximate the structures of the uORF mutants.

3.4.8 Models

Models were fit with simple linear regression to the (luciferase activity)/(luciferase RNA) measurements of the six SERPINA1 constructs and six uORF mutant constructs. R-squared values and predictor p-values were determined by the lm function in R version 3.2.3. The r-squared value shown with each model is the adjusted r-squared value, which adjusts the r-squared value based on the number of predictors used in the model. ΔG values used in these models were derived from the luciferase construct secondary structures (described above). ΔG values of uORF mutants were assumed to be identical to their wild type counterparts. The models we feature in the results are the 'leaky scanning' and the 'structure leaky scanning' models. These models, shown as equations 3.1 and 3.2, were used to calculate translation efficiency metrics for all SERPINA1 constructs and uORF mutants, which were subsequently fit to the luciferase data as above. Equation 3.1 is our expansion of a previously published model [48] to include multiple uORFs. Equation 3.2 is our variation of the model that incorporates structural data, as derived to be exponentially related to ribosomal initiation [140]. TE is "translation efficiency", k, k’, and l are constants, \(P_n\) is the strength of the given Kozak sequence as determined previously [112]. Kozak strengths are converted to probabilities by dividing by the maximum Kozak strength, 150. \(\Delta G_n\) corresponds to the ΔG of unfolding +/-15 bases around the given Kozak sequence (calculation described above). The subscripts of \(P_n\) and \(\Delta G_n\) indicate either the coding sequence or the nth uORF, numbered 5’-3’ in each transcript. \(P_n\) and \(\Delta G_n\) values for a transcript without an nth uORF are simply zero. Constants k’ and l were optimized in the
'structure leaky scanning' model and the constant $k$ is the original published value [48].

$$TE \sim kP_{cds}(1 - kP_1)(1 - kP_2)(1 - kP_3)$$  \hspace{1cm} (3.1)

$$TE \sim k'P_{cds}e^{-\Delta G_{cds}}(1 - k'P_1e^{-\Delta G_1})(1 - k'P_2e^{-\Delta G_2})(1 - k'P_3e^{-\Delta G_3})$$  \hspace{1cm} (3.2)

### 3.4.9 Models in tissues

Total tissue SERPINA1 concentrations (TPM) were fit to their $\alpha$-1-antitrypsin protein concentrations (ppm) (Figure 3.5A) with simple linear regression with the lm function in R version 2.3.2. The SERPINA1 transcript concentrations are described above (Figure 3.1B) and the $\alpha$-1-antitrypsin protein measurements are derived from a mass spectrometry data on the human proteome [73]. $\alpha$-1-antitrypsin protein data were unavailable for skeletal muscle and leukocytes. SERPINA1 translation efficiency in each tissue was measured by dividing $\alpha$-1-antitrypsin protein concentration by the total SERPINA1 concentration. To test a model with the tissues’ translation efficiency, the model’s translation efficiency estimates of all eleven SERPINA1 transcripts were used to predict the average translation efficiency of each tissue as in equation 3.3. $TE_j$ is the model-predicted translation efficiency of tissue $j$, $TPM_{i,j}$ is the transcript abundance in transcript per millions (TPMs) of SERPINA1 transcript $i$ in tissue $j$ and $m(i)$ is the function for the translation efficiency of transcript $i$ with parameters from fitting the model to the luciferase data.

$$TE_j \sim \frac{\sum_i m(i)TPM_{i,j}}{\sum_i TPM_{i,j}}$$  \hspace{1cm} (3.3)

The model-predicted values for tissue translation efficiency were then fit to the measured tissue translational efficiencies with simple linear regression. Models fit best to the log of the measured tissue translational efficiencies. Secondary structure information ($\Delta G$ values) is not available for transcripts NM_001127701.1, NM_001127702.1, NM_001127703.1, NM_001127705.1, NM_001127706.1 and NM_001127707.1. In their case, $\Delta G$ values were assigned based on the most similar transcript with available secondary structure data. $\Delta G$ values used in tissue predictions were derived from the
SERPINA1 transcript secondary structures (described above). Percent error of the ‘leaky scanning’ and ‘structure leaky scanning’ models (Figure 3.5B-C) in each tissue was calculated according to equation 3.4, where TE is ‘translation efficiency’.

\[
\text{Error}(\%) \sim \frac{|\text{modelPredictedTE} - \text{measuredTE}|}{\text{measuredTE}}
\]  

(3.4)
CHAPTER 4

Conclusion

In the RNA World hypothesis, the beginnings of life were run by RNA molecules which served the role of both DNA and protein [81]. I will argue that life still harbors an RNA world: in the past decades an explosion of different RNA classes have been discovered and characterized [4, 7, 46, 65, 136, 165]. Even messenger RNAs (mRNAs), once thought to simply carry information from gene to protein, constitute a complex post-transcriptional world. In eukaryotes mRNAs undergo splicing, poly-adenylation, export from the nucleus, localization in the cytoplasm, translation and degradation, where at each step they interact with RNA binding proteins and other RNAs. Both mRNA sequence and secondary structure play a role in these processing steps, but it is unclear how functional mRNA secondary structures are and how general their mechanisms. Accurate structure determination techniques will be crucial in answering these questions.

In Chapter Two I described a benchmark on computational methods that predict riboSNitches, structural differences between RNAs with single nucleotide variants. Generally we find that with computational prediction, structural changes were most accurately predicted by algorithms that use the base pairing probability matrix over the more popular minimum free energy structure. Thus it is important to consider the ensemble of structures an RNA may form when predicting structure in the absence of experimental data. If experimental data is to be had, however, we know that this greatly improves minimum free energy structure predictions [34].

Experimental techniques that probe RNA structure rely on reagent(s) that differentially react with paired nucleotides versus unpaired nucleotides, returning a structural profile across the nucleotides in an RNA molecule. This sort of structural profile does not indicate the particular partner each base pairs with, but as mentioned, can be incorporated into computational structure modeling to greatly improve base pair predictions. One caveat of the hybrid experimental-computational approach is that its accuracy was determined by benchmarking against RNAs with strong, well-defined structures [34].
Thus it is unclear how its predictions fare with less rigid mRNAs. A more recent structure probing method, PARIS [93], isolates and sequences double stranded RNA to identify specific base-paired regions at near nucleotide resolution. While this method does not necessarily indicate how isolated base-paired regions are arranged in a full RNA molecule, one can imagine that PARIS data combined with computational prediction could very powerfully direct RNA structure models. The implications of knowing exact structures for any transcript of interest are enormous for post-transcriptional RNA research.

Secondary structures in mRNAs have been shown to affect post-transcriptional regulation, especially translation, in a number of cases [6, 21, 41, 132, 162]. In Chapter Three I described a set of mRNAs transcribed from the gene SERPINA1 whose translation efficiency is determined by upstream open reading frames (uORFs) and their secondary structure contexts. In our quantitative model, the translation efficiency of the main coding sequence is determined by its start codon sequence context and secondary structure strength and penalized by the translation efficiency of any uORFs, as determined by the same rules. Developing this model required the accuracy of SHAPE-MaP structural probing data. However, validating the structural component of translation initiation will require additional experiments. We have designed additional mRNAs with either unstructured or strongly structured uORF contexts and will measure these mutants’ translation efficiencies. If a uORF’s structural context does indeed affect its translation efficiency, then we will expect to see a change in the mutants’ translation efficiencies as their uORFs are either strengthened or abolished. More generally, uORFs represent important regulatory regions that could potentially cause large changes in protein output if altered by a single nucleotide variant (SNV) [17]. A SNV that changes the local structural context (a riboSNitch) could affect protein output as well, if uORF structural context does indeed affect translation efficiency.

RNA secondary structure is an important characteristic across all classes of RNAs. As structural determination techniques become more sophisticated we will be able to ask more questions about the roles of secondary structures in the cell. For many classes of RNAs, including mRNAs, their secondary structures remain mostly unstudied and underestimated. Future questions include how conserved certain structures are, how RNAs form and maintain their structures and how structures may be disrupted in disease states.
## APPENDIX A

### Supplementary Material to Chapter 2

Table A.1: AUC values for algorithms’ ROC performance on the entire riboSNitch dataset and two subsets—the best 25% and best 5% as ranked by riboSNitch average fdr-adjusted p-value. 95% confidence intervals shown below each AUC value.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>All</th>
<th>Best 25%</th>
<th>Best 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentroidFold</td>
<td>0.532</td>
<td>0.565</td>
<td>0.552</td>
</tr>
<tr>
<td></td>
<td>(0.507-0.556)</td>
<td>(0.517-0.614)</td>
<td>(0.440-0.664)</td>
</tr>
<tr>
<td>CONTRAfold</td>
<td>0.535</td>
<td>0.566</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>(0.510-0.559)</td>
<td>(0.517-0.615)</td>
<td>(0.435-0.659)</td>
</tr>
<tr>
<td>MC-Fold</td>
<td>0.493</td>
<td>0.483</td>
<td>0.487</td>
</tr>
<tr>
<td></td>
<td>(0.464-0.522)</td>
<td>(0.439-0.528)</td>
<td>(0.381-0.563)</td>
</tr>
<tr>
<td>remuRNA</td>
<td>0.537</td>
<td>0.560</td>
<td>0.579</td>
</tr>
<tr>
<td></td>
<td>(0.513-0.562)</td>
<td>(0.532-0.613)</td>
<td>(0.480-0.637)</td>
</tr>
<tr>
<td>RNAfold</td>
<td>0.534</td>
<td>0.557</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td>(0.509-0.559)</td>
<td>(0.504-0.585)</td>
<td>(0.424-0.576)</td>
</tr>
<tr>
<td>RNAmute</td>
<td>0.511</td>
<td>0.508</td>
<td>0.455</td>
</tr>
<tr>
<td></td>
<td>(0.487-0.535)</td>
<td>(0.466-0.538)</td>
<td>(0.386-0.513)</td>
</tr>
<tr>
<td>RNAmutants</td>
<td>0.509</td>
<td>0.502</td>
<td>0.417</td>
</tr>
<tr>
<td></td>
<td>(0.485-0.533)</td>
<td>(0.464-0.539)</td>
<td>(0.397-0.540)</td>
</tr>
<tr>
<td>RNAsnp</td>
<td>0.533</td>
<td>0.564</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>(0.508-0.558)</td>
<td>(0.512-0.593)</td>
<td>(0.453-0.617)</td>
</tr>
<tr>
<td>RNAstructure</td>
<td>0.536</td>
<td>0.543</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>(0.511-0.560)</td>
<td>(0.502-0.582)</td>
<td>(0.440-0.588)</td>
</tr>
<tr>
<td>SNPfold</td>
<td>0.528</td>
<td>0.553</td>
<td>0.525</td>
</tr>
<tr>
<td></td>
<td>(0.504-0.553)</td>
<td>(0.494-0.574)</td>
<td>(0.416-0.570)</td>
</tr>
<tr>
<td>UNAFold</td>
<td>0.526</td>
<td>0.543</td>
<td>0.530</td>
</tr>
<tr>
<td></td>
<td>(0.502-0.551)</td>
<td>(0.493-0.592)</td>
<td>(0.418-0.643)</td>
</tr>
</tbody>
</table>
Table A.2: The sensitivity and specificity values and corresponding threshold of the most top left point in each ROC curve.

<table>
<thead>
<tr>
<th>Category</th>
<th>CentroidFold</th>
<th>contraFold</th>
<th>MC-Fold</th>
<th>remuRNA</th>
<th>RNAfold</th>
<th>RNAmute</th>
<th>RNAmute</th>
<th>RNAsnp</th>
<th>RNAstructure</th>
<th>SNPfold</th>
<th>UNAFold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probed Threshold</td>
<td>0.44</td>
<td>0.79</td>
<td>6.00</td>
<td>2.82</td>
<td>3.97</td>
<td>19.00</td>
<td>20.00</td>
<td>0.39</td>
<td>1.77</td>
<td>0.92</td>
<td>3.08</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.82</td>
<td>0.64</td>
<td>0.67</td>
<td>0.64</td>
<td>0.73</td>
<td>0.55</td>
<td>0.64</td>
<td>0.83</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.45</td>
<td>0.64</td>
<td>0.38</td>
<td>0.82</td>
<td>0.73</td>
<td>0.64</td>
<td>0.64</td>
<td>0.73</td>
<td>0.55</td>
<td>0.73</td>
<td>0.45</td>
</tr>
<tr>
<td>Validated Threshold</td>
<td>0.89</td>
<td>0.62</td>
<td>44.00</td>
<td>1.65</td>
<td>4.48</td>
<td>6.00</td>
<td>11.00</td>
<td>0.30</td>
<td>3.14</td>
<td>0.93</td>
<td>2.22</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.56</td>
<td>0.56</td>
<td>0.34</td>
<td>0.56</td>
<td>0.59</td>
<td>0.62</td>
<td>0.46</td>
<td>0.48</td>
<td>0.51</td>
<td>0.60</td>
<td>0.59</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.60</td>
<td>0.48</td>
<td>0.80</td>
<td>0.56</td>
<td>0.62</td>
<td>0.51</td>
<td>0.52</td>
<td>0.75</td>
<td>0.63</td>
<td>0.60</td>
<td>0.52</td>
</tr>
<tr>
<td>Symmetric Threshold</td>
<td>0.75</td>
<td>0.60</td>
<td>6.00</td>
<td>7.00</td>
<td>2.38</td>
<td>5.22</td>
<td>5.00</td>
<td>7.00</td>
<td>0.47</td>
<td>2.76</td>
<td>0.93</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.55</td>
<td>0.49</td>
<td>0.52</td>
<td>0.45</td>
<td>0.50</td>
<td>0.54</td>
<td>0.48</td>
<td>0.56</td>
<td>0.48</td>
<td>0.54</td>
<td>0.50</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.58</td>
<td>0.42</td>
<td>0.49</td>
<td>0.64</td>
<td>0.65</td>
<td>0.53</td>
<td>0.53</td>
<td>0.59</td>
<td>0.62</td>
<td>0.61</td>
<td>0.60</td>
</tr>
<tr>
<td>Asymmetric Threshold</td>
<td>0.71</td>
<td>0.50</td>
<td>17.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>0.49</td>
<td>2.52</td>
<td>0.95</td>
<td>1.98</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.52</td>
<td>0.54</td>
<td>0.45</td>
<td>0.50</td>
<td>0.58</td>
<td>0.54</td>
<td>0.51</td>
<td>0.53</td>
<td>0.48</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.52</td>
<td>0.44</td>
<td>0.56</td>
<td>0.56</td>
<td>0.47</td>
<td>0.51</td>
<td>0.51</td>
<td>0.53</td>
<td>0.57</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>All Threshold</td>
<td>0.65</td>
<td>0.50</td>
<td>17.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>0.47</td>
<td>2.52</td>
<td>0.94</td>
<td>2.09</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.55</td>
<td>0.52</td>
<td>0.45</td>
<td>0.51</td>
<td>0.53</td>
<td>0.54</td>
<td>0.50</td>
<td>0.52</td>
<td>0.49</td>
<td>0.55</td>
<td>0.51</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.49</td>
<td>0.44</td>
<td>0.56</td>
<td>0.55</td>
<td>0.52</td>
<td>0.50</td>
<td>0.51</td>
<td>0.55</td>
<td>0.58</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td>25p tails Threshold</td>
<td>0.106</td>
<td>0.068</td>
<td>31.000</td>
<td>4.711</td>
<td>0.224</td>
<td>31.000</td>
<td>31.000</td>
<td>0.084</td>
<td>10.479</td>
<td>0.674</td>
<td>8.748</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.631</td>
<td>0.460</td>
<td>0.662</td>
<td>0.516</td>
<td>0.622</td>
<td>0.502</td>
<td>0.557</td>
<td>0.507</td>
<td>0.556</td>
<td>0.674</td>
<td>8.748</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.548</td>
<td>0.458</td>
<td>0.664</td>
<td>0.517</td>
<td>0.656</td>
<td>0.514</td>
<td>0.616</td>
<td>0.505</td>
<td>0.554</td>
<td>0.549</td>
<td>0.487</td>
</tr>
<tr>
<td>5p tails Threshold</td>
<td>0.017</td>
<td>0.221</td>
<td>42.000</td>
<td>7.192</td>
<td>17.319</td>
<td>42.000</td>
<td>42.000</td>
<td>0.999</td>
<td>9.637</td>
<td>0.999</td>
<td>0.005</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.636</td>
<td>0.461</td>
<td>0.666</td>
<td>0.605</td>
<td>0.257</td>
<td>0.252</td>
<td>0.252</td>
<td>0.527</td>
<td>0.438</td>
<td>0.549</td>
<td>0.487</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.602</td>
<td>0.478</td>
<td>0.678</td>
<td>0.578</td>
<td>0.567</td>
<td>0.513</td>
<td>0.621</td>
<td>0.540</td>
<td>0.554</td>
<td>0.513</td>
<td>0.562</td>
</tr>
</tbody>
</table>

Table A.3: AUC values with 95% confidence intervals for algorithms tested with a Pearson correlation coefficient metric. Previous values from the RNApdist metric shown for comparison.

<table>
<thead>
<tr>
<th>Software</th>
<th>Metric</th>
<th>Probed</th>
<th>Validated</th>
<th>Symm</th>
<th>Asymm</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentroidFold</td>
<td>RNApdist</td>
<td>0.579</td>
<td>0.561</td>
<td>0.569</td>
<td>0.529</td>
<td>0.532</td>
</tr>
<tr>
<td></td>
<td>Pearson</td>
<td>0.628</td>
<td>0.559</td>
<td>0.563</td>
<td>0.532</td>
<td>0.530</td>
</tr>
<tr>
<td>CONTRAfold</td>
<td>RNApdist</td>
<td>0.463</td>
<td>0.562</td>
<td>0.567</td>
<td>0.528</td>
<td>0.535</td>
</tr>
<tr>
<td></td>
<td>Pearson</td>
<td>0.587</td>
<td>0.573</td>
<td>0.568</td>
<td>0.532</td>
<td>0.538</td>
</tr>
<tr>
<td>RNAstructure</td>
<td>RNApdist</td>
<td>0.612</td>
<td>0.578</td>
<td>0.567</td>
<td>0.527</td>
<td>0.536</td>
</tr>
<tr>
<td></td>
<td>Pearson</td>
<td>0.603</td>
<td>0.561</td>
<td>0.546</td>
<td>0.522</td>
<td>0.529</td>
</tr>
<tr>
<td>UNAFold</td>
<td>RNApdist</td>
<td>0.471</td>
<td>0.537</td>
<td>0.548</td>
<td>0.524</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>Pearson</td>
<td>0.455</td>
<td>0.518</td>
<td>0.526</td>
<td>0.517</td>
<td>0.520</td>
</tr>
</tbody>
</table>
Table A.4: Parameters used for each algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Unix Command Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentroidFold</td>
<td>$ centroid_fold –engine McCaskill -o output.txt –posterials 0 seq.fa</td>
</tr>
<tr>
<td>CONTRAfold(For MFE)</td>
<td>$ ./contrafold predict –posterials 0 output.txt seq.fa  $ ./contrafold predict –parens output.txt seq.fa</td>
</tr>
<tr>
<td>MC-Fold</td>
<td>$ export QUERY_STRING=&quot;pass=lucy&amp;sequence=&quot;$seq&quot;&amp;top=1&quot;;./mcfold.static.exe</td>
</tr>
<tr>
<td>remuRNA</td>
<td>$ ./remuRNA seq.fa</td>
</tr>
<tr>
<td>RNAfold(For MFE)</td>
<td>$ cat seq.fa — RNAfold p  $ cat seq.fa — RNAfold</td>
</tr>
<tr>
<td>RNAmutants</td>
<td>$ ./RNAmutants -l /programs/RNAmutants-master/lib -f seq.fa –mutation 0 –sample-number 0</td>
</tr>
<tr>
<td>RNAmute</td>
<td>$ java -jar RNAmute.jar seq.txt</td>
</tr>
<tr>
<td>RNAsnp</td>
<td>$ ./RNAsnp -f seq.fa -s snp.txt -w 100 -l 10 -c 0.0</td>
</tr>
<tr>
<td>RNAstructure(For MFE)</td>
<td>$ partition seq.fa out.pfs  $ Fold -mfe seq.fa out.ct</td>
</tr>
<tr>
<td>SNPfold</td>
<td>$ SNPfold_commandline.py &quot;$seq&quot; &quot;$ref&quot;51&quot;$alt&quot;</td>
</tr>
<tr>
<td>UNAFold(For MFE)</td>
<td>$ UNAFold.pl –model=PF seq.fa  $ hybrid-ss-min –suffix DAT seq.fa</td>
</tr>
</tbody>
</table>
Figure A.1: ROC performance of selected algorithms (A-D) on different input sequence lengths from the "probed" category of riboSNitches. Sequence lengths ranged from 21 to 201 in 6 base increments. The graphs indicate large variance with sequence length, although for some algorithms, performance remains robust within certain sub-ranges.
Figure A.2: ROC performance across riboSNitch categories on a range of sequence lengths for SNPfold, RNAstructure and a random function (A-C). The random function sampled scores from the combined set of all SNPfold riboSNitch and non-riboSNitch scores. The trend of decreasing AUC values from the "probed" category to the "all" category remains evident with SNPfold and RNAstructure and is absent in the more symmetrical random function.
Figure A.3: Algorithms’ distance score distributions for all riboSNitches and matched non-riboSNitches. The 5% tails of the distributions are shaded grey, though are too small to see in many cases. remuRNA returns the relative entropy between variant RNAs, which can be negative. SNPfold returns Pearson correlation coefficients instead of true distance scores and RNAsnp returns p-values. The structure distance functions RNAdistance and RNApdist were used with the rest of the algorithms, generating distance scores greater than or equal to zero. A) Score distributions for specialized algorithms. B) Score distributions for general algorithms.
APPENDIX B

Supplementary Material to Chapter 3

Figure B.1: Schematic of eleven (SERPINA1) transcripts. The eleven known SERPINA1 transcript isoforms are composed of combinations of the exons shown in Figure 3.1A in the main text. Blue regions make up the 5’ and 3’UTRs. Red exons are the α-1-antitrypsin coding sequence. uORFs in each transcript are indicated with light red boxes in the 5’UTR. Transcripts are ordered from largest to smallest. NCBI names of each transcript are listed to the right and Ensembl names to matching transcripts are listed to the left.

Table B.1: StructureLeakyScanning model performance using different window sizes to calculate ΔG around each open reading frame’s Kozak sequence.

<table>
<thead>
<tr>
<th>Window Size</th>
<th>R-squared</th>
<th>Optimal k’</th>
<th>Optimal i</th>
</tr>
</thead>
<tbody>
<tr>
<td>+/- 2</td>
<td>0.61</td>
<td>0.8</td>
<td>0.01</td>
</tr>
<tr>
<td>+/- 5</td>
<td>0.50</td>
<td>0.9</td>
<td>0.03</td>
</tr>
<tr>
<td>+/- 10</td>
<td>0.46</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>+/- 15</td>
<td>0.94</td>
<td>0.4</td>
<td>0.04</td>
</tr>
<tr>
<td>+/- 20</td>
<td>0.79</td>
<td>1.0</td>
<td>0.04</td>
</tr>
<tr>
<td>+/- 30</td>
<td>0.47</td>
<td>1.0</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Figure B.2: Translation efficiency of six SERPINA1 uORF-mutants relative to control, as measured by luciferase reporter assays. The uORF mutants are named according to their corresponding wild type SERPINA1 transcript, followed by "X" and the number of the uORF that was mutated, as numbered from 5’ to 3’ in a given transcript’s 5’ UTR.

Figure B.3: Correlation coefficients between SHAPE-MaP duplicates for each transcript. Correlation coefficients were calculated in a sliding window of length 100 over the entire length of each transcript.
Figure B.4: SHAPE-directed Minimum Free Energy (MFE) structures predictions around all Kozak sequences and uORFs in four SERPINA1 transcripts. Schematics for each transcript are shown, where different boxes represent different exons and grey boxes indicate the uORFs and coding sequence, which are labeled with Greek letters and 'CDS', respectively (CDS not to scale). The structures shown for each transcript are the structures around each Kozak sequence in the uORFs and CDS. Kozak sequences are outlined. Bases in the structures are colored by their SHAPE reactivity, as measured by SHAPE-MaP. Bases colored grey have unknown SHAPE reactivity.
Figure B.5: SHAPE-directed MFE structures predicted for five SERPINA1 transcripts (5’UTR and coding sequence). Vertical bars indicate the start of coding sequence. The transcripts have identical coding sequence and show very similar predicted structures in this region. SHAPE data from SHAPE-MaP experiments were used to predict each structure. Each base is colored to reflect its SHAPE reactivity. Bases colored grey have unknown SHAPE values.

Figure B.6: Model performance predicting translation efficiency in human tissues. uORF\(^\gamma\) in transcript NM_000295.4 assumed to be functional here. A) Under this assumption, the 'leaky scanning' model of translation efficiency is non-predictive. B) Adding secondary structure information in the 'structure leaky scanning' model greatly improves predictions.
REFERENCES


