COMPUTATIONAL TOOLS FOR CLASSIFYING AND VISUALIZING RNA STRUCTURE CHANGE IN HIGH-THROUGHPUT EXPERIMENTAL DATA

Chanin Tolson Woods

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

Chapel Hill
2017

Approved by:

Alain Laederach

Terry Furey

Elizabeth Shank

David Gotz

Shawn Gomez

ABSTRACT

Chanin Tolson Woods: Computational tools for classifying and visualizing RNA structure change in high-throughput experimental data
(Under the direction of Alain Laederach)

Mutations (or Single Nucleotide Variants) in folded RiboNucleic Acid (RNA) structures that cause local or global conformational change are riboSNitches. Predicting riboSNitches is challenging, as it requires making two, albeit related, structure predictions. The data most often used to experimentally validate riboSNitch predictions is Selective 2' Hydroxyl Acylation by Primer Extension, or SHAPE. Experimentally establishing a riboSNitch requires the quantitative comparison of two SHAPE traces: wild-type (WT) and mutant. Historically, SHAPE data was collected on electropherograms and change in structure was evaluated by "gel gazing." SHAPE data is now routinely collected with next generation sequencing and/or capillary sequencers. We aim to establish a classifier capable of simulating human "gazing" by identifying features of the SHAPE profile that human experts agree "looks" like a riboSNitch.

Additionally, when an RNA molecule folds, it does not always adopt a single, well-defined conformation. The folding energy landscape of the RNA is highly dependent on sequence and the molecular environment. Endogenous molecules, especially in the cellular context, will in some cases completely alter the energy landscape and therefore the ensemble of likely low-energy conformations. The effects of these energy landscape changes on the

conformational ensemble are particularly challenging to visualize for larger RNAs including most messenger RNAs (mRNAs). We propose here a robust approach for visualizing the conformational ensemble of RNAs particularly well suited for *in vitro* vs. *in vivo* comparisons.

To my parents, James and Pollie, who from my earliest days pushed me to always do the “hard thing”, and encouraged me whenever I came up short. Those lessons stayed with me and without them I would not be writing this dissertation.

To my husband, Makieal, who tries to remind me to laugh, to take naps, to eat the extra fish stick and to let it all go. You have been my sanity in an otherwise insane process. I love you.

And to the girls, who sent a box of macarons at just the right time.

ACKNOWLEDGEMENTS

There are many people who have helped me navigate my time in graduate school. The BBSP program staff has been vital in providing me with resources to succeed during my year in the umbrella program and long after. I especially want to acknowledge Jessica Harrell and Ashalla Freeman for helping me through a particularly rough patch. I also want to thank Olga Gonzalez-Lopez for mentoring me during this time. The BCB curriculum director, Tim Elston, and program administrators, John Cornett and Cara Marlow, have helped me to tailor the program's resources to best supplement my learning. My committee members have been assets in my development as a graduate student. They held me to a high standard and pushed me to understand the broader context of my work, particularly my committee chair Terry Furey. And David Gotz has provided guidance on data visualization and encouraged me to iterate over my design. I would like to acknowledge our collaborators Nikolay Dokholyan and Benfeard Williams for providing me with their 3D modeling expertise.

I want to thank two of the post-docs in my lab, Amanda Solem and Lela Lackey, for always answering my questions with patience and providing a sounding board for ideas. I also want to thank my fellow graduate students who have been on this journey with me: Meredith Corley, Aaztli Coria, Matthew Halvorsen, Katrina Kutchko and Wes Sanders. Finally, I want to thank my advisor Alain Laederach. He has been patient but persistent in helping me to transition from thinking like an "engineer" to thinking like a "scientist". As a bioinformatician, my advisor

is smart and creative, pushing me to ask more questions and take more risks in my research. Alain is a jovial, kind-hearted person and his lab has been a welcoming place where I always felt like I could succeed. I am grateful to have him as a mentor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RNA	RiboNucleic Acid
DNA	Deoxyribonucleic Acid
mRNA	Messenger Ribonucleic Acid
poly(A)	Poly Adenine
UTR	Untranslated Region
tRNA	Transfer Ribonucleic Acid
eRF1	Eukaryotic Translation Termination Factor 1
PARS	Parallel Analysis of RNA Structure
FTL	Ferritin Light Chain
IRE	Iron Response Element
IREBP	Iron Response Element Binding Protein
NMR	Nuclear Magnetic Resonance
SHAPE	Selective 2'-Hydroxyl Acylation by Primer Extension
SHAPE-MaP	Selective 2'-Hydroxyl Acylation by Primer Extension & Mutational Profiling
cDNA	Complementary Deoxyribonucleic Acid
MFE	Minimum Free Energy
PCA	Principal Component Analysis
MDS	Multi-dimensional Scaling

CHAPTER 1: INTRODUCTION

Machine learning has become an integral part of our lives. These algorithms help us to do everything from shopping to communicating to working out. Every technological device we own, our computers, our phones, even our watches are involved in forming a representation of our lives in data. In the biomedical sector, these algorithms are revolutionizing healthcare diagnostics. These types of tools are being used to predict spiking blood pressure levels in intensive care patients and to monitor a patient's neurological condition in real-time (Kohn *et al.*, 2014). And with the emergence of high-throughput technology, we can leverage these learning algorithms to glean insight into hidden patterns and complex systems found in large, complex biological data sets as well. Machine learning algorithms are being applied to topics in biology as varied as protein structure classification and estimating bias in microarray data (Qi, 2012). In this work, we describe the development of two computational tools that help analyze the role of RNA structure in human health.

1.1 Messenger RNA

Ribonucleic acid (RNA) is involved in many different functions in a cell, such as regulating gene expression or catalyzing reactions (Lee and Young, 2000). RNA is transcribed or copied from deoxyribonucleic acid (DNA) (Lee and Young, 2000). In eukaryotes, transcription factors, often proteins, along with RNA polymerase recognize a promoter sequence in double stranded DNA upstream from the gene start site (Lee and Young, 2000). Moving from the 3' end

to the 5' end of the DNA template strand, the RNA polymerase unwinds the DNA and transcribes the complementary RNA strand (Lee and Young, 2000). The growth of the RNA strand is called elongation (Lee and Young, 2000).

While an RNA is being transcribed, modifications to the nascent RNA are occurring co-transcriptionally (Shatkin and Manley, 2000). A methylguanylate cap is added to the 5' end when the nascent RNA has grown to about 20 nucleotides in length (Shatkin and Manley, 2000). The 5' cap protects an mRNA from degradation, but the cap is most important in recognition of the mRNA by the ribosome for translation (Shatkin and Manley, 2000). Also occurring co-transcriptionally is splicing of the nascent RNA (Moreno *et al.*, 2015). Splicing removes introns from the messenger RNA encoding for a particular protein (Moreno *et al.*, 2015). By including some exons and excluding others, alternative splicing increases the diversity of mRNAs and proteins that can be created (Moreno *et al.*, 2015). For most RNAs, splicing factors and the spliceosome recognize splice sites at the 5' and 3' ends of an intron, however, some RNAs called ribozymes are able to self-splice (Moreno *et al.*, 2015). There are strong and weak splice sites that coupled with a fast or slow elongation rate can lead to the inclusion or exclusion of certain exons in alternative splicing (Moreno *et al.*, 2015). Weak splice sites less effectively recruit splicing factors and the spliceosome, so typically the inclusion of that exon is decreased if the elongation rate is fast (Dujardin *et al.*, 2014; Moreno *et al.*, 2015). However, there have been observed cases where weak splice sites and fast elongation rates have increased inclusion when other factors are involved (Dujardin *et al.*, 2014; Moreno *et al.*, 2015). mRNA transcription terminates when the nascent RNA is cleaved, releasing the upstream messenger RNA (Dever and Green, 2012). The RNA polymerase may continue transcribing beyond this point (Dever and Green, 2012). An exonuclease digests the remainder of the RNA being transcribed until it

reaches the RNA polymerase and detaches it from the DNA (Dever and Green, 2012). An adenine tail is added to the newly formed precursor mRNA (pre-mRNA) (Shatkin and Manley, 2000). This poly(A) tail is added to the 3' end of the pre-mRNA. Polyadenylation can play a role in nuclear export and stability, but is particularly important in degradation (Shatkin and Manley, 2000). Once the poly(A) tail has been added, the now mature mRNA can be exported from the nucleus to the cytoplasm through nuclear pore complexes (Strambio-De-Castillia *et al.*, 2010).

Once in the cytoplasm, mRNAs are ready for translation into protein. Not all of the mRNA is translated into protein; there are untranslated regions (UTRs) at the 5' and 3' ends of the RNA that control gene expression and mRNA degradation. In eukaryotes, initiation factor eIF-3 recognize the small subunit of the ribosome and GTP-binding initiation factor eIF-2 recognizes the methoionine initiator transfer RNA (tRNA) (Jackson *et al.*, 2010). These two complexes join together (Jackson *et al.*, 2010). After the mRNA is exported to the cytoplasm the 5' cap is recognized by initiation factor eIF-4 (Jackson *et al.*, 2010). The complex containing the small subunit and the tRNA is then guided to the 5' end of the mRNA (Jackson *et al.*, 2010). Alternatively, translation initiation can occur independently of the 5' cap through recognition of a structural motif in the 5' untranslated region of the mRNA (Jackson *et al.*, 2010). The small subunit-tRNA complex then scans along the mRNA from 5' to 3' until it reaches a specific three nucleotide start sequence or codon (Jackson *et al.*, 2010). Once a start codon has been reached, the large subunit binds to the 40S subunit. Initiation factor eIF-5 releases the remaining initation factors (Jackson *et al.*, 2010). The ribosome consists of three sites: E, P and A (Yonath, 2010). The initiator tRNA resides in the middle P-site after the ribosome subunits bind (Dever and Green, 2012). For translation elongation, tRNAs enter the A site of the ribosome (Yonath, 2010). A tRNA enters the ribosome at the A-site (Dever and Green, 2012). A conformational

change occurs in the ribosome and if the anticodon sequence of the tRNA is not complementary to the three mRNA nucleotides at that position, particularly the first two positions, the tRNA will be ejected from the site (Yonath, 2010). Once a complementary tRNA has entered the A-site a peptide bond will be formed between the amino acid and the growing polypeptide chain (Yonath, 2010). The RNA portion of the ribosome catalyzes this reaction (Yonath, 2010). The ribosome then moves the tRNA in the A-site to the P-site, and the P-site tRNA to the E-site through a ratcheting motion (Dever and Green, 2012). The tRNA is then able to exit the ribosome from the E-site (Dever and Green, 2012). The growing polypeptide chain exists the ribosome through the ribosomal tunnel (Yonath, 2010). The tunnel is involved in ensuring that the protein properly folds once it has exited the ribosome (Yonath, 2010). When the stop codon is reached on the mRNA, eukaryotic translation termination factor 1 (eRF1) releases the newly formed protein from the ribosome and the two subunits dissociate (Dever and Green, 2012). The ribosome subunits can then be recycled for use on another mRNA (Dever and Green, 2012). It is also important to note that a single mRNA can have many ribosomes attached at once (Dever and Green, 2012).

RNA performs many functions in a cell and key to all of these functions is its structure. For mRNA, structure can regulate gene expression during translation initiation and elongation. Structure in the coding region can also allow for proper folding of proteins. Variants or polymorphisms in RNA sequence that have been copied from DNA can lead to differences in RNA structure. These differences can ultimately lead to differences in function for RNAs.

1.2 RNA Structure

A single-stranded RNA can fold to adopt specific conformations that are key to the functions it performs in a cell (Weeks, 2010). Over a millisecond timeframe, RNA secondary structure develops from base-pair interactions (Weeks, 2010). During the second and minute timeframe an RNA can form long-range tertiary interactions further increasing an RNAs stability (Weeks, 2010). The structure of RNA is dynamic, which may be important for function, like in the case of riboswitches that form at least two different structures (Lemay *et al.*, 2009; Lemay *et al.*, 2006; Lipert *et al.*, 2007; Tucker and Breaker, 2005). Riboswitches are RNAs found in the 5' untranslated region of an mRNA that regulate gene expression (Lemay *et al.*, 2009; Lemay *et al.*, 2006; Lipert *et al.*, 2007; Tucker and Breaker, 2005). These riboswitches bind a ligand that induces a conformational change that either promotes or inhibits translation of that mRNA (Lemay *et al.*, 2009; Lemay *et al.*, 2006; Lipert *et al.*, 2007; Tucker and Breaker, 2005). Another example of RNA structure playing an important role in function is gene regulation by microRNAs, where double stranded regions are precursors for Dicer recognition to further process the microRNAs for use in the RISC complex (Mortimer *et al.*, 2014; Wilson and Doudna, 2013).

RNA structure also plays a role in translational control in mRNAs (Mortimer *et al.*, 2014). Structuredness in the 5' untranslated region of an mRNA, particularly around the translational start codon can reduce translational efficiency (Ingola *et al.*, 2009; Mortimer *et al.*, 2014; Shabalina *et al.*, 2006). RNA structure particularly in the 5' untranslated region can also increase the stability of an mRNA under stress conditions, such as heat shock (Wan *et al.*, 2012). RNA structure in the coding region of an RNA may induce pausing (Meyer, 2005; Mortimer *et al.*, 2014; Wolin and Walter, 1988). This allows for proteins that fold co-translationally to form

intermediates required for proper folding (Komar, 2009; Mortimer *et al.*, 2014). Structuredness in the coding region of an RNA has also been linked to increased translational efficiency (Kertesz *et al.*, 2010; Li *et al.*, 2012; Mortimer *et al.*, 2014). Increased structuredness in the 3' untranslated region of mRNAs can help localize them to the correct location in the cell (Kertesz *et al.*, 2010; Mortimer *et al.*, 2014). Structured 3' untranslated regions of an mRNA can lead to longer half-lives, with less flexible 3' ends being less targeted by the exosome complex for degradation (Wan *et al.*, 2012).

It is important to note that while some ribonucleic acid (RNA) molecules have evolved to adopt a single conformation, a majority of RNAs are thought to adopt multiple conformations (Matthews, 2006). RNA molecules may exist in multiple conformations in order to perform a function, such as promoting or inhibiting expression of an associated gene (Lemay *et al.*, 2009; Lemay *et al.*, 2006; Lipert *et al.*, 2007; Tucker and Breaker, 2005). In order to perform these functions there must be several energetically similar and easily accessible structures that RNA molecules can form (Matthews, 2006). The possible structures that RNA may take in a cell constitute its structural ensemble (Matthews, 2006).

Structural changes in RNA may lead to a functional consequence, a phenomena referred to as a riboSNitch. RiboSNitches can be created through polymorphisms and variation in DNA that is transcribed into the RNA. A recent study compared RNA structure in a human family trio (mother, father and child) on a genome wide scale using parallel analysis of RNA structure (PARS) (Wan *et al.*, 2014). This study identified riboSNitches in 15% of over 12,000 single nucleotide variants, and 22 unique riboSNitches associated with human phenotypes and diseases, including multiple sclerosis and asthma (Wan *et al.*, 2014). These findings indicate that riboSNitches may play a role in gene regulation and disease (Wan *et al.*, 2014). The ferritin light

chain (FTL) is a protein subunit that is associated with Hyperferritienemia Cataract Syndrome, a genetic disorder that causes cataracts in infancy (Ritz *et al.*, 2012). A mutation in the 5' untranslated region of the mRNA for FTL contains an iron response element (IRE) that binds the iron response element binding protein (IREBP) (Ritz *et al.*, 2012). A mutation in the 5' untranslated region that does not directly alter the sequence of the IRE, changes the structure of the 5' untranslated region preventing the IRE from binding to the IREBP (Ritz *et al.*, 2012).

1.3 RNA Structure Probing

There are several methods for determining RNA structure. With X-ray crystallography and Nuclear Magnetic Resonance (NMR) secondary and tertiary structure can be determined, however, for many longer RNAs these methods cannot be used (Kubota *et al.*, 2015). An alternative set of methods for determining RNA structure is chemical probing (Ehresmann *et al.*, 1987; Peattie and Gilbert, 1980). For these methods a chemical probe reagent interacts with a portion of the RNA and that interaction can be measured to give information on RNA structure (Weeks, 2010). There are base-specific reagents that form adducts with one or more RNA bases that provide information on base stacking, hydrogen bonding and the electrostatic environment adjacent to the modified base (Tijerina *et al.*, 2007; Weeks, 2010). Alternatively, hydroxyl radicals can be generated to cleave RNA backbone giving information on solvent accessibility of the backbone (Tullius and Greenbaum, 2005; Weeks, 2010). Using reagents that are tethered to one section of an RNA that can react with distant regions, long-range tertiary interactions can be measured (Sigurdsson *et al.*, 1995; Weeks, 2010). Some methods interact with specific functional groups on the RNA background and measure local nucleotide flexibility and dynamics (Regulski and Breaker, 2008; Weeks, 2010).

One such method that interacts with the 2'-hydroxyl of the RNA backbone is selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) (Wilkinson *et al.*, 2006). SHAPE reagents react with the 2'hydroxyl group to form an adduct (Wilkinson *et al.*, 2006). These reagents preferentially form adducts in more flexible regions (more likely unpaired), because these regions are more likely to adopt conformations that are favorable to reaction with the SHAPE reagent (Wilkinson *et al.*, 2006). Radiolabeled complementary DNA is annealed to the modified RNA by reverse transcriptase (Wilkinson *et al.*, 2006). Once an adduct is reached, the reverse transcriptase stops, leaving a complementary DNA fragment (Wilkinson *et al.*, 2006). These fragments can then be size separated by gel electrophoresis or capillary electrophoresis (Mitra *et al.*, 2008; Wilkinson *et al.*, 2006). Positions with high modification, indicated by 3' fragment ends, are more flexible positions (Wilkinson *et al.*, 2006). Advancements to the SHAPE method have utilized high-throughput sequencing technology to measure hundreds of RNA sequences for several experiments at once (Lucks *et al.*, 2011). Reagents have also been developed to allow for *in vivo* SHAPE analysis (Spitale *et al.*, 2013).

SHAPE experiments can also be performed genome-wide using SHAPE and mutational profiling (SHAPE-MaP) (Figure 1.1) (Siegfred *et al.*, 2014). In this method, an adduct induces a mutation that the reverse transcriptase can read through creating differences in the complementary DNA sequence at more flexible regions (Siegfred *et al.*, 2014). An untreated sample (background mutation rate at each nucleotide) is subtracted from the modified sample (experimental mutation rate at each nucleotide) and normalized by the denatured control (all nucleotides unpaired) (Siegfred *et al.*, 2014). The untreated sample, modified sample and

denatured control can all be run in parallel using high-throughput sequencing technology (Siegfred *et al.*, 2014). This method can be applied on a genome wide-scale by using random primers to synthesize the complementary DNA at any position (Siegfred *et al.*, 2014).

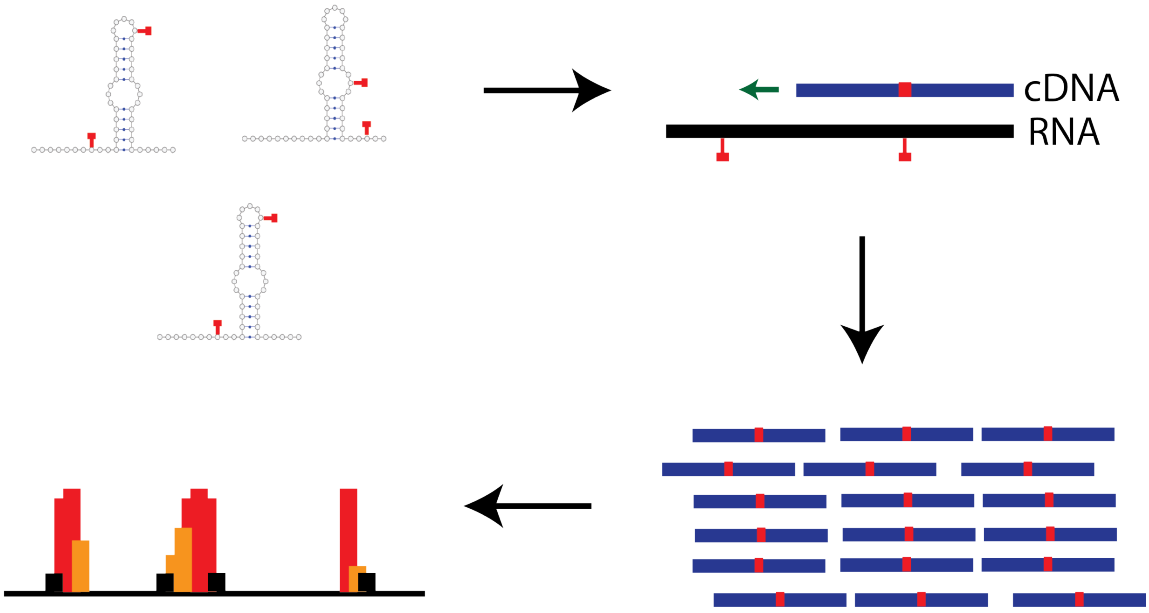


Figure 1.1. SHAPE-MaP Methodology.

RNAs are modified by chemical reagents that preferentially form an adduct at the backbone 2'-hydroxyl for flexible positions (Siegfred et al., 2014). Reverse transcriptase produces complementary DNA (cDNA) strands inducing a mutation at the location of a mutation (Siegfred et al., 2014). High-throughput sequencing technology aligns the cDNA to the reference sequence (Siegfred et al., 2014). A higher mutational rate at a position indicates a nucleotide that is more flexible or more likely unpaired (Siegfred et al., 2014).

Many techniques have been developed to experimentally determine RNA structure. However, structures for every RNA are not always readily available or easily obtained. In these cases it may be useful to rely on computational methods for the prediction of RNA structure.

1.4 RNA Structure Prediction

For RNAs where an experimentally determined structure is unavailable, structure prediction may be a valuable tool in determining RNA secondary structure. RNA secondary structure can be predicted without knowing the tertiary structure, because secondary interactions are typically stronger and occur faster than tertiary interactions (Banerjee *et al.*, 1993; Matthews *et al.*, 1997; Woodson, 2000).

One of the first methods used for predicting RNA secondary and tertiary structure, comparative sequence analysis, compared multiple homologous sequences between organisms with shared ancestry (Gutell *et al.*, 2002; Michel *et al.*, 2000). Base-pairs were inferred by determining canonical pairs that are common among the sequences (Gutell *et al.*, 2002; Michel *et al.*, 2000). Compensating base-pair changes further provided support for a base-pairing (Gutell *et al.*, 2002; Michel *et al.*, 2000). Comparative sequence analysis has been shown to be most successful when many homologous sequences are available (Gutell *et al.*, 2002).

Among the most popular algorithms for determining RNA secondary structure is free energy minimization (Hofacker, 2003; Hofacker *et al.*, 1994; Zuker, 2003). For an RNA, at equilibrium there is an equilibrium between strands folded in structure S_1 and unstructured strands S_{rc} with an equilibrium constant of K_1 (Eq. 1 and Eq. 2) (Matthews, 2006). The stability of structure S_1 is determined by the free energy change ΔG^o_{37}, where R is the gas constant and T is the absolute temperature (Eq. 3) (Matthews, 2006). The relationship between the stabilities for

two structures S_1 and S_2 is given by the ratio of these equilibrium constants. From these equations, the lowest free energy structure is the most common conformation at equilibrium (Matthews, 2006).

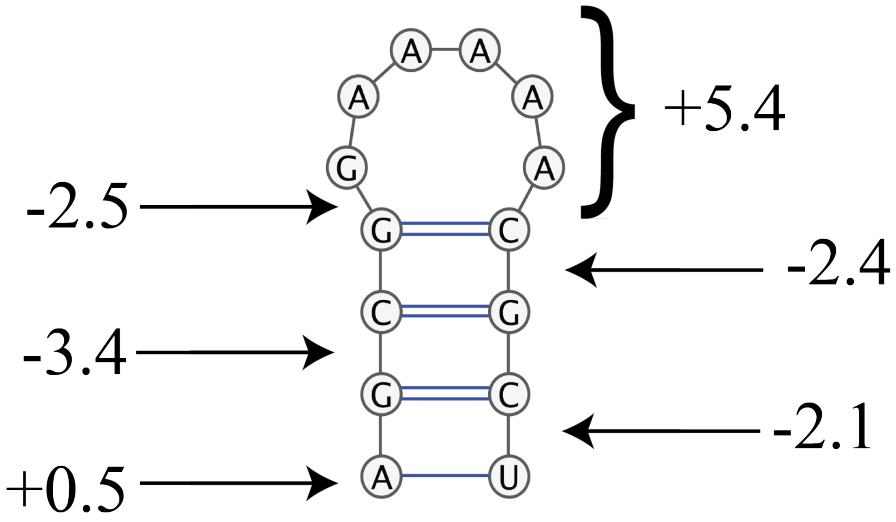
$$S_{rc} \rightleftharpoons S_1 \tag{1}$$

$$K_1 = \frac{[S_1]}{[S_{rc}]} \tag{2}$$

$$K_1 = \mathrm{e}^{-\Delta G^o_{37}(1)/R\mathrm{T}} \tag{3}$$

$$\frac{[K_1]}{[K_2]} = \frac{[S_1]}{[S_2]} = \mathrm{e}^{(\Delta G^o_{37}(2) - \Delta G^o_{37}(1)/RT} \tag{4}$$

The most common method for predicting the folding free energy of a secondary structure is an empirical nearest-neighbor model (Matthews *et al.*, 1999; Matthews *et al.*, 2004; Xia *et al.*, 1998). For a nearest-neighbor model, the free energy change is determined by the sequence and the most adjacent base pairs (Figure 1.2) (Matthews, 2006). The entropic contributions for the free energy change include loops and bulges, while the enthalpic contributions include base pairs and stacking (Matthews *et al.*, 1999; Matthews *et al.*, 2004; Xia *et al.*, 1998). The parameters for each of these contributions have been determined experimentally (Matthews *et al.*, 1999; Matthews *et al.*, 2004; Xia *et al.*, 1998).



$$\Delta G^{o}_{37} = 0.5 - 2.1 - 3.4 - 2.4 - 2.5 + 5.4 = -4.5 \text{ kcal/mol}$$

Figure 1.2. Nearest-neighbor model for a stem loop.

The free energy change calculation for a model RNA stem loop is depicted (Matthews, 2006). A penalty is given for a pair terminating a helix (Matthews *et al.*, 1999; Matthews *et al.*, 2004; Xia *et al.*, 1998). Stacking interactions are included as an additional favorable increment (Xia *et al.*, 1998). An entropic penalty is included for constraining nucleotides in a loop (Matthews *et al.*, 2004).

The number of secondary structures N_{ss} grows exponentially with length N (Eq. 5) (Giegerich *et al.*, 2004; Matthews, 2006). To find the lowest free energy structure, dynamic programming can be implemented in order to avoid generating all possible structures (Figure 1.3) (Nussinov and Jacobson, 1980; Nussinov *et al.*, 1978). In the fill step, the lowest free energy is determined for each sequence fragment starting with the smallest fragment and then increasingly longer fragments (Eddy, 2004; Nussinov and Jacobson, 1980; Nussinov *et al.*, 1978). The additive nature of the free energy calculation allows for the longer fragments to be calculated recursively (Eddy, 2004; Nussinov and Jacobson, 1980; Nussinov *et al.*, 1978). The longest fragment is the entire sequence, so once the fill step is complete the minimum free energy for the sequence is now known (Matthews, 2006). The traceback step starts at the minimum free energy and traces backward through the fragments generated in the fill step to determine the interactions that contributed to the minimum free energy are determined (Eddy, 2004; Nussinov and Jacobson, 1980; Nussinov *et al.*, 1978). This method guarantees the lowest free energy structure is found (Matthews, 2006).

$$N_{ss} \approx 1.8^{N} \tag{5}$$

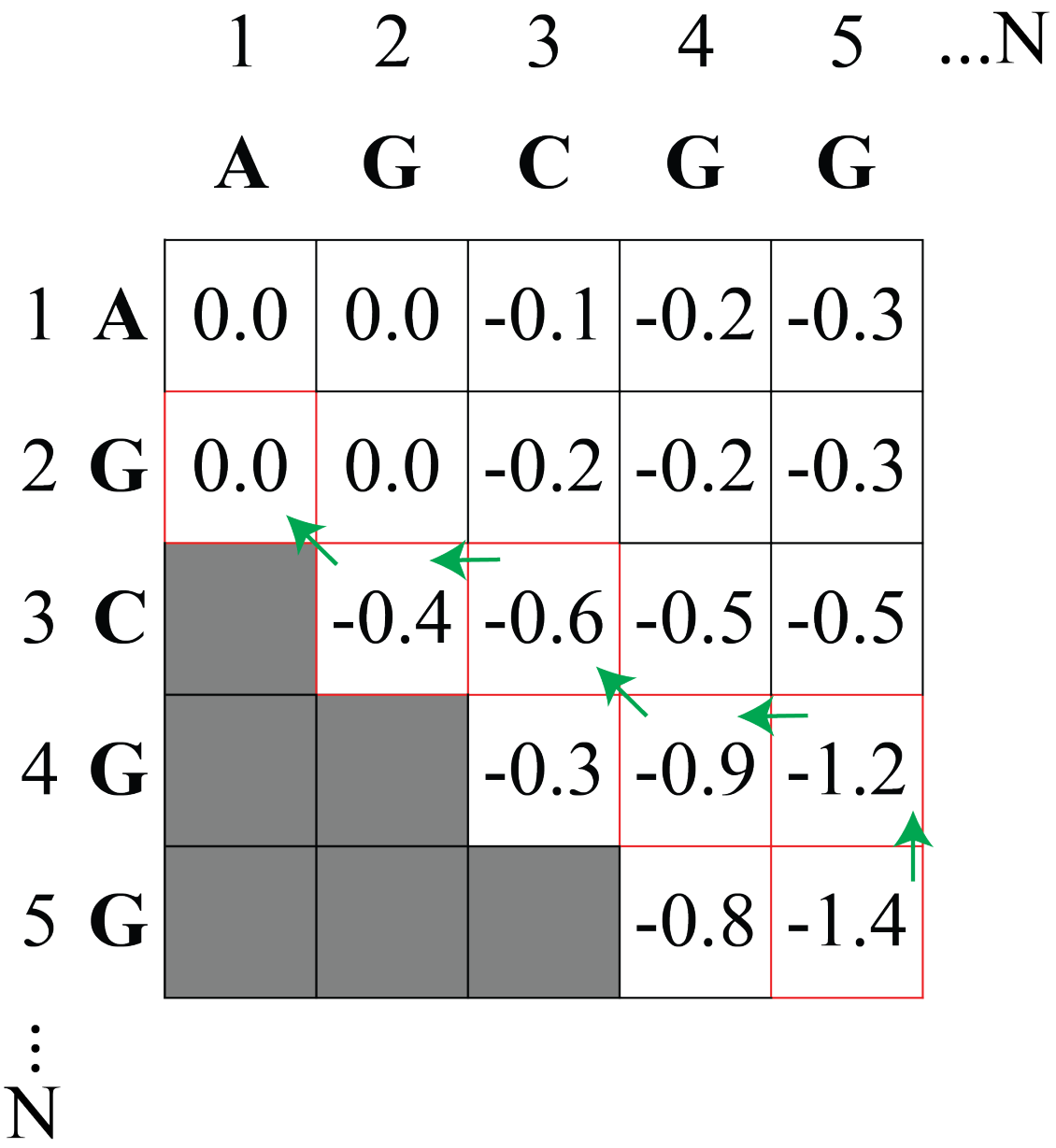


Figure 1.3. Dynamic programming methodology.

The fill and traceback steps for dynamic programming are depicted for a model RNA fragment. The matrix shows the fill step that calculates the lowest free energy. The green arrows show the traceback step that calculates which fragments are included in the minimum free energy structure. The traceback step starts at the minimum free energy.

RNA structure prediction can be improved with the inclusion of structure probing data (Diegan *et al.*, 2008). An extra free energy term can be added to the nearest neighbor free energy model in order to account for the empirical information (Eq. 6) (Diegan *et al.*, 2008). The intercept b represents a reward for pairing nucleotides with a low SHAPE reactivity, and the slope m represents a penalty for pairing nucleotides with high SHAPE reactivity(Diegan *et al.*, 2008). Parameters b and m were determined empirically using the prediction 23S ribosomal RNA as a model RNA, where the secondary structure is known from comparative sequence analysis (Diegan *et al.*, 2008).

$$\Delta G_{shape}(i) = m \ln[SHAPE(i) + 1] + b \quad (6)$$

RNA structure prediction can be useful for understanding how RNA structure relates to function, particularly when a crystal structure is unavailable. While the free energy minimization using dynamic programming provides an efficient method to calculate a representative structure for an RNA, this prediction can be improved with the addition of chemical mapping data. Sampling from the entire ensemble of structures that an RNA may take in a cell, may further improve RNA secondary structure prediction, allowing for a more accurate representation of RNA structure.

1.5 Boltzmann Suboptimal Sampling

Reliance on only the minimum free energy structure is problematic (Matthews, 2006). Free energy nearest neighbor models are incomplete, and many interactions are non-nearest neighbor (Chen *et al.*, 2004; Kierzek *et al.*, 1999; Longfellow *et al.*, 1990; Matthews, 2006;

Schroeder *et al.*, 1999). Some interactions cannot be modeled using dynamic programming (Matthews, 2006; Matthews and Turner, 2002). The assumption that the RNA is at equilibrium does not account for how folding kinetics may play a role in determining secondary structure (Heilman-Miller and Woodson, 2003; Matthews, 2006). Most importantly many RNAs, like riboswitches, have evolved to form multiple conformations, which can be important to their function (Lemay *et al.*, 2009; Lemay *et al.*, 2006; Lipert *et al.*, 2007; Martin *et al.*, 2012; Schultes and Bartel, 2000; Tucker and Breaker, 2005). All of the conformations a single RNA may sample in a cell are its structural ensemble (Matthews, 2006). For these reasons, calculation of a set of low free-energy suboptimal structures provides more information on RNA structure than just the minimum free energy structure.

The first algorithms that allowed for the prediction of suboptimal secondary structures allowed for an arbitrary starting point for the traceback step in finding the minimum free energy structure (Steger *et al.*, 1984; Zuker, 1989). The traceback step then created a suboptimal secondary structure (Steger *et al.*, 1984; Zuker, 1989). This method was efficient, but not all possible secondary structures could be explored (Matthews, 2006). Subsequent algorithms determined all suboptimal structures within an energy range from the minimum free energy (Williams and Tinoco, 1986; Wuchty *et al.*, 1999). All secondary structures are calculated without redundancy in the fill step from minimum free energy calculation (Wuchty *et al.*, 1999). The trace back step then determines all structures within a specified range of the lowest free energy structure (Wuchty *et al.*, 1999).

Current algorithms sample suboptimal secondary structures from the Boltzmann ensemble of structures (Ding and Lawrence, 2003; Ding and Lawrence, 1999). The fill step uses the partition functions devised for minimum free energy calculation (Ding and Lawrence, 1999;

McCaskill, 1990). The traceback step generates base-pairs according to the partition functions for all possible sequence fragments (Ding and Lawrence, 2003; Ding and Lawrence, 1999). This creates a set of suboptimal structures that is a statistical sample of the RNA structural ensemble (Ding *et al.*, 2005). The statistical sample is remarkably stable; a messenger RNA with over 1000 nucleotides a sample size of 1000 structures is sufficient to produce nearly the same base pairing probabilities for each run (Ding *et al.*, 2006).

With the ability to create a statistical ensemble of RNA structures, we can more accurately identify structural elements that are playing a role in an RNAs function (Ding *et al.*, 2006; Ritz *et al.*, 2012). We can also better determine how variants contribute to differences in RNA structure and how that potentially leads to differences in phenotype.

With the advent of new technologies, information on RNA structure can be gathered on a genome-wide scale. However, this large amount of data can be difficult to analyze. This difficulty particularly exists in identifying structure change in RNAs caused by variants or polymorphisms. Algorithms in the machine learning field have been developed to address the problem of finding patterns in large data sets, and can be utilized to address this problem.

1.6 Machine learning in classification

Machine learning is the exploration and development of algorithms that can learn from and make predictions on existing data (Hua *et al.*, 2009; Libbrecht and Noble, 2015). The field of machine learning has led to the development of algorithms that can analyze complex data sets and improve performance based on new information (Libbrecht and Noble, 2015). Many applications in a wide variety of areas, including marketing, finance and telecommunications, utilize machine learning (Hua *et al.*, 2009; Libbrecht and Noble, 2015).

There are three subgroups of machine learning techniques: unsupervised learning, supervised learning, and semi-supervised learning (Hua *et al.*, 2009; Libbrecht and Noble, 2015). Unsupervised learning finds hidden structure in unlabeled data (Raychaudhuri *et al.*, 2009). Examples of unsupervised learning algorithms include clustering and principal component analysis (Ding *et al.*, 2006; Raychaudhuri *et al.*, 2009). The semi-supervised learning uses an incomplete training set for learning, because the use of even small amounts of labeled data improves the accuracy of learning (Libbrecht and Noble, 2015).

Supervised learning infers function from labeled data (Libbrecht and Noble, 2015). This subgroup includes two categories: classification, which predicts categories and regression that allows the prediction of values (Liaw and Weiner, 2002; Libbrecht and Noble, 2015). Both of these infer function from a set of known examples, and predict the function for future samples (Liaw and Weiner, 2002).

Currently, many supervised learning techniques exist that can be utilized for classification of the change in RNA structure. One classifier that can classify RNA structure change is random forest. Such a classifier would be built on a set of features that characterize RNA structure change in chemical mapping data. A set of labeled samples would be required for use in random forest supervised learning.

1.7 Random Forest and Classification of RNA Structure Change

Random Forest is a supervised learning technique that can be used for classification, regression or anomaly detection (Breiman, 2001). This technique has been widely used in bioinformatics including for the analysis of microarray data, drug screening, and genome-wide association studies (Chen and Ishwaran, 2012; Riddick *et al.*, 2011; Wu *et al.*, 2003; Yang *et al.*,

2010). Inputs for random forest consist of a set of N samples characterized by a set of M features. Each of these samples has a label or value (Figure 1.4A) (Breiman, 2001; Liaw and Weiner, 2002).

The random forest technique forms decision trees, which group the samples into different nodes according to the feature being measuring (Figure 1.4B) (Breiman, 2001; Chen *et al.*, 2011). The root node is the feature that includes all of the samples (Liaw and Weiner, 2002). The features of the samples using random forest are selected at random to split the node (Liaw and Weiner, 2002). If multiple features are selected, a linear combination of the features will be used (Liaw and Weiner, 2002). The tree grows by choosing the best split based on the selected features and breaking the samples into two new nodes (Liaw and Weiner, 2002). A node that can no longer be split, is called a leaf node, because all samples are identical or the node only contains a single sample (Liaw and Weiner, 2002). Random forest uses a set or forest of decision trees. Each tree samples with replacement from the input and then grown to the fullest extent (Breiman, 2001; Liaw and Weiner, 2002). This bootstrap sampling ensures that some samples are always left out of each tree. The most common label for samples in a leaf node determines the class for everything found in that node (Breiman, 2001; Liaw and Weiner, 2002). When a decision tree assigns a class to a sample, the decision is casting a vote for that sample. Across all trees, the most common class vote for a sample determines the final classification (Breiman, 2001; Liaw and Weiner, 2002). Each tree can then predict the classification for out of bag samples or those that were not included by the bootstrap sampling. The classification error for the out of bag samples gives the generalization error (Breiman, 2001; Liaw and Weiner, 2002). The classification for new samples can be determined from predictions with an existing forest of trees.

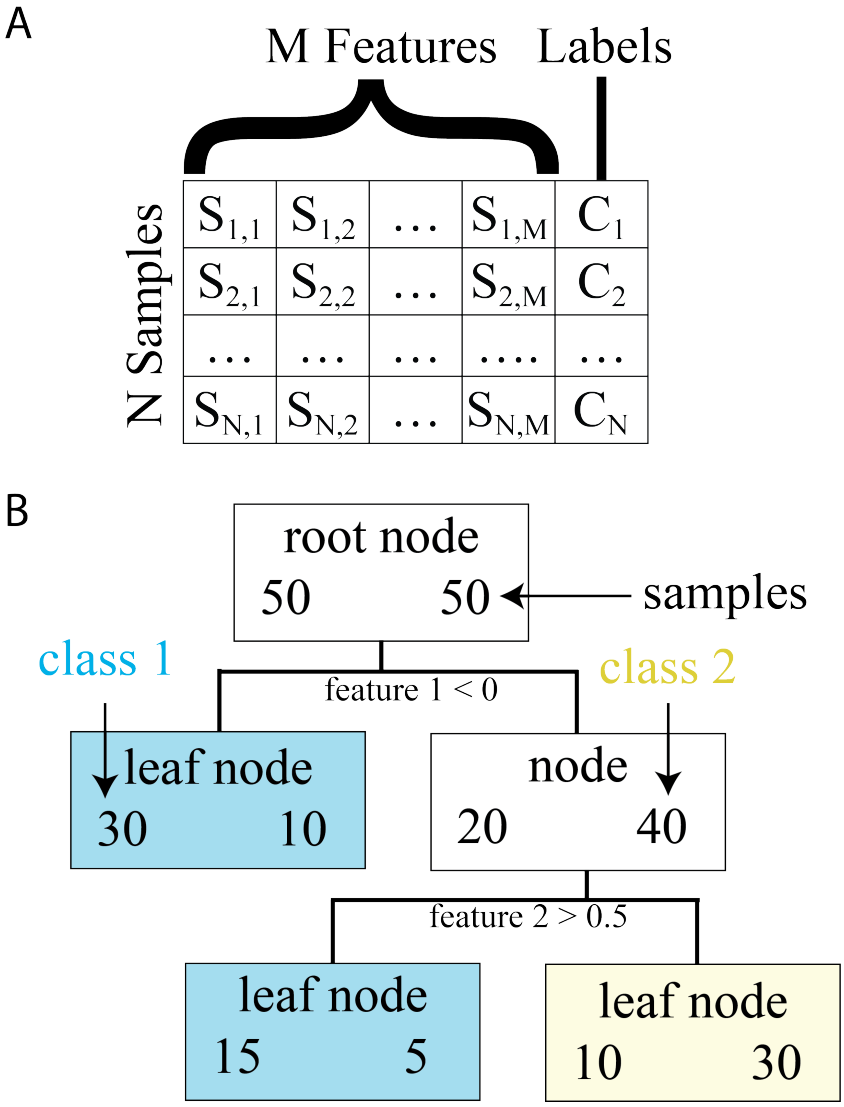


Figure 1.4 A random forest model decision tree

A) This matrix is an example of a random forest input with *N* samples, *M* features and *N* class labels (Qi, 2012). B) This diagram shows an example decision tree with 100 samples and 2 classes. Each node is split on a single feature. The most common class label in a leaf node determines the classification for every sample in the node.

The importance of individual features in building a random forest classifier can be determined using the Gini importance or the permutation importance (Breiman, 2001). Gini importance calculates the average node purity or how mixed the labels are for each node (Breiman, 2001). This measure is sensitive to categorical data features with more variables (Breiman, 2001). The permutation importance determines how much the prediction accuracy decreases when a variable is removed, which is sensitive to differences in scale between features (Breiman, 2001). The fraction of trees where two elements are located in the same leaf node gives a proximity measure (Breiman, 2001). The proximity measure can be used to create a similarity matrix.

A random forest classifier produces the best error rate when the correlation between trees is low and the strength or accuracy of each tree is high (Breiman, 2001). The performance of a classifier can be improved by reducing the number of features selected to split each node (Breiman, 2001). A large number of trees is required for stable estimates (Breiman, 2001). For an unbalanced population, where one class has many more samples, the class with more samples can be under-sampled to produce better error rates (Chen *et al.*, 2004). An alternative is increasing the percentage of tree votes required to determine the class with more samples (Chen *et al.*, 2004).

Random forest is widely used for several reasons; (1) the classifier is efficient on large data sets, because it is easily parallelized, (2) the generalized error rate is unbiased so there is no need for cross-validation, and (3) the method is non-parametric, so there is no assumption about the underlying population distribution (Breiman, 2001; Touw *et al.*, 2013). Random forest performs well with many features and few cases or with few features and many cases for classification (Breiman, 2001; Touw *et al.*, 2013). Despite the many advantages of random forest

there are several drawbacks. Individual trees are not useful, making the interpretation of the forest more difficult (Breiman, 2001; Touw *et al.*, 2013). Correlated features are problematic and especially for determining feature importance (Breiman, 2001; Touw *et al.*, 2013). The generalization error has an upper bound, but it is still possible that the error rate for a training set is much better than the error rate for a test set (Breiman, 2001; Touw *et al.*, 2013).

Random forest can be a useful tool in identifying structure change in chemical mapping data. Once differences in RNA structure have been identified, it is important to determine which structural elements are changing and how these relate to function. Secondary and tertiary structural information for an RNA can be experimentally determined by techniques such as x-ray crystallography (Holbrook and Kim, 1997). However, structures resolved using such techniques are not always available. Our goal is to use computational prediction as a valuable alternative to determine how changes in RNA affects their structure.

1.8 RNA Structure Visualization

One way that we can utilize algorithms that statistically sample an RNA structural ensemble is to ompare the predicted ensembles using data visualization techniques. Data visualization is used for two purposes: data analysis and communication (Few). Visualization can be a powerful tool in identifying important structural elements in an RNA and comparing how these elements differ between variants (Ding *et al.*, 2005; Ritz *et al.*, 2012). RNA structure visualization also enables scientists to effectively communicate how differences in RNA structure may affect phenotype (Ritz *et al.*, 2012).

Typically RNA structure is represented as a single best representative, either the minimum free energy structure (MFE) or the ensemble centroid structure (Ding *et al.*, 2005;

Zuker and Stiegler, 1981). These single structure representatives have been shown to inadequately describe RNA molecules, as they exist in an ensemble (Ding *et al.*, 2006; Matthews, 2006). The MFE representation assumes that at equilibrium an RNA molecule folds into a unique lowest energy state (Ding *et al.*, 2005; Matthews *et al.*, 1999; Zuker and Stiegler, 1981). However, the MFE structure is not always the most common structure in an ensemble (Ding *et al.*, 2005; Ding *et al.*, 2006). The ensemble centroid structure is the structure with the minimum base-pair distance to all other structures in the ensemble(Ding *et al.*, 2005; Ding *et al.*, 2006). This representation is a more accurate representation of an RNA structural ensemble, but does not account for different clusters of structures (Ding *et al.*, 2005; Ding *et al.*, 2006). In some cases a single structure representation is not sufficient to describe an ensemble of RNA structures (Figure 1.5A).

An alternative to single structure representations is multi-dimensional scaling (MDS) (Ding *et al.*, 2005; Ritz *et al.*, 2012; Torgerson, 1952). Classical MDS calculates the Euclidean distance matrix for n-dimensional data, d_{ij} (Abdi, 2007; Ding *et al.*, 2005). Eigen decomposition is performed on the cross product matrix transformed from the distance matrix (Abdi, 2007). Projection of the data onto the first two or three eigenvectors, those with the highest eigenvalues, allows for the visualization of the RNA ensemble in two or three-dimensional space (Abdi, 2007). Metric MDS optimizes the data points to recapitulate the distance matrix (Abdi, 2007; Torgerson, 1952). This algorithm sets the initial positions for the data points, x, in 2-dimensional space, i and j, using classical MDS. From this configuration, metric MDS evaluates the stress function in Eq. 7 (Abdi, 2007; Torgerson, 1952). The data points are reconfigured in the direction of steepest descent. This process is repeated to minimize the stress function (Abdi, 2007; Torgerson, 1952). Minimizing the stress function, attempts to find the configuration with

the smallest residual sum of squares when compared to the original distance matrix (Abdi, 2007; Torgerson, 1952). Non-metric MDS also optimizes the data point configuration but instead preserves the order of the data points not their distances (Abdi, 2007; Torgerson, 1952). This algorithm is similar to metric MDS except an additional step before optimization occurs where isotonic regression is used to relate the distance matrix to the lower dimensional projection (Abdi, 2007; Torgerson, 1952). The positioning between the structures is well maintained in all three MDS algorithms. However, new structures cannot be projected onto the space without recalculating the eigenvectors, which could greatly change the visualization (Abdi, 2007). The lack of a consistent space for projection of structures would make the addition of new structures or the comparison between structural ensembles difficult (Figure 1.5B).

$$Stress = \sqrt{\sum_{ij} \frac{(d_{ij} - \|x_i - x_j\|)^2}{\sum d_{ij}^2}} \quad (7)$$

Another alternative to single structure representation that does create a consistent space for projection of structures is principal component analysis (PCA) (Pearson, 1901). Currently, PCA has been performed on the binary structural information for an RNA ensemble, a base is given a value of paired or unpaired (Ritz *et al.*, 2012). This binary representation for RNA is inappropriate for use in principal components, which is designed for continuous data (Abdi and Williams, 2010). Also, different structures may result in the same binary representation because secondary structure information about the base pairs, which bases are paired together, is lost (Figure 1.5C).

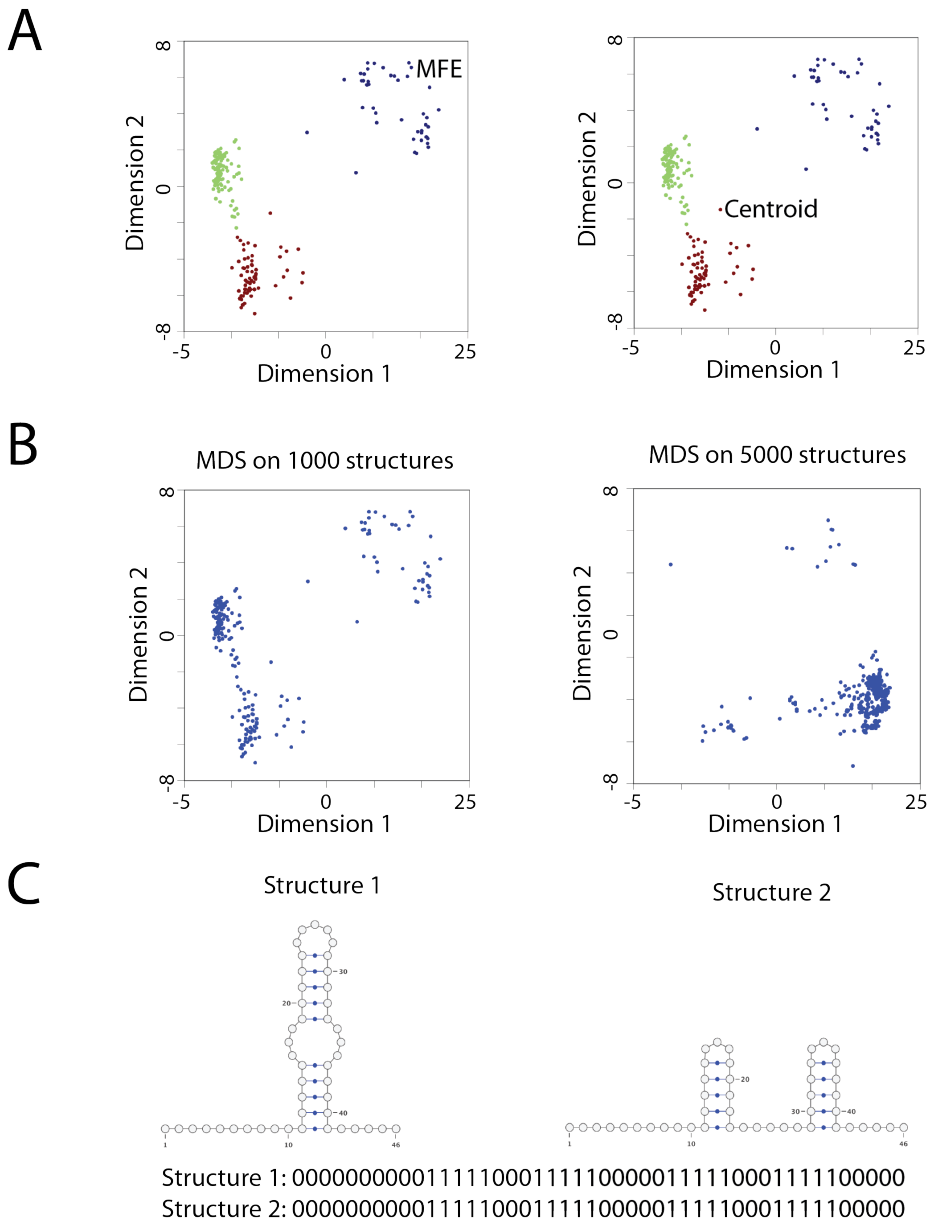


Figure 1.5 Current methods for RNA structure visualization.

A) Multidimensional scaling for a model RNA ensemble. Each point represents a structure. The MFE structure is not in the densest cluster and the centroid structure does not represent any cluster. B) Multidimensional scaling for the same model RNA ensemble with 1000 structures and with 5000 structures C) RNA structure models that have the same binary representation.

A useful tool for the visualization of RNA structural ensembles would include the following: (1) the grouping of similar structures into larger clusters based on relevant structural information, (2) a visual representation of the relationship between different clusters, (3) the frequency that a particular structure or cluster of structures is present, and (4) the ability to compare RNA ensembles. Current tools available for the visualization of RNA structure do not include one or more of these important features. A useful method for visualizing and comparing RNA structure remains a challenge.

CHAPTER 2: CLASSIFICATION OF RNA STRUCTURE[1]

2.1 Introduction

A persistent challenge in the field of structural biology is accurately predicting the conformational and ultimately functional consequences of a mutation on a protein or nucleic acid (Chauhan and Woodson, 2008; Cheng *et al.*, 2005; Churkin *et al.*, 2011; Russell *et al.*, 2002). For both nucleic acids and proteins, accurately predicting the extent of disruption is generally more challenging than predicting the entire structure (Miao *et al.*, 2015; Waldispuhl and Reinharz, 2015; Wan *et al.*, 2014). Indeed it requires making two, albeit related structure predictions. The data most often used in conjunction with RiboNucleic Acid (RNA) structure prediction algorithms are chemical and enzymatic probing experiments (Corley *et al.*, 2015; Ritz *et al.*, 2012; Solem *et al.*, 2015). These experiments, in particular Selective 2' Hydroxyl Acylation by Primer Extension (SHAPE) provide nucleotide resolution structural information and are exquisitely sensitive to structure change (Cruz *et al.*, 2012; Kutchko *et al.*, 2015; Rice *et al.*, 2014; Siegfried *et al.*, 2014). Recent technological advances enable this data to be collected with unprecedented throughput (Siegfried *et al.*, 2014); traditionally this data was carefully

[1] This chapter previously appeared as an article in Bioinformatics
doi: 10.1093/bioinformatics/btx041.

human curated to ensure accuracy, which is simply not possible in the genomic context (Ritz *et al.*, 2012; Rocca-Serra *et al.*, 2011; Sansone *et al.*, 2012).

Chemical and enzymatic probing techniques have long been used in structural, kinetic and thermodynamic characterizations of nucleic acids (Brenowitz *et al.*, 1986; Brenowitz *et al.*, 1986; Deras *et al.*, 2000; Sclavi *et al.*, 1997). Until the advent of capillary sequencing and more recently next generation sequencing, the experiments were carried out using traditional gel electrophoresis (Brenowitz *et al.*, 1986; Brenowitz *et al.*, 1986; Petri and Brenowitz, 1997). Although informatics tools were developed to rapidly quantify these complex electropherograms, most structural insight was still gleaned by "gel gazing;" for an effect to be robust the scientist had to be able to visualize it (Das *et al.*, 2008; Das *et al.*, 2005; Russell *et al.*, 2002; Takamoto *et al.*, 2004). With high-throughput probing experiments rapidly becoming the norm, it is impossible to systematically visualize all the data.

In this manuscript we are specifically interested in mutation induced structure change in RNA and in particular the detection of riboSNitches using chemical and enzymatic probing data (Corley *et al.*, 2015; Halvorsen *et al.*, 2010; Lokody, 2014; Martin *et al.*, 2012; Ritz *et al.*, 2012; Solem *et al.*, 2015; Wan *et al.*, 2014). Accurately detecting riboSNitches experimentally is essential to establishing robust benchmarks (Corley *et al.*, 2015; Ritz *et al.*, 2012). Moreover, as transcriptome-wide structure probing experiments rapidly become the norm (Martin *et al.*, 2012; Wan *et al.*, 2012; Wan *et al.*, 2014), efficiently detecting riboSNitches is likely to become an important component of personalized medicine (Solem *et al.*, 2015). The main premise for the work presented in this manuscript is in the history of chemical and enzymatic probing techniques and in particular the value of expert human decision making in the determination of whether a

structural change is significant. In particular, the distinction between a local structural change affecting several residues and a global structure change affecting a majority of residues.

Human ability to visually detect patterns in data is exceptional; even in the field of RNA structure, humans readily design better RNA folds than purely automated programs (Lee *et al.*, 2014; Rowles, 2013; Treuille and Das, 2014). Interestingly, with enough examples machines can then learn the rules used by humans to make these designs (Lee *et al.*, 2014). In this manuscript, we aim to automate some of the human skills associated with "gel gazing" and apply these to the problem of identifying riboSNitches from high-throughput SHAPE data. We are particularly interested in understanding how humans interpret SHAPE data and what features of the signal they use to classify structure change. We are also interested in determining whether there is a consensus among users of SHAPE data as to what constitutes a small or large change in RNA structure. We therefore created a platform for easily visualizing SHAPE traces and asked experts in the field to classify traces and structures. As will be shown below, there is surprising agreement in human appreciation of the data and from these classifications we are able to identify novel metrics that reproduce the manual classifications. We are therefore able to report a structural classification scheme that quantitatively reproduces the process of "gel gazing." Our classifier allows us to simulate human eyes on high-throughput data sets and identify important differences in specific RNAs' sensitivity to mutation.

2.2 Methods

2.2.1 Data Set

SHAPE traces for 17 mutate-and-map experiments were obtained from the publicly available RNA Mapping DataBase (RMDB) (Cordero *et al.*, 2012; Kladwang *et al.*, 2011; Kladwang *et*

al., 2011). These 17 RNA database entries had a total of 2019 WT and single-point mutant trace pairs (Table 2.3). Of these trace pairs, 200 pairs were chosen for manual evaluation by 14 experts. Due to incomplete survey results we were able to obtain a majority consensus from at least 14 experts on 167 of the pairs.

2.2.2 Data normalization and noise reduction

Each WT trace was normalized to a mean reactivity of 1.5. A multiplier was used to normalize the respective mutant trace. The multiplier was chosen that minimized the difference between the WT and mutant traces. We reduced noise by setting mutant SHAPE values equal to the WT value, if both reactivities were outliers as defined by (Karabiber *et al.*, 2013). To remove end effects, 8% of the data was trimmed from the 5' and 3' ends. Normalization and noise reduction are further explained in Methods Supplementary, Subsection 2.5.2.2.

2.2.3 Human expert evaluations

An online survey was created for the manual evaluation of 200 WT/mutant trace pairs. A trace pair consisted of a single WT trace and a mutant trace. The same WT trace could be used in multiple pairs with different mutants. The WT structure determined from the mutate-and-map experiments was provided, along with the WT SHAPE trace, the mutant SHAPE trace, the overlay of the WT and mutant traces, and the difference between the WT and mutant trace (Kladwang *et al.*, 2011; Kladwang *et al.*, 2011). Survey participants were asked to label each WT/mutant pair as having: (1) no differences or small differences, (2) local differences, or (3) global differences (Methods Supplementary, Subsection 2.5.2.3). For the purpose of this survey, local differences were considered to be close to the mutation site in sequence space. Under this definition, local changers in secondary structure space may be misclassified as global changers. Similarly, global changers in secondary structure space may be misclassified as local changers.

Therefore, it is useful to consider secondary structure in structure change prediction, but the true secondary structure for an RNA is difficult to obtain experimentally. To address this we compared the expert classification to secondary structure prediction guided by SHAPE data. It is important to note that using predicted secondary structures in lieu of experimental structures is imperfect and likely increases the perceived secondary structure classification error by the experts. The experts did occasionally classify local changers in predicted secondary structure as global changers. However, the experts rarely classified global changers in secondary structure as local changers. (Table 2.10). Experts were filtered using a set of questions that gauged their familiarity with the biological sciences, RNA, RNA structure and SHAPE experiments. We identified 14 respondents in our survey results who self-identified as experts.

2.2.4 Feature and algorithm selection

Twenty-three features were initially used to quantify WT and mutant SHAPE trace differences and are reported in Table 2.2 and Table 2.4. These features rely solely on the experimental data and are completely independent of any structure prediction. Recursive feature elimination, using the caret package in R (Kuhn, 2008; Saeys *et al.*, 2007) identified 8 features from the set of 23 that optimally classified the human consensus. In addition we used the WEKA suite to execute thirty-five classification algorithms using the default settings with 5-fold cross-validation (Hall *et al.*, 2009). From these algorithms, random forest was selected as the most accurate for classification (Table 2.5) based on the number correctly predicted for non-changers. Assuming a tie at this level, we then selected the most accurate based on local changers and then global changers. We used this ranking because the distinction between change and no change is the most biologically important in our opinion. Further visual analysis of specific traces suggests that the random forest algorithm better distinguishes between local and non-changers than the

next best performing algorithms, Multilayer Perceptron and Kstar. This is particularly true for WT/mutant pairs with minimal differences in pattern, but sizeable differences in magnitude such as the G55U mutation in the 16S four-way junction, which we illustrate in Figure 2.6. KStar and Multilayer Perceptron mislabel the pair as a local changer, while Random Forest correctly identified the pair as a non-changer in agreement with the majority vote of experts. Although these minor differences in classification do not indicate that random forest is statistically better than Kstar and Multilayer Perceptron, the correct classification by random forest on these particularly difficult comparisons led us to choose it for implementation in the classSNitch approach. We built a random forest classifier on the set of 167 trace pairs using the randomForest R package with 5001 trees and default settings (Breiman, 2001; Liaw and Wiener, 2002). The random forest classifier was used to predict the classes for the entire set of 2019 normalized and noise reduced WT/mutant trace pairs. Feature selection, algorithm selection, and model building are further explained in Methods Supplementary, Subsection 2.5.2.4. The model's robustness to noise was tested using both simulated noise and repeated experiments (Figure 2.7).

2.2.5 classSNitch package

An R package was created for the identification of RNA structure change in large amounts of SHAPE data. The package includes methods for normalization, noise reduction, and calculating features. Feature calculations include pattern change, dynamic time warping, change contiguousness, Pearson correlation, Euclidean norm, change variance, eSDC and change range. The package can identify structure change in new SHAPE data sets based on an existing classifier. classSNitch is currently available at R-Forge. Documentation for classSNitch can be found at http://classSNitch.r-forge.r-project.org.

2.2.6 WT SHAPE improved SNPfold

We modified the SNPfold scoring scheme, which is based on the WT and mutant Pearson correlation coefficient (Halvorsen *et al.*, 2010), to include the WT SHAPE prediction as follows:

$$\text{Score} = \text{-SNPfold}_{\text{score}} + \text{SHAPE}_{\text{paired}} + \text{GorC} \quad (8)$$

where $SHAPE_{\{0,1\}}$ is 1 if the WT SHAPE reactivity is above the median value of the trace, 0 if it is below; $GorC_{\{0,1\}}$ is 1 if the WT nucleotide is a G or C, 0 otherwise. SNPfold is further explained in Methods Supplementary, Subsection 2.5.2.6.

2.3 Results

2.3.1 The "obvious" riboSNitch

Figure 2.1A illustrates the published secondary structure of the apo Glycine riboswitch based on multiple probing experiments, phylogenetic analysis and partial crystal structures (Butler *et al.*, 2011; Kladwang *et al.*, 2011). The nucleotides are color coded according to SHAPE reactivity (red high, yellow medium, and black low). In Figure 2.1B, the corresponding experimental SHAPE data for the WT RNA is plotted as a black line. A qualitative relationship between the structure and experimental data is evident when the data is presented in this way; in general paired nucleotides have low SHAPE reactivity, while unpaired bases have a "peak" in the profile. In a gel electropherogram, the peaks would be darker, and the paired nucleotides lighter. Figure 2.1C illustrates the experimental SHAPE data and corresponding SHAPE-directed structure prediction for the A125U mutation in the Glycine riboswitch. The overlay of the two traces reveals no visible difference between the WT (WT, black) and mutant (MUT, blue) trace;

the structure prediction is nearly identical to that of the WT. Not surprisingly, mutating A125 in domain 2 (P3) does not affect structure, as this nucleotide is not paired.

In Figure 2.1D we report the SHAPE-directed prediction for the A116U mutation, which occurs in the P3 helix of domain 2. In this case we see a local difference in the SHAPE trace, and the predicted structure does not contain this region of P3. This mutation has disrupted a single hairpin. It is important to note that the resulting SHAPE differences are readily visualized with the difference of the two traces (green trace, right panel). Figure 2.1E shows the effect of disrupting a base in the P2 stem in domain 2 with the A94U mutation. This results in a change in the P1 helix of domain 2 as well and is considered a global change. We chose to illustrate these three mutations from the 158 available for the Glycine riboswitch (Cruz *et al.*, 2012) as they are visually striking. As will be revealed below, not all mutation induced RNA structure change is as clear to visualize.

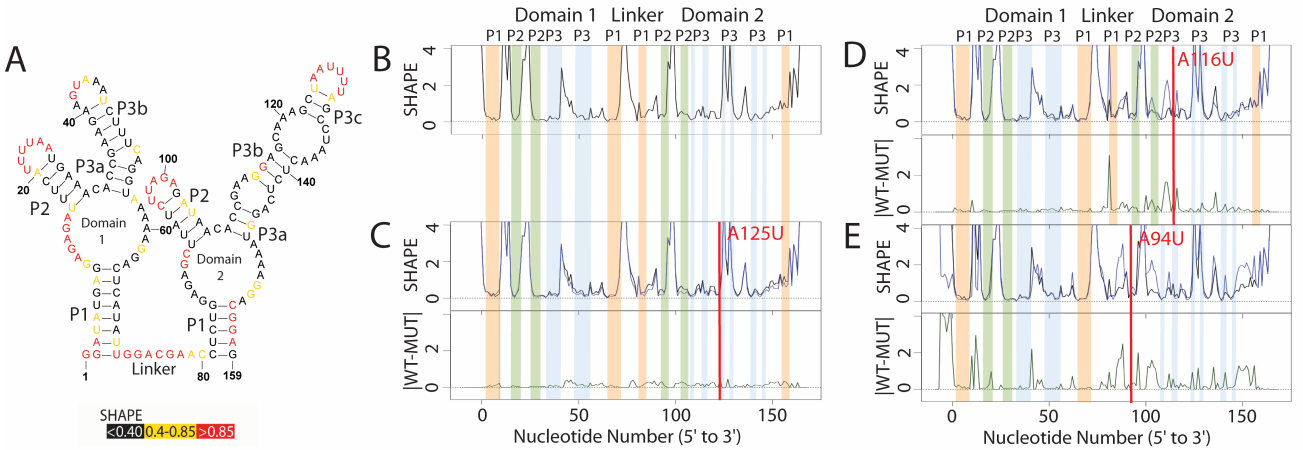


Figure 2.1 Structure change patterns in SHAPE trace data for glycine riboswitch aptamers

A) Published WT structure for the apo glycine riboswitch aptamers consistent with the crystal structure and multiple independent structure probing experiments (Butler *et al.*, 2011; Kladwang *et al.*, 2011). Red nucleotides indicate high SHAPE reactivity, yellow indicates mid-range reactivity, and black indicates low reactivity. B) The individual WT trace is shown in black; the colored bars indicate the structural regions for each of the aptamers: P1 (orange), P2 (green) and P3 (blue). C) The WT trace (black) is overlaid with the mutant SHAPE trace (dark blue), and the absolute difference between the WT and mutant traces is below (dark green). A red bar on the traces shows the mutation site. The A125U mutation is a mutation that leads to no appreciable differences in structure. 100% of experts that classified this mutant labeled it as a non-changer. D) The A116U mutation leads to a local structure change, where the mutant trace reactivity increases at the mutation site disrupting the P3 region of domain 2. 66% of experts that classified the A116U mutant labeled it as a local changer. E) The A94U mutation leads to a global structure change, where the mutant trace reactivity increases at both the mutation site and at nucleotides distant in sequence space disrupting both the P1 and P2 regions of domain 2. 66% of experts that classified the A94U mutant labeled it as a global changer.

2.3.2 Human consensus on local and global structure change

The complexity of interpreting SHAPE traces is illustrated in Figure 2.2. Here we plot the WT structure for the 16S four-way junction from the *E. coli* ribosome, as well as the mutant SHAPE data for A26U, A47U (P2b), and U99A (P1c). In each of these cases, it is not visually evident if the structure change is local, global, or if the data is simply inadequate. It is important to note that these SHAPE data are collected in a high throughput fashion, robotically, and often not replicated (Cheng *et al.*, 2015; Cordero and Das, 2015; Kladwang *et al.*, 2011; Miao *et al.*, 2015). This is one of the main differences in the way in which chemical and enzymatic probing is now collected. Because it can be collected in a very high throughput way, emphasis is placed on multiple experiments (all mutations in an RNA) rather than multiple replicates. Although it would be ideal to replicate these large-scale experiments there is a significant financial cost associated with multiple replicates.

Survey Statistics	
Total Traces	200
Total Experts	14
Total Responses	1427
Mean Trace Coverage	7.24
SD Trace Coverage	2.78
Mean Expert Agreement (%)	79.75
SD Expert Agreement (%)	0.79
Expert Reproducibility (%)	79.70
Total Non-Changers (Majority Consensus)	107
Total Local Changers (Majority Consensus)	40
Total Global Changers (Majority Consensus)	20

Table 2.1 Expert evaluation summary

Human survey statistics on WT/mutant SHAPE trace pair classification.

In visually inspecting traces like the ones illustrated in Figure 2.2A, we observed that in general most people in our lab agreed that A26U does not alter structure, A47U causes a local change, and U99A appears to alter the structure globally. We therefore decided to evaluate if RNA scientists, when presented with these types of traces and the accepted secondary structure of the RNA, agree on the classification of these data into none, local and global change. We recruited 14 volunteers from multiple RNA labs to answer an online survey in which each person would classify up to 200 traces (WT/MUT comparisons) into none, local and global changes. In total 1427 comparisons were manually classified, with an average of seven views for each trace (Table 2.1). From this data we built a consensus human classification of the traces and evaluated each expert's ROC (receiver operator curve) area under the curve (AUC) to the consensus (Figure 2.2B). Since this is a three-way classification we evaluate AUC pairwise for none, local and global change. As can be seen the expert reproducibility is high (AUC average above 0.8) which indicates RNA scientists agree with each other at least with respect to what structure change looks like in a SHAPE trace. We also evaluate human three-way AUC using a cobweb plot (Figure 2.2C). This shows that the largest disagreement between self-reported RNA SHAPE experts is in their classification of local versus global change. The average AUC is still 0.8 (blue) suggesting the disagreement is weak. The green AUC curves in Figure 2.3A, show that for all but distinguishing global vs. none (rightmost graph) eSDC performs quite poorly.

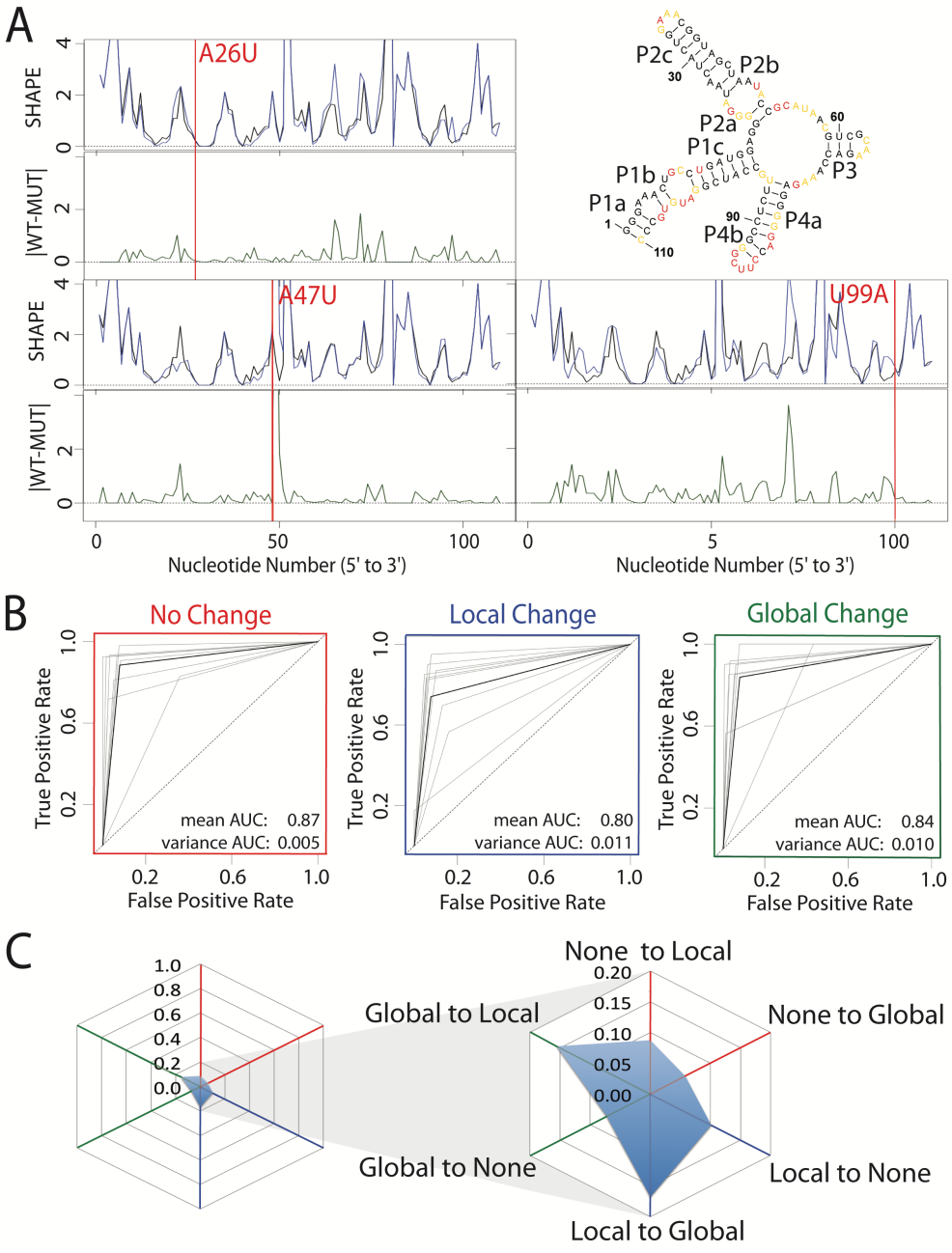


Figure 2.2 Expert evaluation of RNA structure change in SHAPE data

A) Accepted WT structure for the 16S four-way junction domain from the *E. coli* ribosome in agreement with the crystal structure and multiple structure probing experiments (Cordero and Das, 2015; Tian *et al.*, 2014; Zhang *et al.*, 2009). Red nucleotides indicate high SHAPE

reactivity, yellow indicates mid-range reactivity, and black indicates low reactivity. For each mutant, the WT trace (black) is overlaid with the mutant SHAPE trace (dark blue). The absolute difference between the WT and mutant traces is depicted below (dark green). 100% of experts that evaluated the A26U WT mutant pair agree that there is no difference or a small difference. 88% of experts agree that the A47U mutation creates a local difference. Experts are split on the U99A mutation. 37.5% of experts indicated that the mutation creates no difference or a small difference, 37.5% of experts indicated that the mutation creates a local difference and 25% of experts indicated the mutation creates a global or distant mutation. B) ROC curve analysis was used to compare expert classification to the majority vote consensus. The gray curves represent individual expert performances, while the black curves show the average performance among experts. The ROC curves are depicted for performance in identifying non-changers (red), local changers (blue) and global changers (green). C) Cobweb plots show the percentage of mutants mislabeled by the expert majority vote with non-changers on the red axes, local changers on the blue axes, and global changers on the green axes. Expert classification is least consistent on differences between global and local changers with a higher percentage of global changers mislabeled as local changers, and local changers mislabeled as global changers.

We also investigated whether another standard metric, the Euclidean distance (blue AUC) did any better and observed a similar trend. The mean expert performance is shown in black, and is far superior to any single metric. Thus, to achieve consensus, RNA scientists must be looking at other features in the data than simple correlations in the pattern. We set out to discover what these are and to develop an automated classification system of RNA structure change that simulates human consensus calls.

Feature	Formula	Description
Pearson CC	$^{P}CC(SHAPE_{ref}, SHAPE_{alt})$	Pearson correlation coefficient is the covariance between the wild type and mutant trace SHAPE values divided by their standard deviations. Additional descriptions can be found in Figure 2.8.
Pattern CC	$^{P}CC(Change_{ref}, Change_{alt})$	Pattern correlation coefficient is the Pearson correlation coefficient between wild type and mutant trace patterns. The trace pattern is given by increase (+1), decrease (-1) or no change (0) in SHAPE value moving from one nucleotide to the next across the entire length of the RNAs. The pattern change between wild type and mutant traces are positions where the trace patterns different. Additional descriptions can be found in Figure 2.8.
Contiguousness	# of $i_{contiguous}$	Contiguousness is the number of contiguous stretches of pattern change between wild type and mutant traces. See Pattern CC. Additional descriptions can be found in Figure 2.8.
Change Range	$max(i_{diff}) - min(i_{diff})$	Change range is the interval containing all pattern changes between wild type and mutant traces. See Pattern CC. Additional descriptions can be found in Figure 2.8.
Change Variance	$\Sigma_i(i_{diff} - mean(i_{diff}))/N$	Change variance is the spread of pattern change distances between the wild type and mutant traces. The pattern change distance is the distance away from the mutation site (in nucleotides) that a pattern change occurs. See Pattern CC. Additional descriptions can be found in Figure 2.8.
Dynamic time warping	dynamic time warping algorithm	Dynamic time warping is an algorithm to optimally align wild type and mutant traces by "warping" one into the other (Giorgino, 2009). Dynamic time warping aligns two series on the sides of a grid. The distance between each point in the two series is calculated for every position in the grid. Summing over the minimum distance path along the grid gives the overall distance. Additional descriptions can be found in Figure 2.9.
eSDC	$(1-^{P}CC(SHAPE_{ref}, SHAPE_{alt})*sqrt(N)$	Experimental structural disruption coefficient is 1 minus the Pearson correlation coefficient between the wild type and mutant traces, normalized by the square root of the length of the RNA (Ritz, et. al, 2012). See Pearson CC. Additional descriptions can be found in Figure 2.8.
Euclidean Norm	$\Sigma_i(SHAPE_{ref}[i]-SHAPE_{alt}[i])^2$	Euclidean norm is the L2-norm or distance between the wild type and mutant traces. The distance is calculated as the sum over the squared difference between wild type and mutant traces. Additional descriptions can be found in Figure 2.8.

Table 2.2 Features used to quantify differences between WT and mutant traces

Feature formulas and descriptions for the 8 features included in the model. These 8 features were chosen by recursive feature elimination from the total set of 23 features (Methods

Supplementary, Subsection 2.5.2.4). The formula symbol descriptions are included in Table 2.7. Additional descriptions for these methods can be found in Figures 2.8 and 2.9. A list of feature statistics can be found in Table 2.9

2.3.3 Automated classification of mutation induced structure change

To develop an automated classifier for identifying mutation induced structure changes in RNA we began by establishing a list of 23 features commonly used to evaluate quantitative differences between two linear data sets (Table 2.2 and Table 2.4). Using the human survey classification (Table 2.1) for supervised learning, we trained 38 different algorithms and evaluated their accuracy. The results of this training are provided in Table 2.5 and suggest the Random Forest classifier performs the best on this data using the eight features found in Table 2.2. The trained random forest classifier on these eight features is the algorithm used in the classSNitch R package released with this manuscript.

Interestingly no single feature drives the classification, indicating that the human experts are looking at multiple features of the signal to decide what is or is not a change. Nonetheless we performed random feature elimination and did identify that dynamic time warping alone achieves an accuracy of 65% (Figure 2.10A). Dynamic time warping is less sensitive to distortion caused by local misalignments, a quality that makes the technique useful in speech recognition and likely contributes to the feature's success in our classifier (Sakoe and Chibe, 1978). We also ranked the eight features by their importance and see that each feature increases accuracy incrementally when added to the model in approximately equal increments. Plotting the WT to mutant Pearson correlation coefficient and contiguousness versus dynamic time warping (Figure 2.10B) reveals how these features correlate but also illustrates subtle differences in how these different features classify change.

We illustrate the basic dynamic time warping principle in Figure 2.9A and how we score differences based on this trace alignment strategy. The score increases as the two traces differ and is calculated over the entire alignment. Dynamic time warping is visualized on the U99A

data in Figure 2.9B. It identifies the minimum number of insertions and deletions to optimally align the mutant and WT traces. As such, a higher dynamic time warping score indicates greater differences in the traces. It is therefore likely that the expert humans are performing some form of trace alignment combined with pattern matching when evaluating the data. Processing SHAPE data (whether it is obtained by capillary or next generation sequencing) requires an alignment strategy. It is not surprising that humans may choose to ignore small frame shifts in the data (which lead to very high eSDC values) since they know these are most likely errors in trace alignment (Figure 2.11).

Overall, the classSNitch performance (purple line Figure 2.3A) is equivalent to human consensus for none, local and global change. The cobweb plot reveals that the highest error rate in classSNitch classification is false negatives for local change (Figure 2.3B). In comparison to eSDC and the Euclidean distance (green and blue AUC, respectively) our classifier performs significantly better. Thus classSNitch is a good approximation of human expert classification of SHAPE trace differences and applying it to high-throughput mutational data sets can simulate human consensus classification of these data.

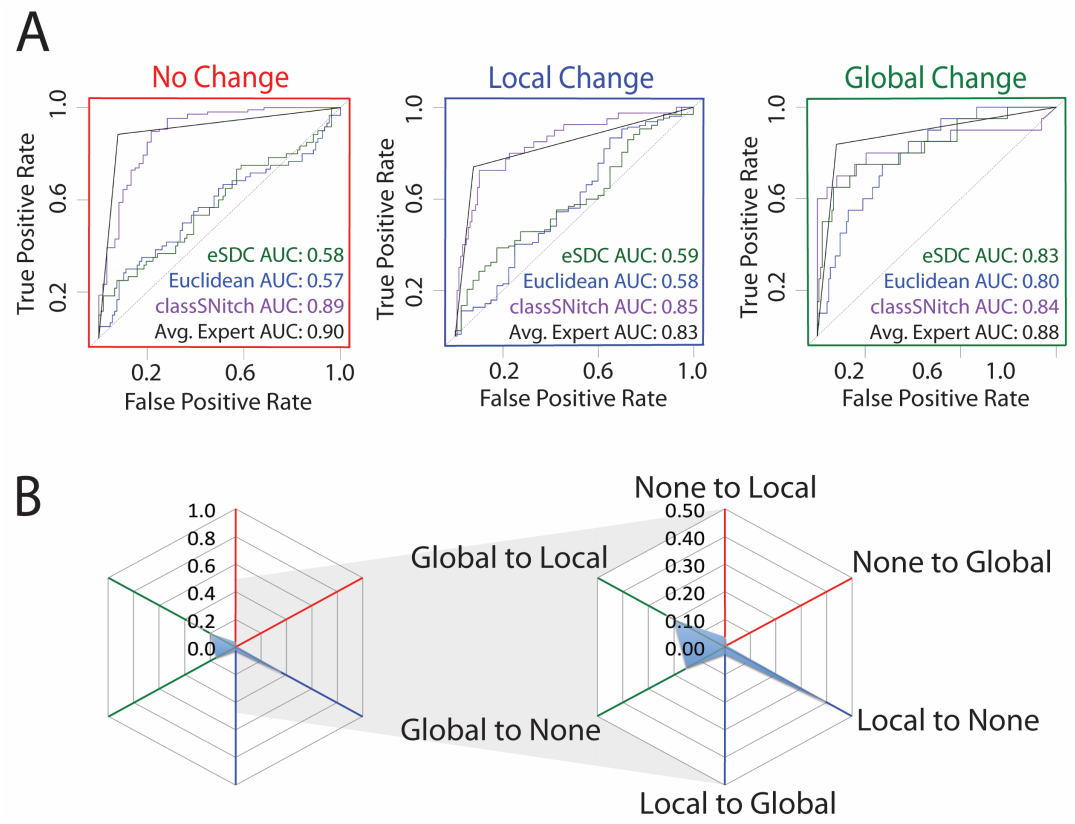


Figure 2.3 classSNitch performance

A) ROC curve analysis comparing methods for classifying structure change to the majority consensus by experts. The ROC curves are depicted for performance in identifying non-changers (red), local changers (blue) and global changers (green). The methods used for experimental classification are classSNitch (purple), eSDC (green), Euclidean norm (blue), and the mean expert human performance (black). Consistently, classSNitch performs comparably to the mean expert evaluation. classSNitch outperforms eSDC and the Euclidean norm, which are the current metrics for classifying RNA structure change in SHAPE data. B) The cobweb plot shows the percentage of traces mislabeled by classSNitch; a higher percentage of local changers are misclassified.

2.3.4 classSNitch analysis of experimental structure change

The training data used for the development of the classSNitch classifier (Table 2.1) represents a small subset of publically available mutational SHAPE data (Cordero *et al.*, 2012). We identified a total of 2019 SHAPE traces for eleven different RNAs (Table 2.3). We classified these using the classSNitch algorithm excluding the training set of 167 RNAs. In this data set we identified 382 local changers (19%), and 111 global changers (5%). When these data are further broken down by RNA (Figure 2.4A) we immediately observe significant differences in the sensitivity of mutation in these RNAs. Some RNAs, like the homeobox (Hox) A9 5'UTR, are more resistant to mutations. The Hox mRNAs are involved in development, and the 5'UTR plays an important role in ribosome-mediated translational control. It is highly structured and folding to a specific conformation is essential to function (Alexander *et al.*, 2009; Xue *et al.*, 2015). Similarly, the phenylalanine-transfer RNA, 16S four-way junction and 5S ribosomal RNA are also relatively resistant to mutation. Other RNAs are more sensitive to mutations, like the synthetic Tebowned aptamer that was designed in the Eterna laboratory as part of their online game (Cordero and Das, 2015; Lee *et al.*, 2014). RNAs folded in different solution conditions, such as aptamers in the absence or presence of their ligand, respond differently to mutation as well (Figure 2.4B). For the adenine and glycine riboswitches, ligand binding increases the RNA's sensitivity to mutations. The synthetic Tebowned aptamer has decreased sensitivity to mutations when in the presence of ligand. The chemical modifier used in chemical mapping experiments also affects the SHAPE data and ultimately sensitivity to structure change (Figure 2.12). N-methlyiastoic anhydride (NMIA) is less reactive and requires a longer time to react than 1-methyl-7 nitroisatoic anhydride (1M7) (Mortimer and Weeks, 2007). Given the kinetics of the

reaction, it is not surprising that 1M7 can detect more subtle differences in structure that could be occurring on a shorter time scale.

Most structure prediction programs have low accuracy when identifying experimental riboSNitches with AUC values ranging from 0.6-0.7 (Corley *et al.*, 2015; Ritz *et al.*, 2012). In these benchmark studies, validation of the experimental data is analyzed using simple metrics like eSDC or the Euclidean distance (Corley *et al.*, 2015; Ritz *et al.*, 2012). One possible explanation for the poor predictive performance of the prediction algorithms in these benchmark studies is misclassification of the experimental data with these simple metrics. Indeed, when we observe the performance of SNPfold on data classified with either eSDC or Euclidean difference, the AUC values indicate the algorithm is barely predictive (Figure 2.5A). We observe a subtle improvement in performance when we use the classSNitch classification of the experimental data. A similar performance increase is observed for the other published algorithms designed for riboSNitch prediction (Figure 2.5B) (Halvorsen *et al.*, 2010; Sabarinathan *et al.*, 2013; Salari *et al.*, 2013). Thus, misclassification of experimental data is likely a confounding factor for the poor performance of riboSNitch prediction algorithms, and the use of classSNitch in future benchmarking studies may improve prediction accuracy. Details on algorithm parameters can be found in Methods Supplementary, section 2.5.3.4.

The mutational strategy data is based primarily on four types of transversion mutations (Kladwang *et al.*, 2011) as seen in Table 2.6. The data presented in this table indicates mutating C or G in the WT sequence is more likely to induce structure change than mutating A or U with an odds ratio of 1.9, $p<0.001$. We also observed that low SHAPE reactivity in the experimentally predicted WT structure is more likely to lead to structure change when mutated (OR=1.4, $p<0.05$).

2.3.5 WT SHAPE informed riboSNitch detection

It is well established that incorporating SHAPE into RNA structure folding algorithms improves secondary prediction performance (Diegan *et al.*, 2009). Since we use SHAPE data to detect riboSNitches, it does not make sense to include experimental data for the WT and mutant in structure predictions. Nonetheless our analysis of sequence composition and WT SHAPE data for local and global changers does suggest an alternative. Can the WT SHAPE trace alone inform riboSNitch predictions? This is an attractive strategy since ultra high-throughput techniques exist to collect WT data on a genome-wide scale (Siegfried *et al.*, 2014).

The major bottleneck in collecting systematic mutational information is the molecular biology required to synthesize and validate each mutant. When we modify the SNPfold algorithm scoring to include WT SHAPE data and to take into account the type of mutation (Eq. 8), we are able to improve the performance of our algorithm further (Figure 2.5B). Thus the WT SHAPE data is useful in increasing the accuracy of riboSNitch prediction.

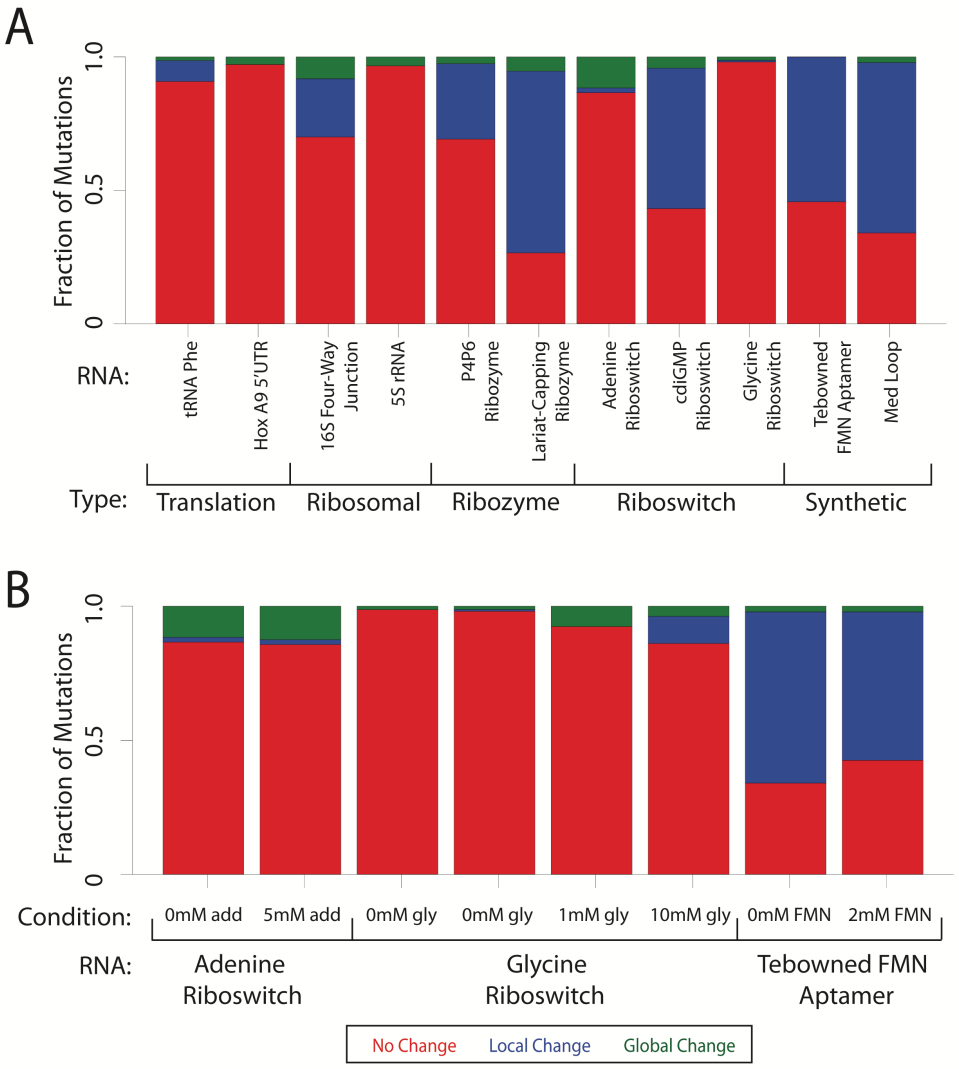


Figure 2.4 Fraction of disruption for individual RNAs

A) The fraction of mutations that cause no change (red), local change (blue) or global change (green) for each RNA as classified by classSNitch. The RNAs are grouped by biological function: translation, ribosomal, ribozyme, riboswitch or synthetic. The experimental conditions for each of these RNAs are listed in Table 2.3B) The fraction of mutations that cause aptamers to change structure in the absence or presence of differing amounts of ligand for the adenine riboswitch, glycine riboswitch and Tebowned FMN aptamer.

2.4 Discussion

Identifying mutations that are likely to lead to changes in RNA structure remains a significant computational and experimental challenge (Chauhan and Woodson, 2008; Cheng *et al.*, 2005; Churkin *et al.*, 2011; Russell *et al.*, 2002). Such predictions are important in the context of personalized medicine since many riboSNitches are now known to be causative of human disease (Solem *et al.*, 2015). Despite the advent of experimental technology enabling us to probe structure on a genome-wide scale, we still rely on structure change prediction algorithms or visual interpretations of the data to detect riboSNitches as there is no ultra-high throughput approach for rapidly mutating an RNA (Ritz *et al.*, 2012; Rocca-Serra *et al.*, 2011; Sansone *et al.*, 2012; Siegfried *et al.*, 2014).

We hypothesized that one reason for the poor performance of RNA structure prediction algorithms (Corley *et al.*, 2015; Ritz *et al.*, 2012) on riboSNitches is the misclassification of the experimental data. We therefore set out to develop novel metrics to evaluate structure change from SHAPE data. This approach did lead to modest improvements in performance suggesting that careful analysis of SHAPE data is essential when using these data as a benchmark. In this age of whole transcriptomic structure probing, manual validation and curation of these data sets is impractical. The classSNitch classifier simulates human consensus on what is and is not a structure change and therefore offers an alternative to simple metrics like eSDC in experimentally describing RNA structure change.

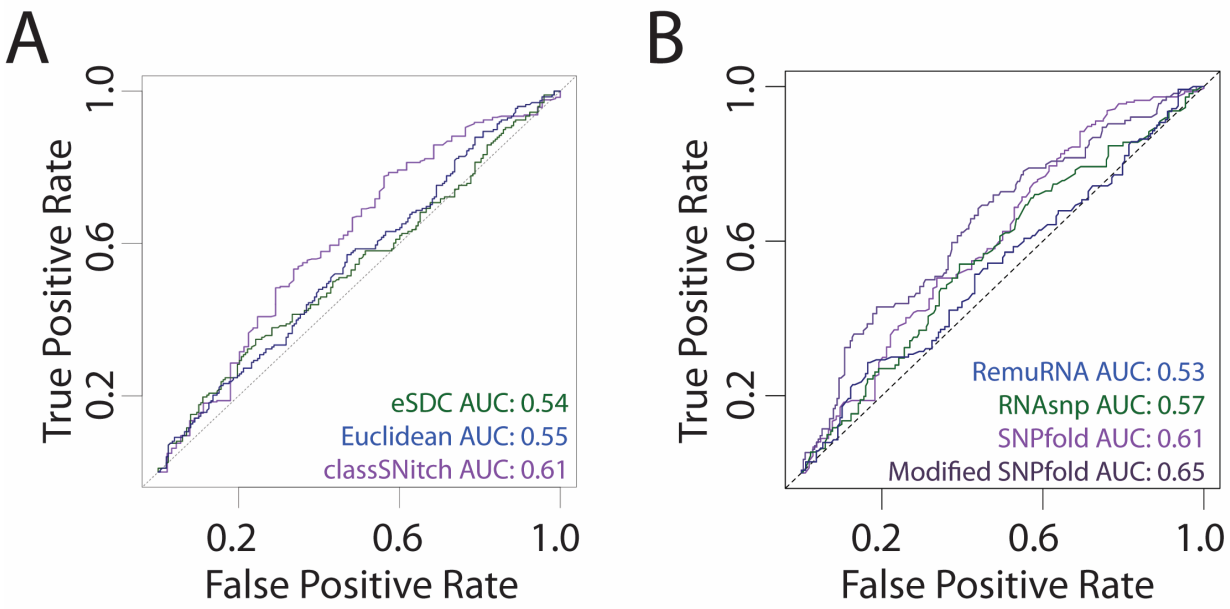


Figure 2.5 Improving the performance of structure change prediction algorithms

A) We performed ROC curve analysis for SNPfold, a structure change prediction algorithm, using classSNitch (purple), eSDC (green) and the Euclidean norm (blue) to classify the experimental data using the 10% tails strategy (Corley *et al.*, 2015). B) We compare the performance of structure change prediction algorithms on the classSNitch classification for SNPfold (purple), RNAsnp (green) and RemuRNA (blue). Each of these algorithms predicts structure change in RNA using only sequence information. SNPfold, remuRNA, and RNAsnp all make *ab initio* predictions on whether a mutation alters the RNA structure; none of the algorithms benchmarked used SHAPE-directed structure prediction since we are using the WT and mutant SHAPE data for experimental validation. We improved the SNPfold prediction (dark purple) using Eq. 8. The ROC curves for local and global change predictions are included in Figure 2.13.

The features that classSNitch uses to classify change reveals some of the subtleties involved in interpreting SHAPE data. Beyond evaluating the magnitude difference between traces, human experts also utilize information on pattern matching and the distribution of change along the length of the RNA (Figures 2.8 and 2.9). We used those features to develop a classifier that successfully mimics expert classification of structure change (Figure 2.3). SHAPE reactivity is correlated with secondary structure, more reactive nucleotides are generally single stranded (Eddy, 2014); however the experiment probes the overall structure of the RNA. The classSNitch classifier does not attempt to model structure, but instead establishes a standard for quantifying change. This is biologically relevant, allowing us to compare different RNAs using a standard vocabulary (Figure 2.4). Although only two synthetic RNAs are included in our data set, there is a striking difference in their sensitivity to mutation (Figure 2.4A). Indeed a much larger fraction of the mutations in these RNAs result in conformational rearrangement. Although with only two RNAs it is impossible to draw statistical conclusions, this observation remains biologically interesting and warrants further investigation as more experimental data is obtained on a wide variety of RNAs (both synthetic and naturally occurring). The idea that RNA sequences under natural evolutionary pressure may evolve a general robustness to mutation warrants further investigation.

The data used for training classSNitch was exclusively collected using traditional capillary methods of electrophoresis. The quantification of this type of data from a capillary trace is a challenge, as it requires alignment to a reference ladder (Das *et al.*, 2005; Karabiber *et al.*, 2013; Mitra *et al.*, 2008). Recent algorithmic developments have further automated this process and increased reliability (Yoon *et al.*, 2011). It is interesting that dynamic time warping is the most significant feature used by classSNitch in reproducing expert classification. If alignment

errors were to persist in the data, one might expect that experts could be correcting these when gazing at the data. As technology has evolved, in particular with the use of next generation sequencing to collect chemical and enzymatic probing data (Kertesz *et al.*, 2010; Mortimer *et al.*, 2012; Rouskin *et al.*, 2014; Siegfried *et al.*, 2014) alignment artifacts may disappear in the data. As such it may become necessary to retrain classSNitch on these newer types of data. In our lab's limited experience with these types of data (currently unpublished), classSNitch performance is similar regardless of the type of data analyzed. However, it will be necessary to continue evaluating classSNitch performance as new experimental modalities are used. SHAPE data measures the selective reactivity of a probe for the 2' OH of the RNA (Diegan *et al.*, 2009). As such, the direct relationship between structure and reactivity is complex and ultimately depends on the 3-D structure of RNA. As a result, differences in SHAPE data due to mutation (or exogenous molecule binding) are notoriously difficult to interpret (Kutchko and Laederach, 2016). This does not however mean that SHAPE data does not contain useful information. Our use of the WT SHAPE data to improve riboSNitch predictions (Eq. 8, Figure 2.5B) indicates that much as including SHAPE as a free energy term in structure prediction (Diegan *et al.*, 2009), aspects of the reactivity can inform predictions. It is likely that the improvement we observe when using Eq. 8, which does not include any free energy terms, is due to the fact that in general, higher SHAPE reactivities are indicative of unpaired nucleotides (Eddy, 2014; Kutchko and Laederach, 2016). The by effectively adjusting the SNPfold score for nucleotides that are likely unpaired in the WT structure, which also are less likely to cause a riboSNitch, we observe a modest improvement in prediction performance. This effect remains modest since the correlation between SHAPE reactivity and base-pair probability is only moderate (Kutchko and Laederach, 2016).

Although classSNitch was trained on riboSNitches and is primarily intended as a tool to evaluate the effect of mutation induced structure change, it is in fact a more general metric for comparing SHAPE data. RNAs will adopt alternative conformations depending on their environment. For example riboswitches adopt different conformations depending on the presence of the ligand. When applied to the WT traces of apo and bound riboswitch data, the algorithm does identify local and global change for a majority of riboswitches, as expected. Protein binding, changes in cellular environment and even counter-ions are known to affect RNA structure (Bai *et al.*, 2005; Frederiksen *et al.*, 2012). The classSNitch classifier provides a common language to describe these differences. For example, it could be used when comparing *in vivo* and *in vitro* probing of the RNA to identify regions where the presence of proteins alters structure locally and globally. It also offers an attractive way to quantify these changes in agreement with expert consensus.

Manual classification of traces remains a laborious process, and is the main reason we developed the classSNitch classifier. We limited our training set to 200 traces and were able to recruit 17 experts to classify a majority of these traces. Certainly, a larger number of manual classifications will further improve the performance and precision of our classifier, especially for difficult cases. As such it is important when using the classSNitch classifier to be aware of the limited size of the training set and exercise care in evaluating the predictions on novel data. In particular, the performance of the classifier was with only 5 cross-validation folds in lieu of an independent test set, and as such is likely still somewhat partial. Nonetheless our data do suggest that it will be possible to arrive at a consensus for what a small and large RNA structure change look like and that the approach we present here is viable for developing a community standard.

The agreement between human experts "gazing" at this data is reassuring. Prior to quantitative methods being widely available to life scientists, significant progress was achieved by carefully looking at the data; the structure of group I introns, tRNA, and the ribosome were correctly predicted manually years before the were crystallized (Michel and Westhof, 1990). The value of automated systems that reproduce human appreciation of data is underutilized in RNA structural research despite the rich history of success in the field. Developing the classSNitch classifier minimally captures dying expert knowledge, while also making this expertise accessible to the community in an automated package.

2.5 Methods Supplementary

2.5.2 Methods

2.5.2.2 Data normalization and noise reduction

Variables:

n = nucleotide position

N = trace length in nucleotides

WT = wild type reactivities

MT = mutant reactivities

WTnorm = normalized wild-type trace

MTnorm = normalized mutant trace

WTreduc = noise reduced

MTreduc = noise reduced

HIGH = normalized "high reactivity" value determined by QuSHAPE

Normalization: We normalized the mean of every WT trace to a mean of 1.5. This step increases the mean of the WT traces so that differences in magnitude are more pronounced.

$$WTnorm = (1.5N/\Sigma WT) * WT \tag{9}$$

We normalized each MT by the multiplier that minimizes the absolute difference between the WT and MT. This step minimizes small differences in magnitude between the WT and MT traces that may be attributable to noise.

$$MTnorm = argmin_x(\Sigma(|WT - x * MT|)) * MT \tag{10}$$

Noise Reduction: For every nucleotide [n], if the value in both WT and MT are higher than the normalized "high reactivity" value determined by QuSHAPE, we set MT[n] equal to WT[n]. This step minimizes differences in magnitude when both WT and MT have high reactivity.

$$\text{if (WTnorm[n] > HIGH \& MTnorm[n] > HIGH)\{MTreduc = WTreduc\}} \tag{11}$$

For every nucleotide [n], if the value is less than -0.5, set them equal to 0. This step minimizes differences in magnitude when both WT and MT have small reactivity.

$$\text{if (WTnorm[n] < -0.5)\{WTreduc = 0\}} \tag{12}$$

$$\text{if (MTnorm[n] < -0.5)\{MTreduc = 0\}} \tag{13}$$

2.5.2.3 Human expert evaluations

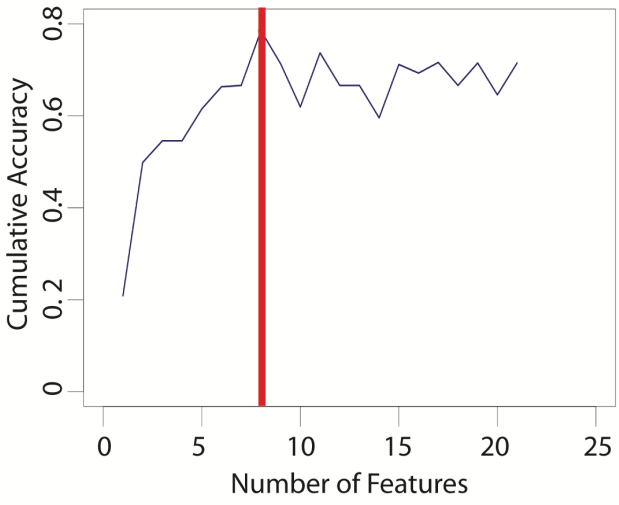
Non-changer: A mutation that leads to no difference between the wild type and mutant SHAPE traces. Small differences at the mutation site or the two nucleotides immediately adjacent to the mutation site may be caused by a difference in reactivity between the modifier and specific nucleotide types. Due to this difference, changes in this region are ignored.

Local changer: A mutation that leads to a difference between the wild type and mutant SHAPE traces in the 20-nucleotide region surrounding the mutation site (10 nucleotides on either side). This is the average region of change around the mutation site for mutants labeled as local changers by the experts.

Global changer: A mutation that leads to differences between the wild type and mutant SHAPE traces beyond the 20-nucleotide region surround the mutation site. The change may be contiguous or separated by some distance and may also include change around the mutation site.

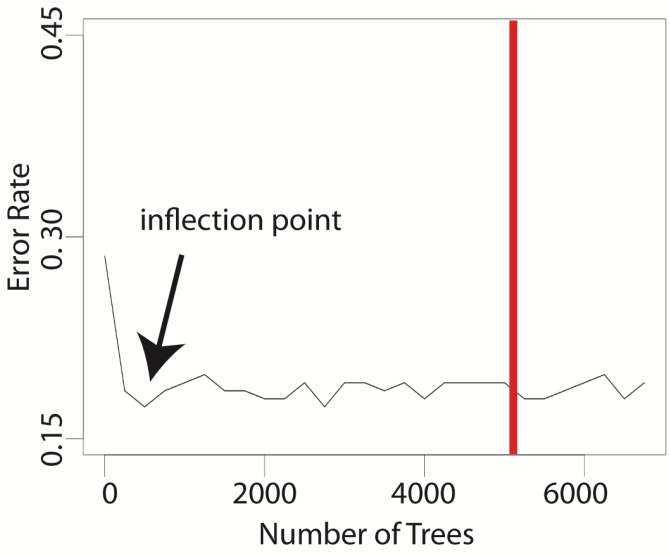
2.5.2.4 Feature and algorithm selection

k-fold cross-validation (CV): CV is a method used for model validation. In CV the data is divided into *k* subsamples. *k*-1 subsamples are used to build the model and the remaining subset is used for testing. This is done for each of the *k* subsamples (Hall *et al.*, 2009). We did cross-validation using 5 subsets where a single RNA data set is always grouped together, but multiple RNA data sets may be in a split. The subsets are as close in size as can be achieved with this constraint. The



same 5 subsets were used for validation of every algorithm. The traces used in the validation folds are further described in Table 2.8. Random forest performed the best across validation for local, global and non-changers. Since the available number of manually classified samples was limited, we decided to average the performance over the 5 cross-validation sets (Brun *et al.*, 2008; Molinaro *et al.*, 2005).

Recursive Feature Elimination (RFE): RFE is a method for choosing a subset of appropriate features for use in a predictive model (Kuhn, 2008; Saeys *et al.*, 2007; Zhou *et al*., 2014). We initially built a random forest model using all 23 features. To rank the features, we used permutation importance, which measures the mean decrease in accuracy when a feature's values are shuffled (Kuhn, 2008; Saeys *et al*., 2007; Zhou *et al*., 2014). We systematically remove one feature and retrain the models. This recursive process effectively ranks the feature's importance when repeated, in our case, 10 times. Each time the order of the features remained the same, but the cumulative accuracy varied. Averaging the cumulative accuracy over 10 runs, we selected the number of features beyond which the cumulative accuracy stabilized (the accuracy no longer increased). Thus this procedure allows us to identify the seven features we ultimately implemented in the classSNitch classifier. To perform recursive feature analysis, we used the rfeControl and rfe functions in the caret R-package with the following parameters: random forest function, bootstrap resampling and 10 iterations.



Random Forest (RF): Random Forest is a supervised learning technique using decision trees that can be used for classification or regression (Breiman, 2001; Liaw and Wiener, 2002). A decision tree groups the samples into different nodes according to the feature being measured. The root node includes all of the samples. In the first step a feature is selected at random and used to split the tree. The tree grows by choosing the best split based on the selected feature and breaking the samples into two new nodes. A node that can no longer be split is called a leaf node, because all samples are identical or the node only contains a single sample. Leaf nodes may contain samples with a mixture of expert class labels. For each sample, the tree "votes" on a class based on the majority of class labels in the leaf node where it is found. This is a supervised learning algorithm where the expert classifications determine the best tree from which to build the classifier. Random forest uses a set or forest of these decision trees (Breiman, 2001; Liaw and Wiener, 2002). Each individual tree samples with replacement from the full set of data, such that each tree is built on a different subset. Due to this sampling, trees may disagree on class votes for individual samples. The class for each sample is determined by the majority vote among all of the trees. Sampling with replacement results in some data being left out of the tree, referred to as "out-of-bag". Each tree can predict the class for its "out-of-bag" samples. The classification error for the out of bag samples gives the generalization error. The classification for new samples can be determined from an existing forest of trees (Breiman, 2001; Liaw and Wiener, 2002). We chose a number of random forest trees greater than the number beyond which the error rate (averaged over 10 runs) stabilized.

Algorithm Parameter Optimization: All algorithms were initially run using the default settings in Weka (Hall *et al.*, 2009). Based on the number of correctly predicted non-changers, the top three performing algorithms were selected for futher optimization. The results after

optimization are those reported in Table 2.5. The optimization for random forest is described above. KStar is an instance-based learner that assigns classes for new instances according to its k-nearest data points (Cleary and Trigg, 1995). Similarity between samples in KStar is determined by entropy. By gradually increasing the global blending parameter, we found the value that resulted in the maximum number correctly classified for non-changers, 22%. Multilayer perceptron is an artificial neural network (Silva *et al.*, 2008). The algorithm consists of layers of nodes in a directed graph, where each node is a nonlinear activation function. In multilayer perceptron the network is trained through backpropagation using gradient descent. We determined the optimal learning rate, momentum number of hidden layers and number of nodes in a given layer by gradually increasing each parameter individually and leaving the others at their default setting. The optimal learning rate was 0.1, the momentum was 0.65, the number of hidden layers was 1, and the single hidden layer had 5 nodes.

2.5.2.6 WT SHAPE improved SNPfold

SNPfold: SNPfold is an algorithm that identifies structure-changing single nucleotide polymorphisms (SNPs). Previously SNPfold has been used to identify structure-changing mutations in the Human Gene Mutation Database that map to untranslated regions of RNA (Halvorsen *et al.*, 2010). This algorithm uses only sequence information to predict the partition function of an RNA. The partition function is a representation of the ensemble of structures that an RNA may form. Summing over the columns in a partition function gives the base-pairing probabilities for a given RNA. The Pearson Correlation coefficient between the base-pairing probabilities is then used to compare each wild type-mutant pair (Halvorsen *et al.*, 2010). Default parameter settings for accurate p-value calculation was used for analysis with SNPfold. We used the p-value for the correlation coefficient as the measure.

SNPfold.py -A <seq file> <mutation> (14)

2.5.2.7 Figures and Diagrams

The minimum free energy structures were predicted using the RNAstructure package suite with SHAPE data to constrain the prediction (Diegan *et al.*, 2009; Matthews, 2004; Matthews *et al.*, 2004). Structure diagrams were created using VARNA (Darty *et al.*, 2009). Cobweb diagrams were used for multi-class performance analysis (Diri and Albayrak, 2008).

2.5.3 Results

2.5.3.4 classSNitch analysis of experimental structure change

RNAsnp: We used default parameter settings for mode 1 (global folding) with RNAsnp (Sabarinathan *et al.*, 2013). The window size for including base pairing probabilities is 200 nucleotides on either side of the mutation. We used the p-value for the correlation coefficient as the measure.

RemuRNA: We used default parameter settings for analysis with RemuRNA (Salari *et al.*, 2013). The default window size was used for calculating the non-localized measure. We used the relative entropy as the measure.

2.6 Supplementary Materials

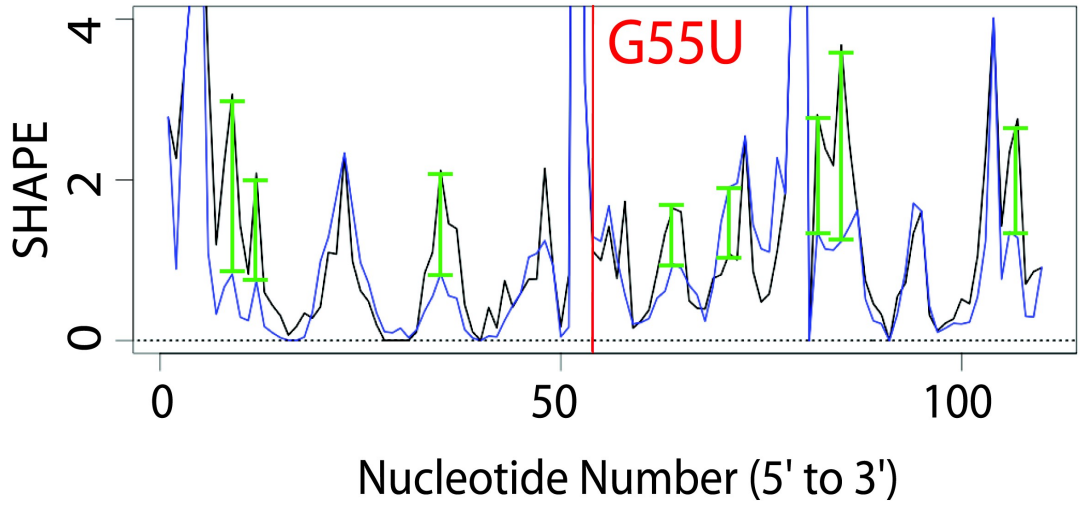


Figure 2.6 Differences between top performing algorithms

The top three performing algorithms are able to distinguish well between global changers and local or non-changers. In most examples, these algorithms are also able to distinguish between local changers and non-changers. However, random forest may have a slight advantage over KStar and Multilayer Perceptron in distinguishing between local and non-changers when there is little or no change in pattern, but a sizeable change in magnitude. An example of this scenario is seen in the 16S four-way junction G55U mutant trace (blue) overlaid with the WT trace (black). Green vertical lines highlight the difference in magnitude. This trace pair was labeled as a non-changer by a majority vote among experts. KStar and Multilayer Perceptron mislabeled the pair as a local changer, while Random Forest correctly identified the pair as a non-changer.

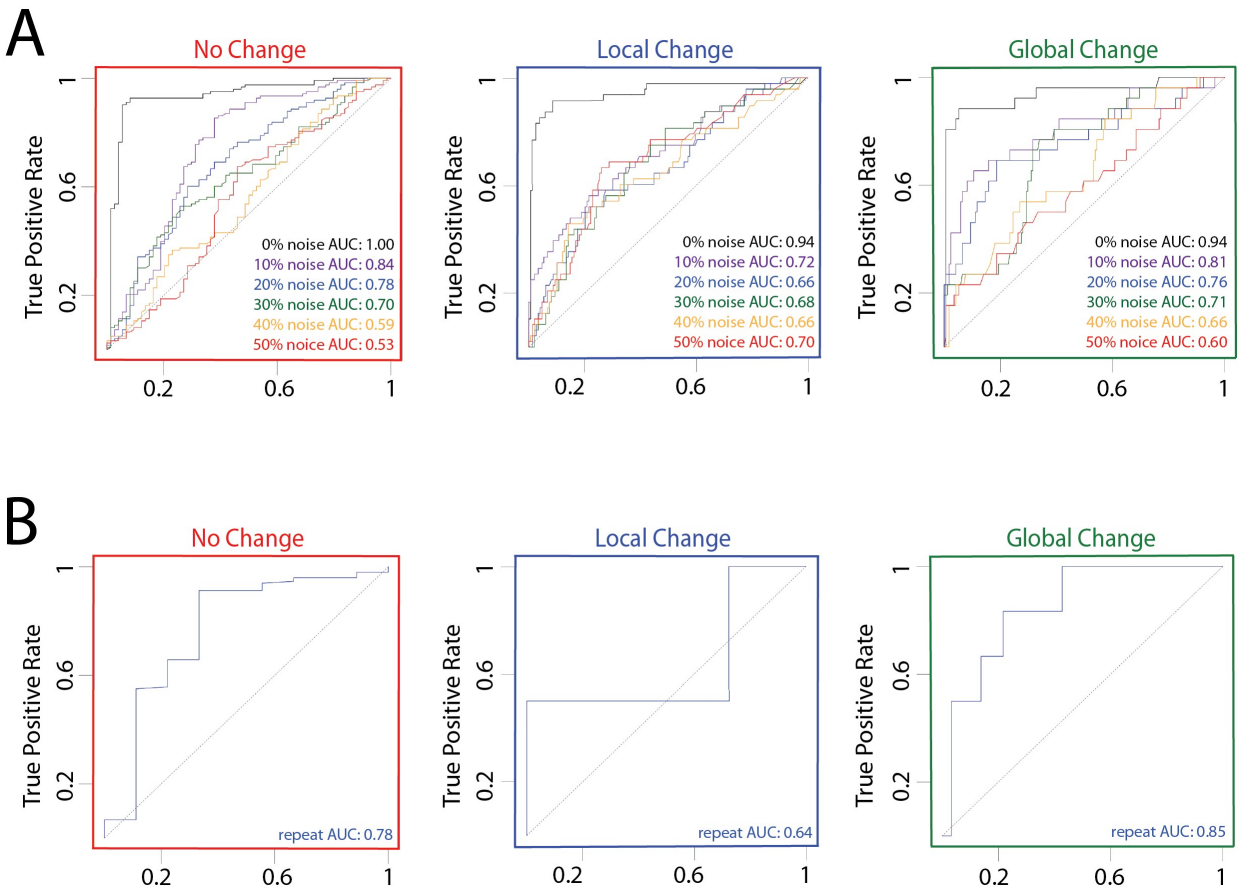


Figure 2.7 Robustness to noise

A) The model's robustness to noise was tested using randomly added simulated noise in 10% increments. Predictions on the noisy data were compared to the expert curated majority consensus. B) We also tested robustness to noise using the experimental repeats from the *F. nucleatum* glycine riboswitch by comparing the model predictions from one repeat against the other.

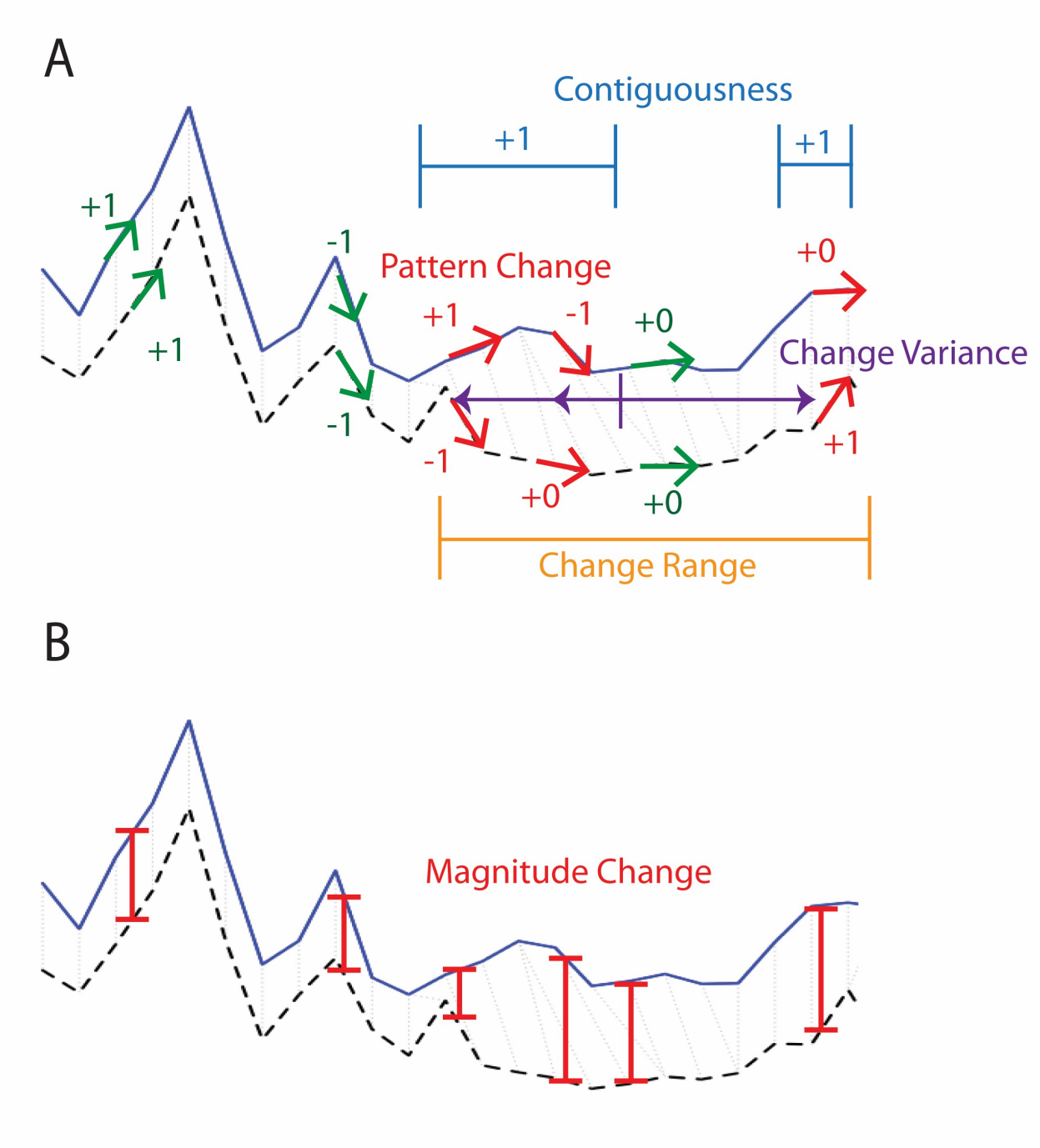


Figure 2.8 Feature descriptions

A) The sign of the slope at each nucleotide determines the trace pattern. If the slope is smaller than 0.1, the pattern at that nucleotide is 0. The trace pattern for the WT (blue) is [1, -1, 1, -1, 0, 0] and the trace pattern for the mutant (black dotted) is [1, -1, -1, 0, 0, 1]. Pattern

CC is the Pearson correlation between the WT and mutant trace patterns. The pattern is the same when the WT and mutant trace patterns are in agreement (green). A pattern change occurs when the WT and mutant trace patterns differ (red). Contiguousness sums the number of contiguous pattern change regions (light blue). Change range is the number of nucleotides between the first pattern change and the last (orange). Change variance calculates the average nucleotide of pattern change (purple vertical line). Change variance then counts the nucleotides between that mean location and each pattern change instance (purple arrows), and finds the variance among those distances. B) Pearson CC, eSDC, and the Euclidean norm are three different ways to measure the difference in magnitude (red) between the WT (blue) and mutant (black dotted). For WT/mutant pairs that are highly similar (Pearson CC greater than 0.9), the relationship between eSDC and Pearson CC is almost linear. For trace pairs from the subset of RNAs classified by experts in this range, the two features have a correlation of -0.99. However, for less similar pairs in this subset (Pearson CC less than 0.9), the correlation between eSDC and Pearson CC drops to -0.55. This is particularly acute with RNAs of longer length (greater than 150 nucleotides). As such these two features, although correlated, do behave differently and we chose to include both of them in our machine learning training.

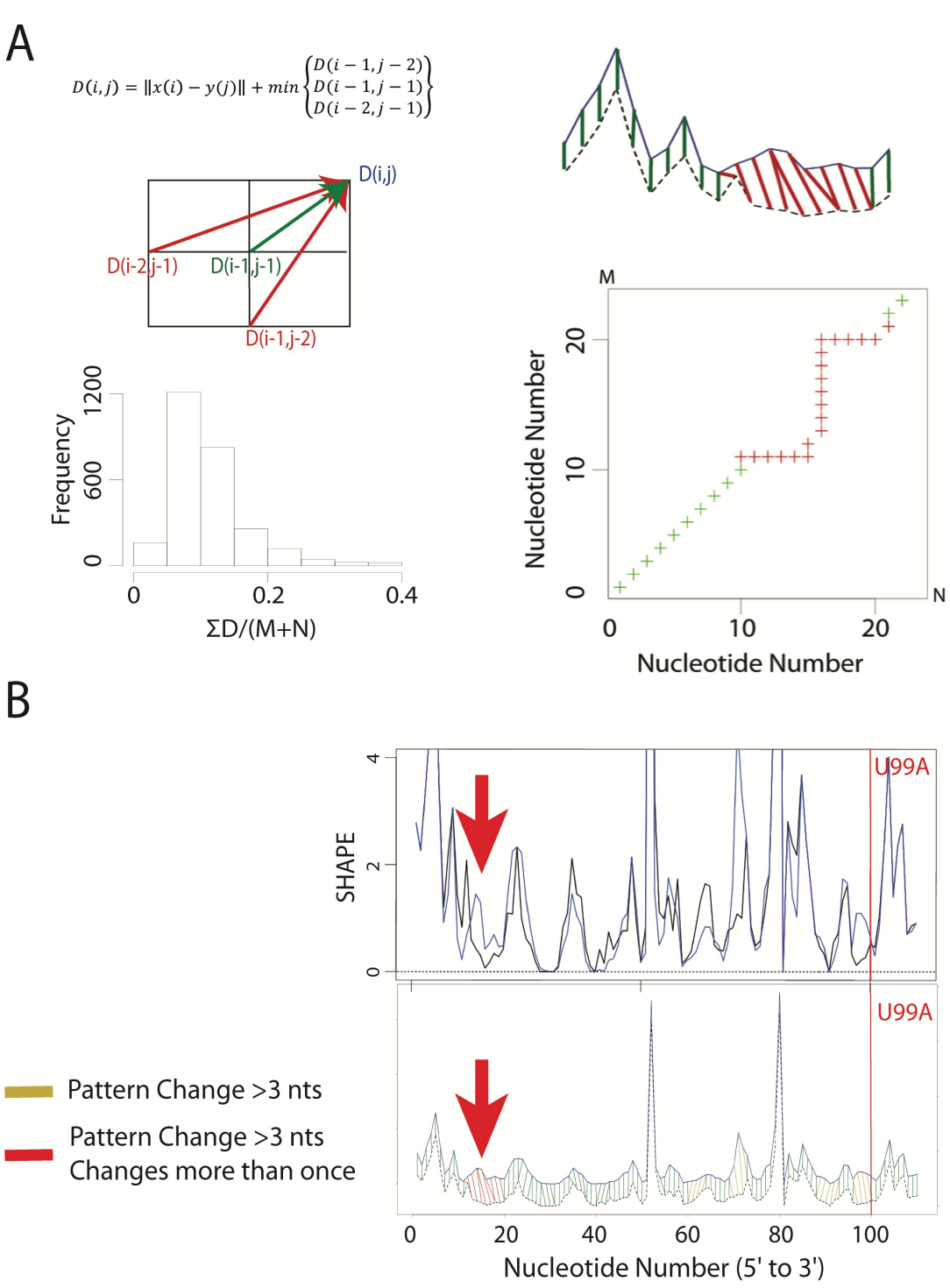


Figure 2.9 Dynamic time warping feature

A) Dynamic time warping calculates the minimum alignment between two time series (Giorgino, 2009, Sakoe and Chibe, 1978). D(i,j) represents the stepping algorithm for a

single step. Each step is constrained (middle right). A histogram of the dynamic time warping scores for the set of 2019 traces (bottom left) suggest most trace pairs have little or no difference between the WT and mutant. We aligned the WT (blue) and mutant (black/dashed) traces for an RNA fragment (top right). In the step profile (bottom right) for the RNA fragment, a diagonal step indicates a match (green), whereas a horizontal or vertical step indicates an insertion or deletion (red). Moving from left to right, each diagonal or horizontal step corresponds to the next nucleotide. Moving from bottom to top, each diagonal or vertical step corresponds to the next nucleotide. B) The SHAPE trace overlay (top right) shows the WT (black) and the U99A mutant (blue) for the 16S Four-Way junction (Cordero and Das, 2015, Tian et al., 2014, Zhang et al., 2009). The red bar highlights the U99A mutation site. On the dynamic time warping overlay (bottom right), green indicates a region with correct matches, yellow indicates a region with at least 3 shifted nucleotides and red indicates a region with at least 3 shifted nucleotides that shifts more than once. The red arrow emphasizes the region with the largest change.

A

Feature	Cumulative Accuracy	Mean Decrease Accuracy	Mean Decrease Gini
Dynamic Time Warping	0.65	0.05	10.77
Pearson CC	0.72	0.06	11.08
eSDC	0.74	0.05	10.19
Contiguousness	0.74	0.03	8.53
Pattern Change	0.75	0.07	13.19
Change Variance	0.77	0.04	11.37
Euclidean Norm	0.77	0.04	8.36
Change Range	0.80	0.05	12.47

B

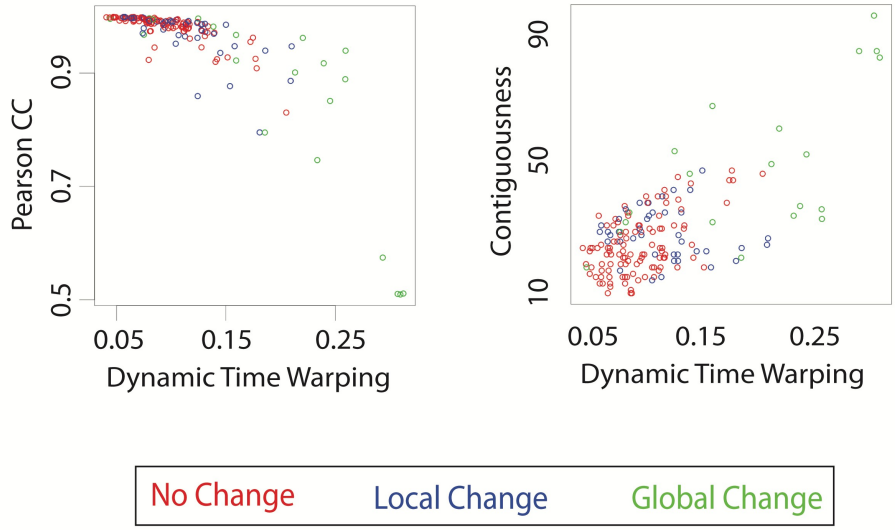


Figure 2.10 Feature selection

A) We performed recursive feature elimination on the set of 23 features from Table 2.1. Beyond the top eight features, additional features did not further contribute to the accuracy of the random forest classifier. This table lists the RNAs by their importance determined when all 23 features are included. Using only the set of eight features, we calculated the mean decrease accuracy and mean decrease Gini. B) The scatter plot depicts the separation of non-changers (red) from local changers (blue) and global changers (green) by comparing dynamic time warping versus Pearson correlation coefficient and dynamic time warping versus contiguousness.

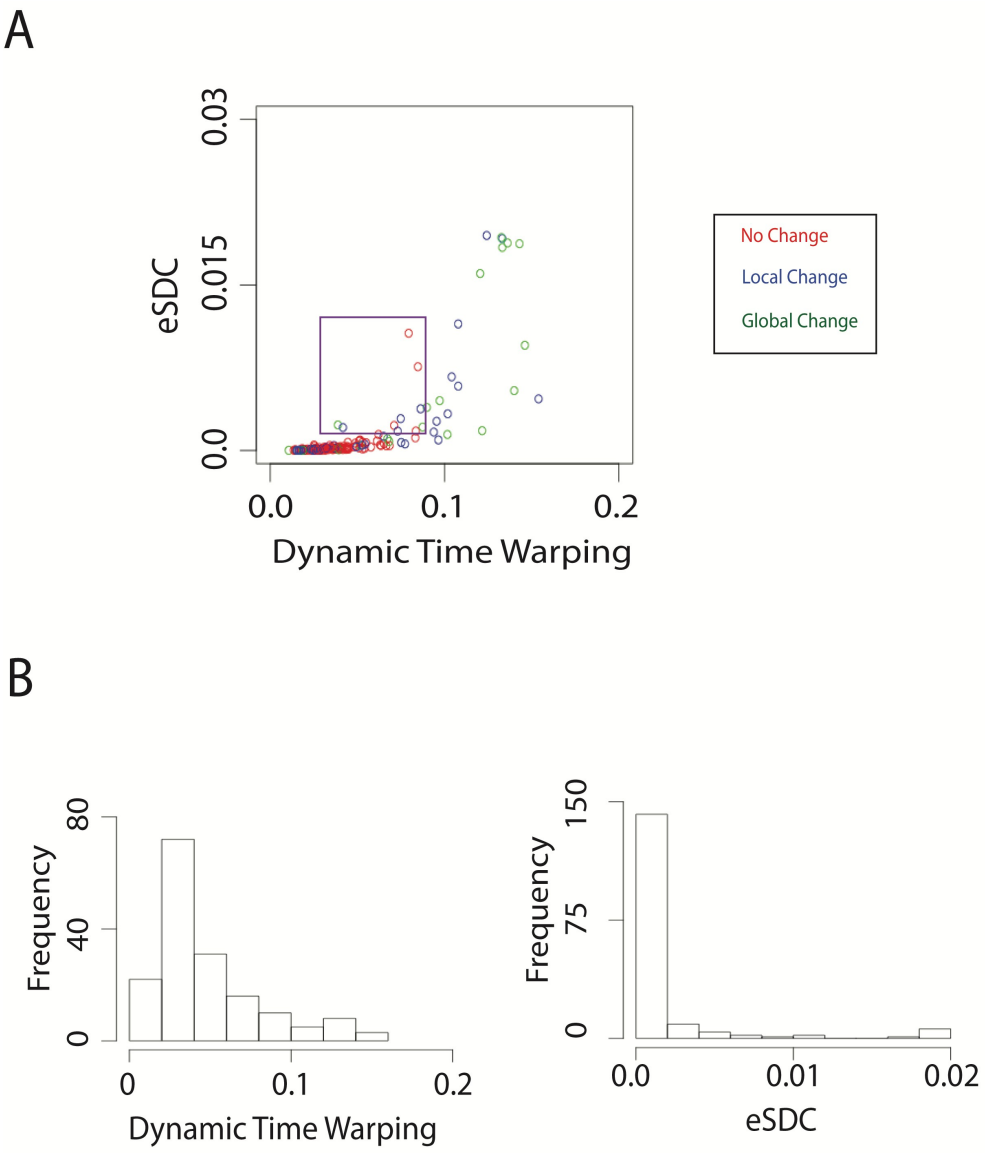


Figure 2.11 Dynamic time warping versus eSDC

A) Scatter plot showing no change (red), local change (blue) and global change (green) for the 167 expert classified RNAs. Dynamic time warping is less sensitive to shifts in the WT or mutant traces that result from misalignment. As such, there are WT/mutant trace pairs with low/mid-range dynamic time warping scores, but high eSDC values (purple box). Any trace pair values beyond the scatter plot range were truncated in order to better highlight this

region. B) Histograms for dynamic time warping (left) and eSDC (right). These histograms indicate that scores below 0.1 for dynamic time warping are low/mid-range, and values above 0.005 for eSDC are high.

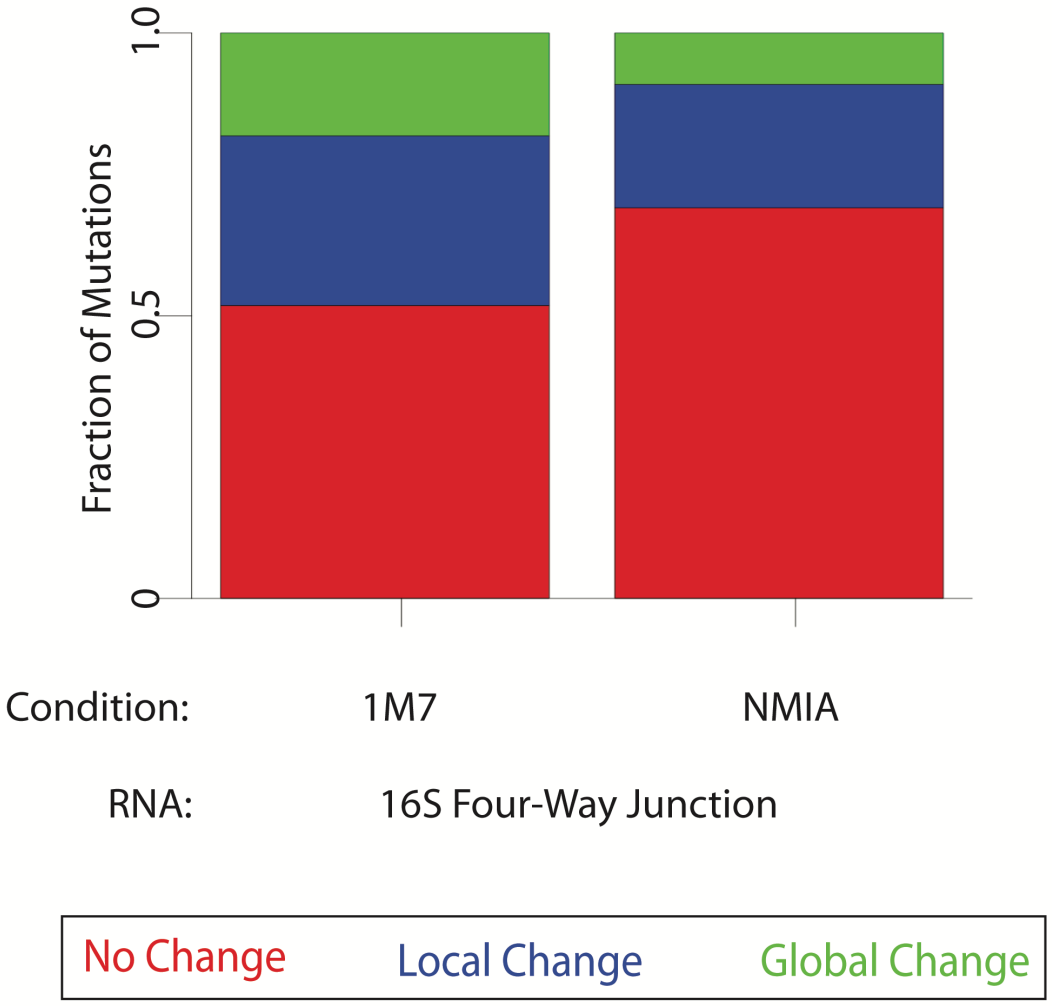


Figure 2.12 Probability of disruption for 16S Four-Way Junction

This bar plot shows the fraction of mutations that cause no change (red), local change (blue) or global change (green) determined by classSNitch. The data sets used either 1M7 or NMIA chemical modifiers on the 16S Four- Way junction (Cordero and Das, 2015). 1M7 reacts faster and detects more structural differences.

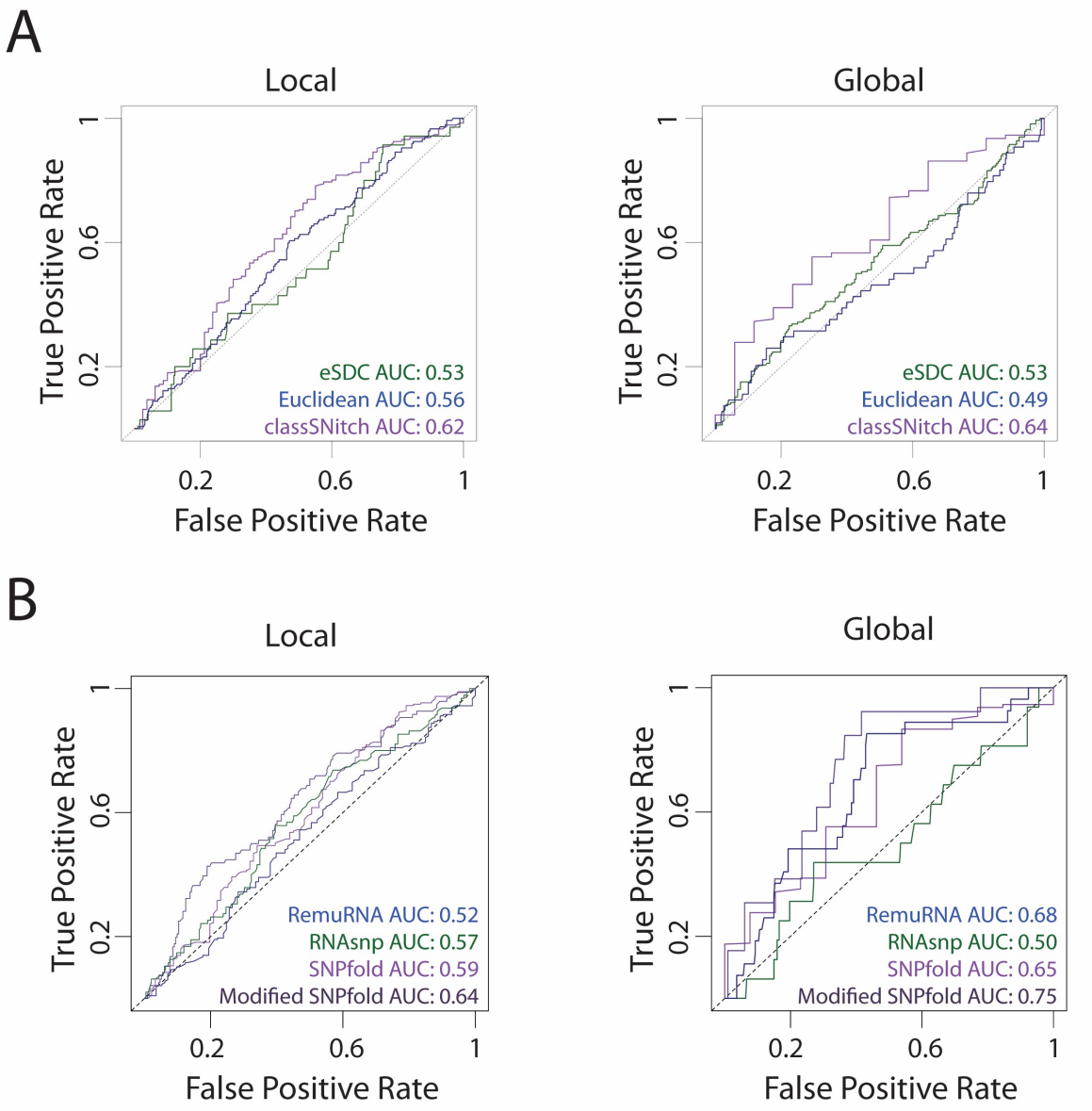


Figure 2.13 Improving the performance of structure change prediction algorithms

A) For local and global change, we performed ROC curve analysis for SNPfold, a structure change prediction algorithm, using classSNitch (purple), eSDC (green) and the Euclidean norm (blue) to classify the experimental data using the 10% tails strategy (Corley *et al.*, 2015). B) For local and global change, we compare the performance of structure change prediction algorithms on the classSNitch classification for SNPfold (purple), RNAsnp

(green) and RemuRNA (blue) (Halvorsen *et al.*, 2010; Sabarinathan *et al.*, 2013; Salari *et al.*, 2013). Each of these algorithms predicts structure change in RNA using only sequence information. We improved the SNPfold prediction (dark purple) using Eq. 8. The ROC curves for no change predictions are included in the text, Figure 2.5.

RNA Name	Organism	Length	SHAPE Traces	SHAPE Modifier	Conditions
16S rRNA four-way junction domain	*Escherichia coli*	109	110	1M7	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
16S rRNA four-way junction domain	*Escherichia coli*	109	110	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
5S rRNA	*Escherichia coli*	120	121	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Adenine-sensing *add* riboswitch aptamer	*Vibrio vulnificus*	71	72	1M7	24C, pH6.5, 10mM MgCl2, 50mM KCl, 25mM K3PO4
Adenine-sensing *add* riboswitch aptamer	*Vibrio vulnificus*	71	72	1M7	24C, pH6.5, 10mM MgCl2, 50mM KCl, 25mM K3PO4, 5mM adenine
cyclic di-GMP-1I riboswitch aptamer	*Vibrio cholerae*	103	115	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES, 10uM cyclic-diguanosine-monophosphate
Glycine riboswitch aptamer	*Fusobacterium nucleatum*	159	159	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Glycine riboswitch aptamer	*Fusobacterium nucleatum*	159	159	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Glycine riboswitch aptamer	*Fusobacterium nucleatum*	159	159	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES, 10mM glycine
Glycine riboswitch aptamer	*Fusobacterium nucleatum*	159	159	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES, 1mM glycine
Hox A9 mRNA 5'UTR	mouse E10.5–12.5	176	105	1M7	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Lariat-capping ribozyme	*Didymium iridis*	188	189	1M7	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Medloop	Synthetic	55	35	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
P4-P6 ribozyme domain	*Tetrahymena thermophila*	164	160	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES, 30pct methylpentanediol
Phenylalanine tRNA	*Sacchromyces cerevisiae*	76	77	NMIA	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Tebowned FMN aptamer	Synthetic	72	48	1M7	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES
Tebowned FMN aptamer	Synthetic	72	48	1M7	24C, pH 8.0, 10mM MgCl2, 50mM Na-HEPES, 2mM FMN

Table 2.3 RNAs for use in analysis

We collected the set of 2019 traces from the mutate-and-map experiments in the RMDB database (Cordero *et al*. 2010, Kladwang *et al*. 2011, Kladwang *et al.* 2011). This table reports basic information for each of the 17 RNAs in this data set, as well as the experimental conditions, the SHAPE modifier and the number of available traces.

Feature	Formula	Description
Abs Diff at Mutation	$abs(SHAPE_{ref}[i_{mut}]-SHAPE_{alt}[i_{mut}])$	Absolute difference in SHAPE values at the mutation site between wild type and mutant traces
Change Points (binary)	change point algorithm	Change point analysis algorithm to detect if there is a location of change outside of the mutation site between wild type and mutant traces (Killick and Eckley, 2014)
Diff around Mutation	$SHAPE_{ref}[mut]-SHAPE_{alt}[mut]$	Difference in SHAPE values at the mutation site for wild type and mutant traces
Flatness	$(N-\Sigma_i(\#\ of\ Change_{alt}))/N$	Number of locations where the mutant trace pattern does not change
Length	N	Length of mutant trace
Max Abs Diff	$max(SHAPE_{ref}[i]-SHAPE_{alt}[i])$	Maximum absolute difference between wild type and mutant traces
Max Peak	$max(SHAPE_{alt}[i])$	Maximum SHAPE value for mutant trace
Mean Change	$\Sigma_i(SHAPE_{ref}[i]-SHAPE_{alt}[i])/N$	Average difference between wild type and mutant traces
Mean Change Distance	$\Sigma_{i\neq i_{diff}}(i_{mut}-i_{diff})/N$	Average distance from mutation site for change between wild type and mutant trace pattern
Mean Derivative	$\Sigma_i(d(SHAPE_{ref}[i]-SHAPE_{alt}[i]))/N$	Slope of difference between wild type and mutant traces
Mean Diff at Mutation	$\Sigma_i(SHAPE_{ref}[i]-SHAPE_{alt}[i])/5$	Average SHAPE values 5 nucleotides around mutation for mutant trace
Median SHAPE	$median(SHAPE_{alt}[i])$	Median SHAPE value for mutant trace
Mutation Change (binary)	change point algorithm	Change point analysis algorithm to detect if there is a location of change at the mutation site between the wild type and mutant traces (Killick and Eckley, 2014)
Mutation Raw	$SHAPE_{alt}[i_{mut}]$	Raw SHAPE value at the mutation site
Peak ratio	$(\#SHAPE_{high}[i])/(\#SHAPE_{low}[i])$	Ratio of "high" SHAPE values (above 1.5) to "low" SHAPE values (below 1.5) for mutant trace

Table 2.4 Formula and descriptions for features describing SHAPE trace pairs

The formula symbol descriptions are included in Table 2.7. After random feature elimination, these features were not included in the final model (Methods Supplementary, Subsection 2.5.2.4).

	No Change				Local Change				Global Change			
	Correct	PPV	TPR	FPR	Correct	PPV	TPR	FPR	Correct	PPV	TPR	FPR
RandomForest	138	0.71	0.90	0.10	138	0.94	0.48	0.53	155	0.98	0.55	0.45
KStar	138	0.76	0.87	0.13	136	0.88	0.63	0.38	153	0.96	0.6	0.4
MultilayerPerceptron	131	0.72	0.83	0.17	120	0.88	0.23	0.78	156	0.99	0.55	0.45
LogitBoost	129	0.61	0.87	0.13	128	0.92	0.30	0.70	154	0.99	0.45	0.55
AdaBoostM1	128	0.46	0.93	0.07	122	0.92	0.15	0.85	155	0.99	0.45	0.55
IBk	128	0.68	0.82	0.18	126	0.84	0.50	0.50	151	0.95	0.55	0.45
FilteredClassifier	126	0.39	0.96	0.04	126	0.99	0.00	1.00	158	1.00	0.55	0.45
J48	126	0.45	0.93	0.07	124	0.88	0.33	0.68	153	0.97	0.55	0.45
PART	126	0.51	0.90	0.10	126	0.89	0.35	0.65	152	0.96	0.55	0.45
REPTree	125	0.57	0.85	0.15	130	0.95	0.23	0.78	152	0.99	0.35	0.65
LMT	123	0.60	0.82	0.18	125	0.94	0.15	0.85	154	0.98	0.5	0.5
RandomSubSpace	123	0.49	0.88	0.12	126	0.96	0.10	0.90	155	0.99	0.5	0.5
ClassificationViaRegression	122	0.45	0.90	0.10	129	0.96	0.18	0.83	156	0.99	0.5	0.5
DecisionTable	122	0.48	0.88	0.12	121	0.93	0.10	0.90	155	0.99	0.5	0.5
Bagging	119	0.60	0.79	0.21	130	0.92	0.35	0.65	156	0.99	0.5	0.5
BayesNet	119	0.51	0.83	0.17	126	0.99	0.00	1.00	155	0.97	0.6	0.4
HoeffdingTree	119	0.39	0.90	0.10	127	1.00	0.00	1.00	150	0.94	0.55	0.45
NaiveBayesUpdateable	119	0.39	0.90	0.10	90	0.47	0.78	0.23	153	0.94	0.7	0.3
RandomTree	119	0.58	0.79	0.21	125	0.87	0.38	0.63	144	0.89	0.65	0.35
NaiveBayes	118	0.37	0.90	0.10	91	0.47	0.80	0.20	153	0.94	0.7	0.3
AttributeSelectedClassifier	116	0.50	0.81	0.19	125	0.91	0.25	0.75	155	0.99	0.45	0.55
SMO	116	0.21	0.97	0.03	127	1.00	0.00	1.00	153	1.00	0.3	0.7
SimpleLogistic	115	0.27	0.93	0.07	125	0.99	0.00	1.00	154	0.98	0.5	0.5
OneR	114	0.55	0.77	0.23	121	0.86	0.30	0.70	157	0.99	0.55	0.45
MultiClassClassifier	113	0.39	0.84	0.16	124	0.96	0.05	0.95	151	0.97	0.45	0.55
JRip	112	0.54	0.75	0.25	129	0.92	0.30	0.70	153	0.97	0.5	0.5
CVParameterSelection	107	0.00	1.00	0.00	127	1.00	0.00	1.00	147	1.00	0	1
MultiScheme	107	0.00	1.00	0.00	127	1.00	0.00	1.00	147	1.00	0	1
NaiveBayesMultinomialText	107	0.00	1.00	0.00	127	1.00	0.00	1.00	147	1.00	0	1
Stacking	107	0.00	1.00	0.00	127	1.00	0.00	1.00	147	1.00	0	1
Vote	107	0.00	1.00	0.00	127	1.00	0.00	1.00	147	1.00	0	1
LWL	105	0.59	0.66	0.34	124	0.97	0.03	0.98	156	0.99	0.55	0.45
DecisionStump	103	0.57	0.66	0.34	125	0.97	0.05	0.95	154	0.97	0.55	0.45
MultiClassClassifierUpdateable	71	0.94	0.13	0.87	127	1.00	0.00	1.00	20	0.00	1	0
ZeroR	22	0.00	1.00	0.00	25	1.00	0.00	1.00	31	1.00	0	1

Table 2.5 Algorithm selection

Using 5-fold cross-validation, we compared the performance of 35 different algorithms on predicting no change, local change and global change using default settings in the Weka suite (Hall et al., 2009). Reported in this table are the number of correct predictions (Correct), the positive predictive value (PPV), the true positive rate (TPR), and the false positive rate (FPR). Random forest, KStar and Multilayer Perceptron parameters were further optimized. Details of this process can be found in Methods Supplementary, Subsection 2.5.2.4.

		Mutant Nucleotide			
		A	C	G	U
Wild type Nucleotide	A	-	-	-	516/97/25 (81%/15%/4%)
	C	-	-	279/91/28 (70%/23%/7%)	-
	G	-	365/123/39 (69%/23%/8%)	-	-
	U	366/71/19 (80%/16%/4%)	-	-	-

Table 2.6 Breakdown of mutations for the mutate-and-map data set

The nucleotide change from WT to mutant for non- changers (red), local changers (blue), and global changers (green) determined by classSNitch for the set of 2019 mutants. The table includes the raw count (above), and the percentage that change structure (below). Mutate-and-map experiments select for transversion mutations that are more likely to change structure.

Formula Symbols	Definitions
N	Length of RNA
^{P}CC	Pearson correlation coefficient
i	Nucleotide position
i_{mut}	Mutation site
i_{diff}	Nucleotide position where there is a difference in trace patterns between WT and mutant
$i_{contiguous}$	Starting nucleotide position of stretches of contiguous change
$SHAPE_{WT}$	SHAPE value for WT
$SHAPE_{mut}$	SHAPE value for mutant
$SHAPE_{high}$	SHAPE value above 1.5
$SHAPE_{low}$	SHAPE value below 1.5
X_{WT-mut}	SHAPE value difference between WT and mutant
μ_{WT-mut}	Average SHAPE value difference between WT and mutant
$Change_{WT}$	Trace pattern for WT
$Change_{mut}$	Trace pattern for mutant
trace pattern	Increase (+1), decrease (-1) or no change (0) in SHAPE value moving from one nucleotide to the next
Q1	First interquartile range
Q3	Third interquartile range
SD	Standard deviation

Table 2.7 Formula symbols

Descriptions of formula symbols used in Tables 2.2 and 2.3.

	Validation Traces		
RNA Name	None	Local	Global
16S rRNA four-way junction domain	20	4	3
16S rRNA four-way junction domain	0	0	0
5S rRNA	30	0	3
Glycine riboswitch aptamer	16	2	0
Glycine riboswitch aptamer	6	1	1
Glycine riboswitch aptamer	0	1	2
Glycine riboswitch aptamer	17	0	1
Adenine-sensing *add* riboswitch aptamer	0	8	2
Adenine-sensing *add* riboswitch aptamer	0	0	2
cyclic di-GMP-1I riboswitch aptamer	0	2	0
Hox A9 mRNA 5'UTR	0	0	4
Phenylalanine tRNA	7	1	0
Lariat-capping ribozyme	0	8	2
Medloop	0	0	0
P4-P6 ribozyme domain	9	10	0
Tebowned FMN aptamer	2	3	0
Tebowned FMN aptamer	0	0	0

Table 2.8 Validation Traces

The set of 200 validation traces were chosen from the 17 data sets. This table lists the majority consensus classification for the subset of 167 RNAs. The uneven distribution of global, local and non-changers among the data sets is a result of some RNAs being more sensitive to change than others, global changers being a fairly rare occurrence and experts being unable to reach a consensus on 33 of the original 200 RNA traces that were classified. Five folds for cross-validation were created from these traces and stratified by RNA, with some sets containing multiple RNAs. Training sets are denoted by dotted lines.

Feature	Min	Q1	Median	Q3	Max	Mean	SD
Abs Diff at Mutation	0.000	0.014	0.088	0.315	15.09	0.312	0.706
Change Points (binary)	0	0	0	0	1	0.136	0.343
Change Range*	22	107	149	162	194	134.1	36.58
Change Variance*	0.055	0.202	0.273	0.370	1.276	0.306	0.150
Contiguousness*	5	23	30	38	103	31.24	12.15
Diff around Mutation	-7.271	-0.090	0.016	0.115	15.09	-0.001	0.772
Dynamic time warping*	0.038	0.083	0.113	0.152	0.567	0.130	0.074
eSDC*	0.010	0.073	0.075	0.082	0.136	0.080	0.015
Flatness	0.000	0.191	0.247	0.302	0.462	0.246	0.078
Euclidean Norm*	0.702	3.174	4.926	7.453	35.28	6.082	4.422
Length	52	115	178	179	199	154.5	38.13
Pearson CC*	0.105	0.961	0.986	0.995	1.000	0.962	0.070
Max Abs Diff	0.334	1.403	2.277	3.820	27.19	3.160	2.849
Max Peak	4.294	9.747	17.44	22.56	59.50	19.71	13.97
Mean Change	0.261	0.915	1.280	1.618	20.21	1.362	0.877
Mean Change Distance	0.952	28.06	61.01	101.0	188.0	67.31	45.36
Mean Derivative	-0.053	0.000	0.000	0.000	0.079	0.001	0.006
Mean Diff at Mutation	-3.820	-0.075	0.000	0.099	3.987	0.016	0.386
Median SHAPE	0.002	0.379	0.523	0.705	1.469	0.572	0.273
Mutation Change	0	0	1	1	1	0.536	0.499
Mutation Raw	-0.076	0.193	0.551	1.349	59.50	1.288	2.695
Pattern CC*	0.040	0.741	0.806	0.858	0.962	0.781	0.118
Peak ratio	0.017	0.265	0.342	0.407	0.926	0.351	0.137

Table 2.9 Feature Statistics

We calculated the statistics for each of the features from the complete set of 2019 RNAs. The eight features included in the model are denoted by an asterisk. Feature descriptions are found in Tables 2.2 and 2.3. Feature symbols are described in Table 2.7.

		Expert		
		No Change	Local Change	Global Change
Prediction	No Change	91	14	9
	Local Change	15	25	5
	Global Change	1	1	6

Table 2.10 Prediction/Expert confusion matrix

For each of the validation traces, we used the SHAPE profile to guide prediction of the minimum free energy secondary structure (Diegan *et al.*, 2009). We compared the predicted structure change to the expert classification. The predicted structures were filtered to remove base pairing regions with fewer than 3 base pairs. We classified the predicted secondary structures into local (differences in base pairing stems that encompass the mutation and in stems that are immediately adjacent to the encompassing stem), global (differences in more distant stems) and non-changers (same number and order of stems regardless of stem length). The experts occasionally classify local changers in secondary structure as global changers. However, the experts rarely classify global changers as local changers.

CHAPTER 3: VISUALIZATION OF THE RNA SUBOPTIMAL ENSEMBLE[2]

3.1 Introduction

RiboNucleic Acid (RNA) 3-dimensional structures are the result of remarkably complex interaction networks that together create emergent biological functions (Chauhan and Woodson, 2008; Mitra *et al.*, 2011; Shcherbakova *et al.*, 2008; Sinan *et al.*, 2011). Although crystal structures reveal these networks with atomic detail, these remain static snapshot models of the conformations existing in the cellular environment (Noller, 2005). RNAs, particularly highly structured RNAs such as ribosomal RNA, exist in multiple conformations, many of which are likely to affect their function(s) (Kutchko *et al.*, 2015; Ritz *et al.*, 2013; Sclavi *et al.*, 2005). Thus, when describing RNA structure, it is more accurate to discuss an ensemble of conformations instead of a single structure (Eddy, 2009; Ponty, 2008; Ritz *et al.*, 2013; Thirumalai and Hyeon, 2005). However, significant biophysical challenges remain, whether at the secondary or tertiary structural level, including visualization of the ensemble of RNA conformations and identification of essential functional elements within the entire ensemble (Das *et al.*, 2003; Martin *et al.*, 2012; Shapiro *et al.*, 2001; Thirumalai and Hyeon, 2005).

[2] This work has been submitted as an original manuscript to the Biophysical Journal and is currently in review.

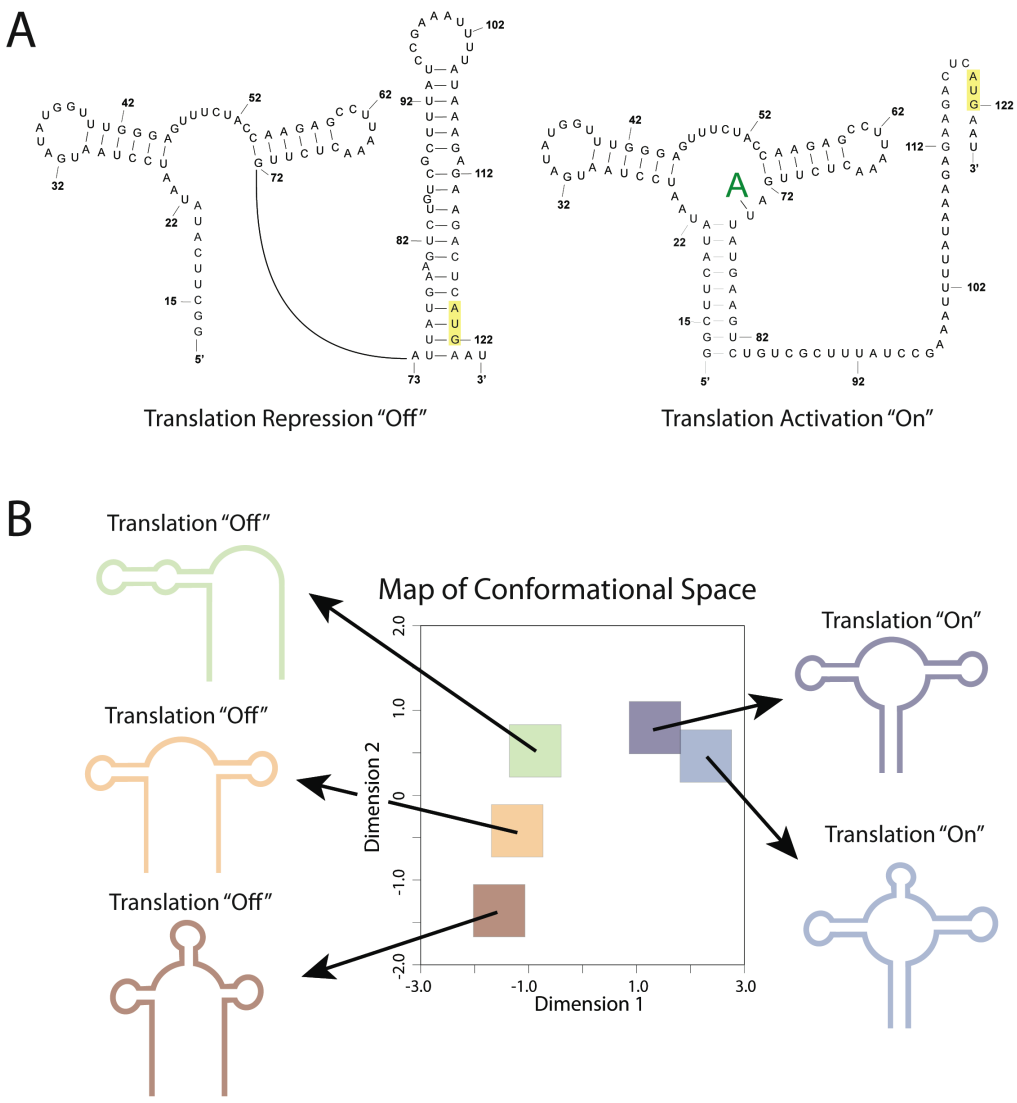


Figure 3.1. The conformational states of the *vibrio vulnificus add* adenine riboswitch.

A) The accepted structures for the bound and unbound states of the riboswitch determined by crystallography and NMR (Serganov *et al.*, 2015). The unbound state represses translation, and the bound state activates translation (Lemay *et al.*, 2009; Serganov *et al.*, 2015). B) The map of conformational space explores five possible structure clusters for the riboswitch. The representative arc diagram is the cluster medoid structure. The orange cluster represents the translation "Off" conformation, and the purple cluster represents the translation "On" conformation, as confirmed by crystallography and NMR (Liu *et al.*, 2015).

The challenge of visualizing an RNA secondary structure ensemble is easily illustrated by the *vibrio vunificus* adenosine deaminase (*add*) adenine riboswitch (Figure 3.1) (Cordero and Das, 2015; Delfosse *et al.*, 2010; Lemay and Lafontaine, 2007; Lemay *et al.*, 2006). Typically RNA is represented as a single structure, but, for the riboswitch, at least two structures are required for function: the "On" and "Off" conformations (Figure 3.1A) (Delfosse *et al.*, 2010; Lemay *et al.*, 2011; Lemay *et al.*, 2006). These two structures interchange, with the "Off" conformation favored without the adenine ligand, and the "On" conformation stabilized by binding adenine (Lemay and Lafontaine, 2007; Lemay *et al.*, 2009; Lemay *et al.*, 2006). Thus, in solution the RNA exists as an ensemble of conformations that interchange (Bokinsky and Zhuang, 2005; Chauhan and Woodson, 2008; Halvorsen *et al.*, 2010; Kutchko *et al.*, 2015; Ponty, 2008; Roh *et al.*, 2010). In visualizing such an ensemble, two salient aspects should be highlighted to understand function: 1) the structural similarity and difference between the two conformations, and 2) the relative abundance of each conformation in the ensemble.

Defining structural similarity requires a representation that captures biologically important structural features of the RNA to facilitate clustering of highly similar conformations. From these clusters, it is then possible to determine the relative abundance of the conformations, which reflects their relative thermodynamic weights in the Boltzmann ensemble. We therefore aim to create a visualization based on a sampling of conformational space like the one illustrated for the *add* riboswitch (Figure 3.1B), which was suboptimally sampled from the Boltzmann ensemble. In Figure 3.1B, we illustrate a map of conformational space, in which each square represents a cluster of similar conformations based on a "nested feature vector" which we define below. This representation is particularly interesting as it reveals several aspects of the *add* riboswitch conformational ensemble that are not apparent when considering only two structures

(Figure 3.1A). First, this visualization suggests that there are more than two classes of conformations in the *add* riboswitch conformational ensemble. Second, the "On" and "Off" conformational change is conveniently captured along dimension 1. The methods we describe below provide a robust approach for identifying specific dimensions that capture biologically informative structural differences, such as those in Figure 3.1B.

In Figure 3.1B, we purposely did not indicate the relative abundance of conformations in each conformational cluster; each square is equal in size. The relative weight of these clusters depends on the underlying thermodynamic parameters of the energy model. Given a nearest-neighbor energy model, it is now computationally efficient to rapidly sample the Boltzmann suboptimal ensemble (Ding *et al.*, 2004; Ding *et al.*, 2005; Hamada *et al.*, 2009; Waldispuhl and Clote, 2007). Furthermore, the nearest neighbor model can be extended to empirically include experimental structure probing data, particularly Selective 2' Hydroxyl Acylation by Primer Extension (SHAPE) data (Deigan *et al.*, 2009; Wilkinson *et al.*, 2009). Inclusion of SHAPE data is relevant because the RNA structure is readily probed under different experimental conditions. For example, the *add* riboswitch can be probed with and without the ligand that causes a structural rearrangement (Cheng *et al.*, 2015; Cordero and Das, 2015; Cordero *et al.*, 2012). As we will show below, the visualization proposed in Figure 3.1B accurately captures this biologically important rearrangement when combined with SHAPE-informed structure probing.

Although visualizing riboswitch ensemble conformations is one important goal of our work, the main motivation for improving the ability to visualize and interpret RNA conformational ensembles stems from our studies of messenger RNA (mRNA) folding *in vitro* vs. *in vivo*. Quantitative comparison of these two conditions effectively enables us to deconvolute the effect of the cellular environment on mRNA folding. The structural ensembles

of these highly regulated RNAs tend to be far more complex than the structural ensembles of riboswitches. As such, we require tools that enable "sorting the forest from the trees" to understand these large and complex molecules. We present here an experimental high-resolution comparison of SHAPE data for the human *β-actin* mRNA that reveals specific regions in which the RNA folds differently *in vitro* vs. *in vivo*. We show how these visualizations enable interpretation of the complex rearrangements of the mRNA conformational ensemble that occur in the cell, thereby obtaining meaningful biophysical and biological insight into the specific structure function relationships of the specific messenger. Together, these novel data and methods establish a robust approach for interpreting chemical and enzymatic probing data in the context of conformational ensembles.

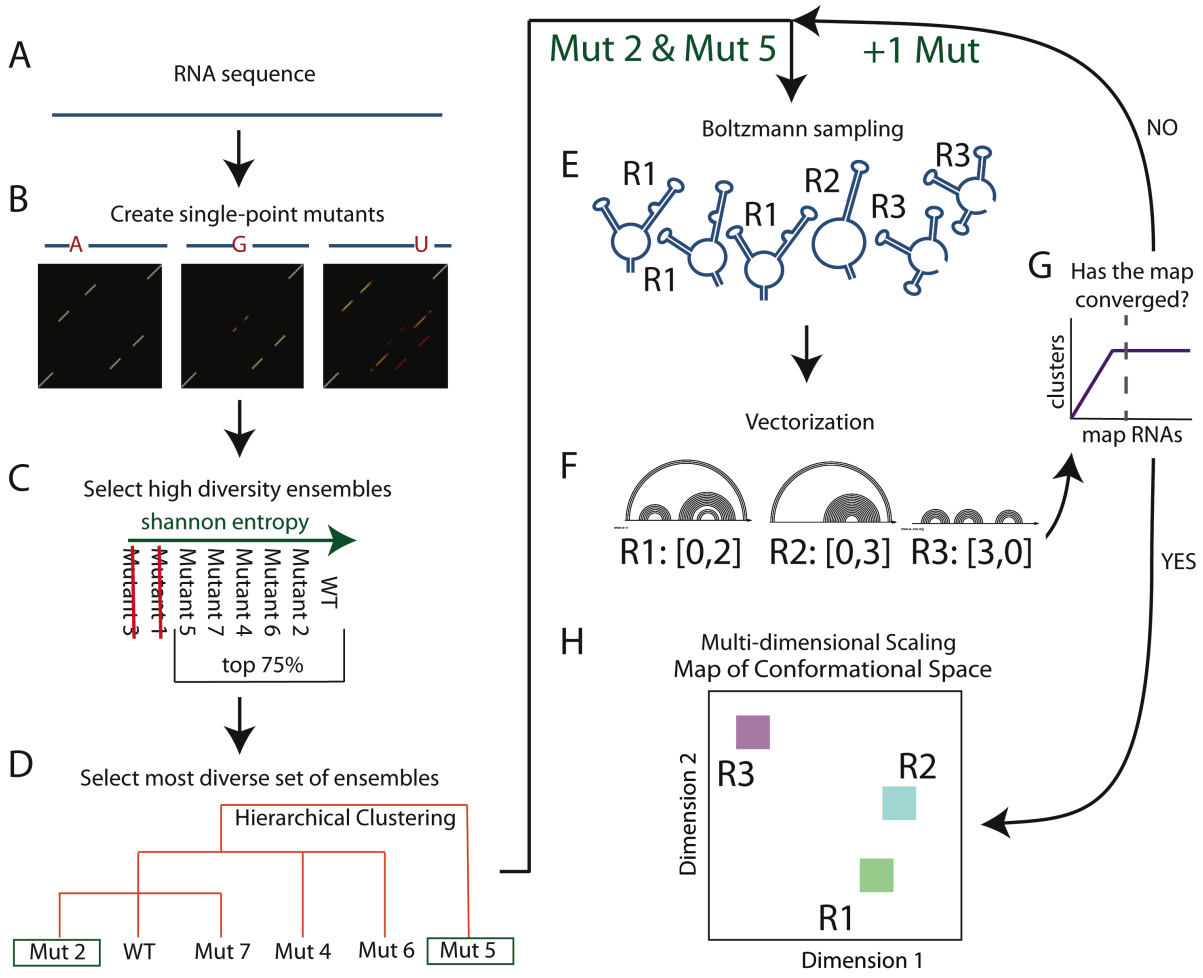


Figure 3.2. Building the map of conformational space.

The map explores the possible structural space for an RNA sequence and its single point mutants. A) A single point mutant was created for every position in the RNA. We used only mutations that were expected to lead to the largest changes in structure based on experimental observations from the mutate-and-map experiments (AtoU, UtoA, CtoG, GtoC) (Kladwang *et al.*, 2011; Kladwang *et al.*, 2011). B) The partition function was generated for the wild type and

single point mutants using established structure prediction methods (Halvorsen *et al.*, 2010; Matthews *et al.*, 1999; Matthews *et al.*, 2004; Matthews and Turner, 2002). C) The RNAs were ranked by Shannon entropy, and the top 75% were retained to filter for individual RNAs with more diverse ensembles (Shannon, 1951). D) We collapsed the partition function for each of the remaining RNAs into their base pairing probabilities, and performed hierarchical clustering on the probabilities (Defays, 1977). This clustering selects the most diverse RNA subsets. E) We selected the most distant RNA and sampled 1000 structures according to their Boltzmann probability (5). F) We used data abstraction to identify the number of unique structure clusters. This data abstraction is further described in Fig. S1 in the Supporting Material. G) We repeated steps E and F until the number of cluster structures converged. H) The structure clusters are projected into 2-dimensional space using metric multidimensional scaling (MDS). By minimizing the stress function for the Euclidean distance matrix, MDS optimizes the positioning of the structure clusters (Abdi, 2007; Torgerson, 1952).

3.2.1 Generating structures for the map of conformational space

Our strategy for establishing a conformational map of an RNA ensemble is illustrated in Figure 3.2. Beginning with the RNA sequence (Figure 3.2A), we compute its partition function (probability of base-pairing (Bernhart *et al.*, 2006; Markham and Zuker, 2008; Mathews, 2004; McCaskill, 1990; Waldispuhl and Clote, 2007)) and the partition functions of all AtoU, UtoA, CtoG, and GtoC single point mutant sequences (Figure 3.2B). These point mutations are experimentally determined to be maximally disruptive of structure (Kladwang *et al.*, 2011). The sum over the rows in the partition function is the base-pairing probability, P, at each nucleotide, x_i (Eq. 15). Our goal is to generate an ensemble of diverse possible conformations and establish a representative 2-dimensional map for visualization. Thus, single point mutants with the highest ensemble Shannon entropy, as defined by Eq. 15, are selected for further analysis. In the first pass, we eliminate the lowest 25% Shannon entropy mutants (Figure 3.2C) (Shannon, 1951). In a second filter, we perform hierarchical clustering of the base-pairing probability $P(x_i)$ vectors based on their Euclidean distance (Defays, 1977) to identify the most divergent partition functions (Figure 3.2D). We then perform Boltzmann suboptimal sampling on the two most divergent partition functions (Figure 3.2E), and create nestedness feature vectors from the sampled structures (Figure 3.2F, and Figure 3.6), to generate a map of conformational space using metric multidimensional scaling (Abdi, 2007; Torgerson, 1952) (Figure 3.2H). We iteratively add additional Boltzmann ensemble samples of divergent single point mutant sequences until the map of conformational space converges (Figure 3.2G).

$$Entropy = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \quad (15)$$

3.2.2 Projection of the map of conformational space

Our projection is based on the representation proposed in the RNAshapes abstraction that captures whether a stem or stack element exists, ignoring the size of that element (Steffen *et al.*, 2006). Biologically, significant variation is observed in stem length but stack elements are generally more conserved (Macke *et al.*, 2001; Rivas and Eddy, 2001; Shapiro, 1988). Thus, we expect that basing our projections on this distance metric will capture important structure/function features in the ensembles. Our representation counts the number of inner loops and stacks and then positions that count according to the location of the outermost stack in the nestedness feature vector (Figure 3.6). Stems and stacks with fewer than three base pairs are ignored. We determine the nestedness representation for every structure in the map of conformational space and collapse the structures into clusters based on unique nestedness representations (Figure 3.2F). Metric multidimensional scaling (MDS) projects the structure clusters into 2-dimensional space (Figure 3.2H) (Abdi, 2007; Torgerson, 1952). MDS calculates the Euclidean distance matrix for n-dimensional data, d_{ij}. Initial positions for the data points, x, are set in 2-dimensional space, i and j. Based on this configuration, MDS evaluates the stress function in Eq. 16 (Abdi, 2007; Torgerson, 1952). The data points are reconfigured in the direction of steepest descent. This process is repeated to minimize the stress function (Abdi, 2007; Torgerson, 1952). Minimization of the stress function finds the configuration with the smallest residual sum of squares when compared with the original distance matrix (Abdi, 2007; Torgerson, 1952). As a result, MDS yields a 2-dimensional embedding of the data points (used for visualization) which optimally reflects the pairwise distances between data points as computed within the original n-dimensional data.

$$Stress = \sqrt{\sum_{ij} \frac{(d_{ij} - \|x_i - x_j\|)^2}{\sum d_{ij}^2}} \quad (16)$$

3.2.3. EnsembleRNA package

A python package, EnsembleRNA, was created for the visualization of RNA structural ensembles. The package produces bubble charts for the map of conformational space and the wild type RNA, and allows for comparison between structural ensembles. The package is available at ribosnitch.bio.unc.edu/software. Documentation for EnsembleRNA can be found at http://ribosnitch.bio.unc.edu/software. The documentation contains additional information on usage, troubleshooting and tutorials.

3.2.4. *In vitro* SHAPE treatment

SHAPE-MaP experiments were performed *in vitro* (Siegfried *et al.*, 2014). We obtained a clone of *β-actin* mRNA (Origene - SC319328) and directly PCR-amplified the coding sequence with a 5' primer containing the T7 promoter for in vitro transcription (Q5® Site-Directed Mutagenesis Kit and T7 RNA Polymerase from NEB). To remove DNA following transcription, we treated the reaction with TURBO™ DNase for 15 minutes at 37º C (ThermoFisher Scientific). Standard bead clean up was performed between each step (Beckman Coulter - Ampure XP). The transcribed RNA was folded at 37º C in buffer containing 100 mM Na-HEPES, pH 8.0, 100 mM NaCl, and 10 mM $MgCl_2$. One ug of RNA was treated for five minutes with either 10% dimethyl sulfoxide (DMSO) or DMSO containing the RNA modifying agent 1-methyl-7-nitroisatoic anhydride (1M7) at a final concentration of 10 mM.

3.2.5. *In vivo* SHAPE treatment

We performed *in vivo* SHAPE-MaP experiments for *β-actin* in the 1000 Genome cells lines GM07037 and GM12003 (Consortium *et al.*, 2015), obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. Approximately 50 million cells were collected by centrifugation, resuspended in 1 mL of folding buffer (as in *in vitro* SHAPE protocol) supplemented with 400 U murine RNAse inhibitor, and sonicated three times at 10% power for 10 seconds (Fisher Scientific Sonic Dismembrator Model 500). These samples were incubated at 37º C for ten minutes, after which either DMSO (10% final concentration) or 500 mM 1M7 in DMSO (final concentration 30 mM) was added for five minutes with three separate additions. RNA was isolated with Trizol reagent (ThermoFisher Scientific), followed by treatment with TURBO™ DNase and removal of the majority of ribosomal RNA (RiboMinus™ Eukaryote System v2 from Life Technologies).

3.2.6. SHAPE data collection and analysis

For all samples, we performed reverse transcription with the specialized reverse transcription conditions for SHAPE-MaP and random nonamer primers (Siegfried *et al.*, 2014). The transcription reactions were purified via Ampure XP beads or G50 columns and dsDNA was made by second strand synthesis (Ampure XP beads from Beckman Coulter, G50 columns from GE Healthcare Life Sciences, NEBNext® mRNA Second Strand Synthesis Module from NEB). To prepare libraries we used the Nextera or Nextera XT kits (Nextera® DNA Sample Preparation Kit, Nextera® XT DNA Sample Preparation Kit and Index Kits from Illumina). Sequencing for the *in vitro* samples was performed on HiSeq2500 as paired end, 50-read multiplex runs. Sequencing for the *in vivo* samples was performed on HiSeq2500 as paired end,

100-read multiplex runs. Analysis was performed with the ShapeMapper pipeline (Siegfried *et al.*, 2014) using either *β-actin* mRNA (NM_001101) to align sequences derived from the *in vitro* samples or the entire genome (hg38) to align sequences derived from the *in vivo* alignment. The *β-actin* data are in the supplementary SNRNASM (S1). SHAPE traces for the wild type *vibrio vulnificus add* riboswitch mutate-and-map experiments were obtained from the publicly available RNA Mapping Database (RMDB)(37-39). To normalize the SHAPE-MaP data, scaled background reactivities were subtracted from the plus reagent reaction reactivities. A multiplier was used to fit the resulting distribution of values to the distribution of values for the normalized reactivities of a reference mRNA.

3.2.7. *β*-actin RNA Structural Modeling

An RNA/protein complex was generated from a starting model of two DNA strands bound to the KH34 protein (Chao *et al.*, 2010). A custom python script was used to convert the DNA strands to the appropriate RNA nucleotide sequence. The resulting RNA/protein complex was equilibrated by discrete molecular dynamics (DMD) simulations (Dokholyan *et al.*, 1998; Dokholyan *et al.*, 2011; Shirvanyants *et al.*, 2012) to accommodate the zipcode binding regions of the RNA strands. The remaining regions of the RNA strands were modeled using coarse-grained DMD simulations (Ding *et al.*, 2008) in which each nucleotide was represented as three pseudo-atoms corresponding to the phosphate backbone, sugar group, and nucleobase. With the replica exchange approach, we efficiently sampled RNA conformations by utilizing replicas of the same RNA system in parallel at different temperatures. Replicas were allowed to exchange simulation temperatures periodically based on a Monte Carlo algorithm. The replica exchange DMD simulations were run for 50 ns with replica temperatures of 0.200, 0.225, 0.250, 0.270,

0.300, 0.333, 0.367, and 0.400 with units kcal/(mol*kB). Free energy bonuses were incorporated between nucleotides to model the *in vivo* base pairing interactions. To select the final RNA model, we used a hierarchical clustering analysis based on the pairwise root mean square deviation (RMSD) of the phosphates and the potential energy as determined by the DMD force field. The coarse-grained RNA model was reconstructed to an all-atom model to combine with the KH34 protein system. We then equilibrated the entire RNA/protein complex using all-atom DMD simulations at a temperature of 0.4 kcal/(mol*kB) and included static constraints on the protein and harmonic constraints on the zipcode binding regions of the RNA strand.

3.2.8. *In vitro* model

We incorporated the *in vitro* secondary structure as constraints in coarse-grained replica exchange DMD simulations, using the same settings as those in the *in vivo* RNA system. We then performed an RMSD-based clustering analysis to determine the centroid and reconstructed an all-atom model at a temperature of 0.4 kcal/(mol*kB).

3.2.9. RNA Dynamics

The dynamics of the 2' hydroxyl groups of the *in vitro* and *in vivo* RNA strands were calculated using the root mean square fluctuation (RMSF) with the Wordom software package (Seeber *et al.*, 2011). RMSF calculations were performed on three 100 ns DMD simulations at a temperature of 0.4 kcal/(mol*kB) for both RNA systems. The *in vivo* system included static constraints on the protein and harmonic constraints on the zipcode binding protein-interacting regions of the RNA. We calculated the mean using 3-nucleotide windows and standard deviation of the RMSF based on the three DMD simulations for each system.

3.3 Results

3.3.1 Generating a robust 2-dimensional representation of an RNA ensemble

Our first goal in creating a visualization of a structural ensemble was to establish a robust and consistent 2-dimensional representation of the conformational space of RNA. Traditionally, principal component analysis (PCA) is used to identify two Eigen vectors for projection (Ding *et al.*, 2004; Ding *et al.*, 2005). One challenge with this approach is that the first three Eigen vectors often fail to capture enough variance to detect major structural elements (Quarrier *et al.*, 2010). If a conformation change is predicted, this limitation of PCA makes it difficult to understand the relative differences in the ensemble. A second challenge is determining which structural features to highlight in the representation to capture important biological aspects of the ensemble. Selecting features to highlight requires picking a specific structural distance representation, which can affect the interpretation as much as which Eigen vectors are used for projection. We propose an approach that provides a stable and robust visualization while also capturing important biological features (*e.g*. the "On" and "Off" conformation of the *add* riboswitch in Figure 3.1B).

Our approach is summarized in Figures 3.2 and 3.7. We begin by computing the partition function of the wild type RNA sequence and all single point mutants. From these partition functions, we select the RNAs that are maximally different, as determined by Shannon entropy and hierarchical clustering on base pairing probability (Defays, 1977; Shannon, 1951). From these partition functions, we sample the Boltzmann suboptimal ensemble and use these structures as the basis to build our visualization (Ding *et al.*, 2005). This strategy effectively allows us to more comprehensively sample the suboptimal ensemble and the strategy does not depend on the approach used to compute the partition function. The visualization creates a stable space for the

comparison of structural ensembles using mutations to explore the possible conformations that an RNA may take (Figure 3.1B). Data abstraction identifies clusters of similar structures that likely have similar function. This cluster representation reduces the map size, thereby creating a more accurate and interpretable visualization of secondary structure. Projecting the structure clusters into 2-dimensions using metric multidimensional scaling (MDS) optimizes their distances (Abdi, 2007; Torgerson, 1952). This approach enables easy interpretation of the visualization, in which clusters that are farther apart are more different. We can project the RNA ensemble of interest onto this space by varying the size of cluster bubbles based on the number of structures that belong to that cluster (Figure 3.1B). Experimental structure probing data can be included to guide the ensemble prediction (Diegan *et al.*, 2008). This method is further described in the Methods and Materials section.

3.3.2. Detecting RNA structure change induced by ligand binding

The *add* riboswitch is found in the 5'UTR of an mRNA that codes for adenosine deaminase (Lemay *et al.*, 2009; Serganov *et al.*, 2015). This riboswitch forms two distinct conformations that control translation of the adjacent coding region (Lemay *et al.*, 2009; Serganov *et al.*, 2015). The adenine-unbound conformation represses translation, and the adenine-bound conformation activates translation. Figure 3.1A shows the accepted secondary structures for the unbound and bound states as determined by crystallography and NMR (Serganov *et al.*, 2015). These secondary structures represent only two of several possible conformations that the riboswitch may adopt in the cell (Ding *et al.*, 2006; Matthews, 2006). Indeed, the map of conformational space produced by our visualization explores a total of five possible structure clusters including the two accepted conformations (Figure 3.1B). This

visualization produces a separation in 2-dimensional space between conformations that can bind adenine and activate translation and conformations that cannot bind adenine.

The structural difference induced by ligand binding for the *add* riboswitch is particularly well suited for the application of SHAPE data. Without experimental data to guide structure prediction algorithms, the accepted bound conformation dominates (Figure 3.3A), and differences in structure that result from changes in environment cannot be discerned. However, including SHAPE data in the ensemble prediction algorithms reveals differences in the *add* riboswitch structure with and without ligand (Figures 3.3B and 3.3C). In each ensemble, the respective structure observed in crystallography and NMR dominates. Thus, our visualization approach combined with SHAPE-directed structural modeling captures key structural features of the ensemble (Lemay *et al.*, 2009; Serganov *et al.*, 2015).

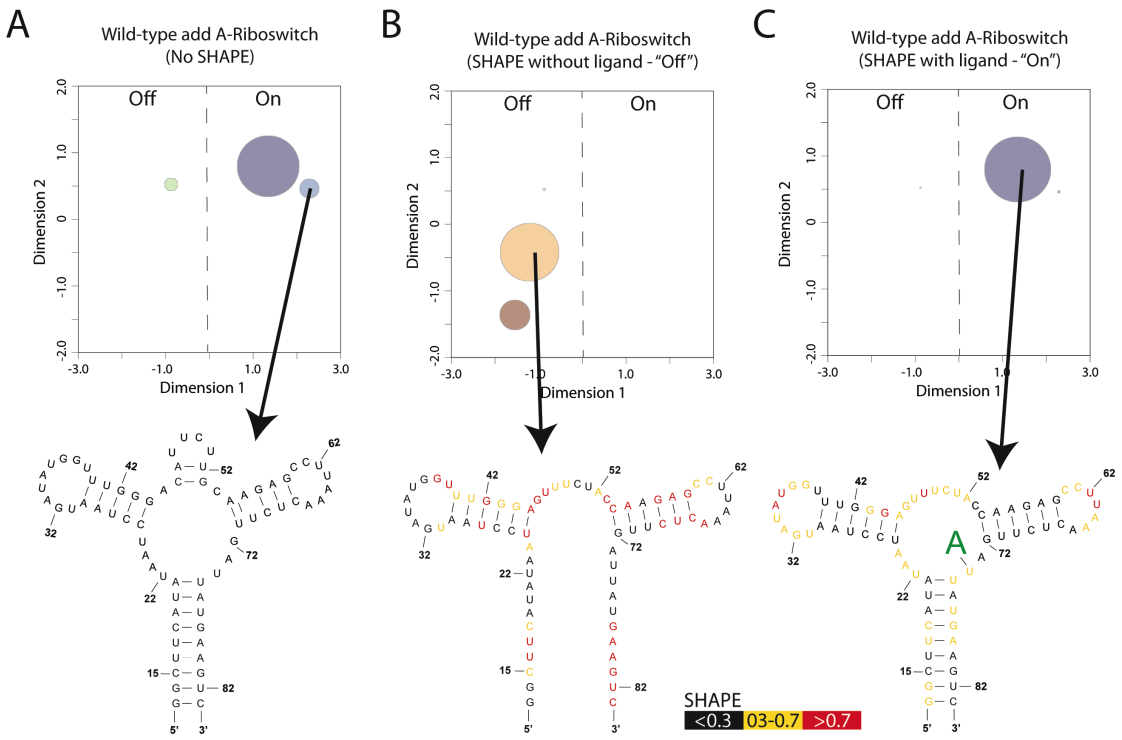


Figure 3.3 Visualization of the *vibrio vulnificus add* adenine riboswitch.

A) Projection of the predicted wild type ensemble without SHAPE data favors the experimentally determined "On" conformation (left). However, alternative conformations are still present (right). B) When the ensemble generation is guided by SHAPE experiments conducted without ligand, "Off" conformations are favored in the projection (left). Particularly, the experimentally confirmed "Off" structure is the most populated conformation. C) When SHAPE data are collected in the presence of ligand, the experimentally confirmed "On" conformation (right) is preferred in the projection (left). Both SHAPE data sets (with and without ligand) are publically available in the RNA Mapping Data Base (Cordero and Das, 2015; Kladwang *et al.*, 2011; Kladwang *et al.*, 2011).

3.3.3. Observing regional structure differences *in vitro* and *in vivo*

β-actin is a cytoskeletal protein involved in cell motility and structure (Guo *et al.*, 2013). The advent of high throughput structure probing methods such as SHAPE-MaP has only recently allowed us to collect information on larger RNAs such as the ~2-kb *β-actin* mRNA (Siegfried *et al.*, 2014). Structure probing data are collected for RNA in the presence of cellular components, *e.g.*, RNA-binding proteins (*in vivo*), and for free RNA (*in vitro*) (Spitale *et al.*, 2013). Thus, it is possible to detect structural differences in long mRNAs caused by differences in environments, such as the presence of ribosomes or RNA-binding proteins in the cell (Smola *et al.*, 2015; Smola *et al.*, 2016). Therefore, we performed SHAPE-MaP structure probing experiments on the *β-actin* mRNA present in *in vitro* and *in vivo* environments (Figure 3.4).

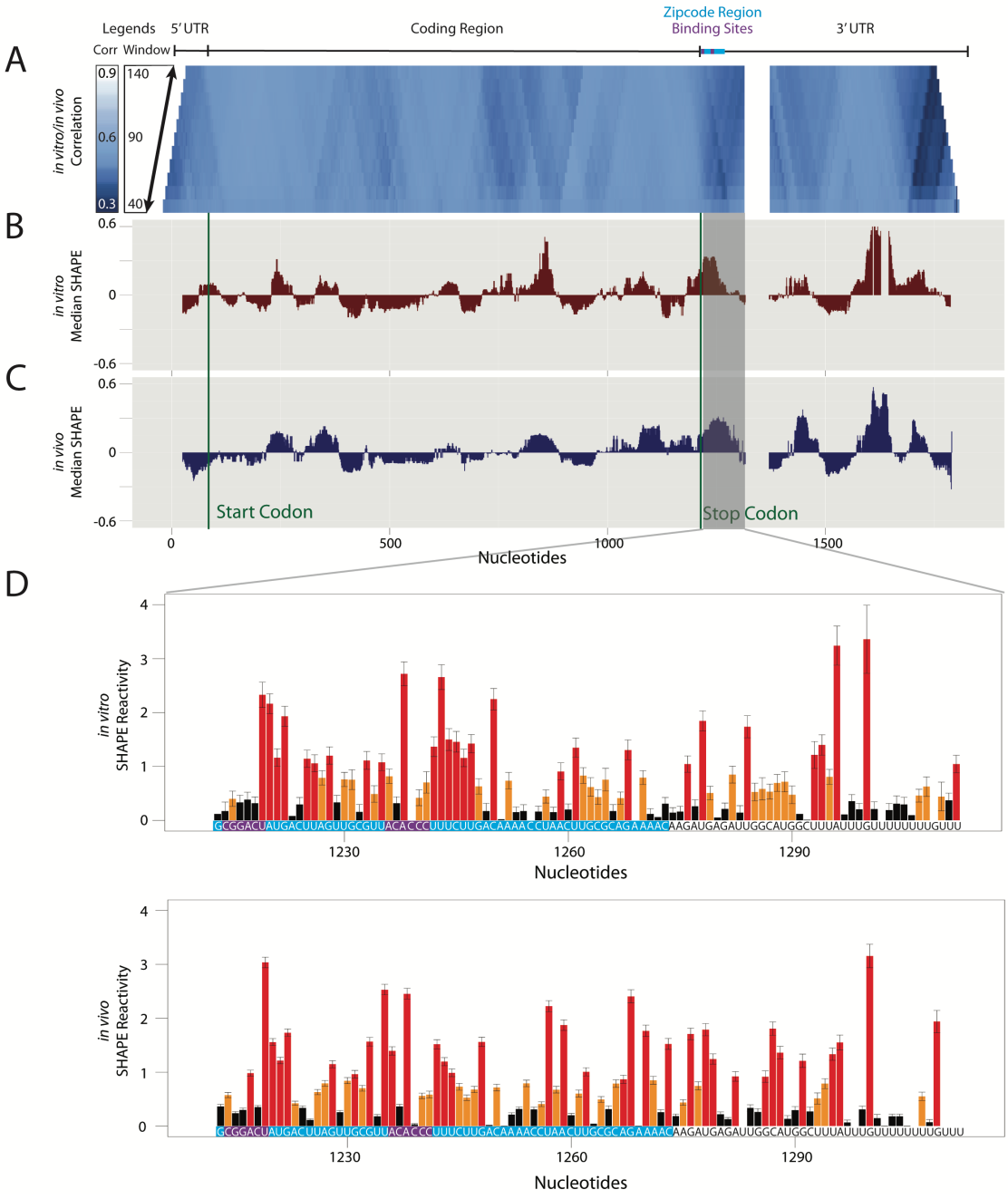


Figure 3.4. Comparison of *in vitro* and *in vivo* structure for the human *β-actin* mRNA.

A) We calculated the Pearson correlation in windows between the SHAPE reactivities collected *in vitro* and *in vivo* for the *β-actin* mRNA. For each step of the trapezoid from bottom to top, the window size increases by five nucleotides from 40 to 140. High correlation (white) corresponds

to areas that are similar in structure and low correlation (blue) corresponds to areas that are different in structure. The distances from the median SHAPE value for B) *in vitro* and C) *in vivo* *β-actin* were calculated in 50 nucleotide windows. Segments with reactivities above the median are less structured than segments with reactivities below the median. The gray panel highlights a region in which the SHAPE reactivity differs between *in vitro* and *in vivo*. D) This difference is seen in the SHAPE traces for *in vitro* (top) and *in vivo* trace (bottom). Structure probing was performed using the high throughput SHAPE-MaP technique. Red nucleotides correspond to high SHAPE reactivity, yellow corresponds to medium reactivity, and black corresponds to low reactivity. The ZBP1-binding region (bright blue) and two zipcode binding protein-interacting sites (purple) are labeled above the windowed correlation and at the bottom of the SHAPE traces. The overlay for the SHAPE traces is in Figure 3.10.

Because we are specifically interested in differences between the two environments (*in vivo* and *in vitro*), we compute the windowed SHAPE correlation coefficient between the two data sets and plot this correlation in Figure 3.4A for a range of window sizes (40 to 140 nucleotides). Overall, we observe high correlation between the two data sets for a majority of the mRNA's span, with a mean correlation coefficient of 0.88. This result can be seen clearly in Figure 3.8, in which we plot raw data for a highly similar window in the coding region of the gene. We begin our structural analysis by performing SHAPE-directed Boltzmann suboptimal sampling of nucleotides 200 to 400, which we identified as having high *in vitro* to *in vivo* correlation (Figure 3.8). We expect to observe only small changes in the suboptimal sampling because the SHAPE data in this region are highly similar. As expected, the visualization confirmed only small differences, but it identified a remarkably complex ensemble with 24 structural clusters (Figure 3.9). This result agrees with the high median SHAPE data (Figure 3.4B and 3.4C) observed for this region; high median SHAPE is correlated with higher ensemble entropy, *i.e.*, multiple alternative conformations (Siegfred *et al.*, 2014).

The region with the lowest correlation is at the 3' end of the mRNA. The *in vitro*-probed mRNA was transcribed in the absence of a polyA polymerase, therefore it was not polyadenylated, which likely explains the differences near the 3' end because the *in vivo* mRNA is most likely polyadenylated (and 5'-capped). The region of difference we chose to further characterize structurally occurs 3' of the stop codon. This region in the mRNA contains functional elements known as the Zipcode Protein Binding Protein Sites (ZPBS1 and ZPBS2). Binding of the zipcode binding protein (ZBP1) mediates mRNA localization and translation, hence the name of the protein (Kislauskis *et al.*, 1994; Lawrence and Singer, 1986). We used our ensemble visualization approach to characterize the *in vivo* conformational rearrangements

occurring in ZPBS1 and ZBPS2 within the ZBP1 binding region of the mRNA and to understand these rearrangements in the context of this region's function. The 54-nucleotide region we model below was previously identified as necessary and sufficient for localization of *β-actin* mRNA to the cell periphery (Kislauskis *et al.*, 1994; Lawrence and Singer, 1986). We therefore decided to specifically focus on the ensemble structure of this region.

Boltzmann suboptimal sampling for the ZBP1 binding regions *in vivo* visualized using our approach revealed a shift in the structural ensemble away from the preferred *in vitro* conformation toward an alternative conformation (Figures 3.5A and 3.5B). Nonetheless, the dominant conformation *in vitro* (Figure 3.5A) is still significantly populated *in vivo* (Figure 3.5B). Thus, our visualization suggests a more complex ensemble of conformations *in vivo.* To further understand the structural context of the shift in ensemble, we visualized the secondary structure medoid for each of the largest structure clusters *in vivo* and *in vitro*. Although in both conformations the Zipcode Binding Protein Sites (ZBPS) are unpaired, *in vivo* the dominant confirmation shows ZBPS1 and ZBPS2 in a contiguous unpaired region, consistent with the larger *in vivo* SHAPE values. Importantly, the SHAPE reagent is not a "footprinting" reagent and is only minimally affected by nucleotide accessibility (Merino *et al.*, 2005; Wilkinson *et al.*, 2006). Thus, it is not surprising that we observed higher SHAPE values surrounding the ZBPS. In fact, the ZBP1 is divalent, and it has been shown to simultaneously bind the two ZBPS motifs separated by a linker portion of the RNA, although the precise occupancy of the second site is not known (Chao *et al.*, 2010; Patel *et al.*, 2012). Nonetheless, binding to this region is essential for correct *β-actin* mRNA localization and translational control (Hüttelmaier *et al.*, 2005; Ross *et al.*, 1997). To accommodate the ZBP1 protein, the RNA likely has to become more open and flexible, consistent with the higher SHAPE data we observed.

To further understand the *in vivo* structural rearrangement, we performed molecular simulations of apo and bound mRNA conformations (Figures 3.5C and 3.5D). By using the secondary structure as initial constraints, we aimed to estimate the root mean square fluctuations (RMSF) of the RNA backbone. We show these data for the apo and bound simulations in Figures 3.5E and 3.5F, overlaid with the *in vitro* and *in vivo* SHAPE data, respectively. We observed qualitative agreements between the experimental SHAPE data and the simulations, suggesting these molecular models captured overall aspects of the conformational ensemble. One important aspect of these comparisons, especially in the case of the *in vivo* data, is that the SHAPE data are an ensemble average over all of the *β-actin* mRNA molecules in the cell. Because ZBP1 binding represses translation, some message molecules are likely not bound by ZBP1, a situation that may explain why a shift to multiple conformations is observed *in vivo* as opposed to observing only the bound conformation. Nonetheless, these data demonstrate the value of visualizing the structural ensemble to explain structure/function relationships in an mRNA.

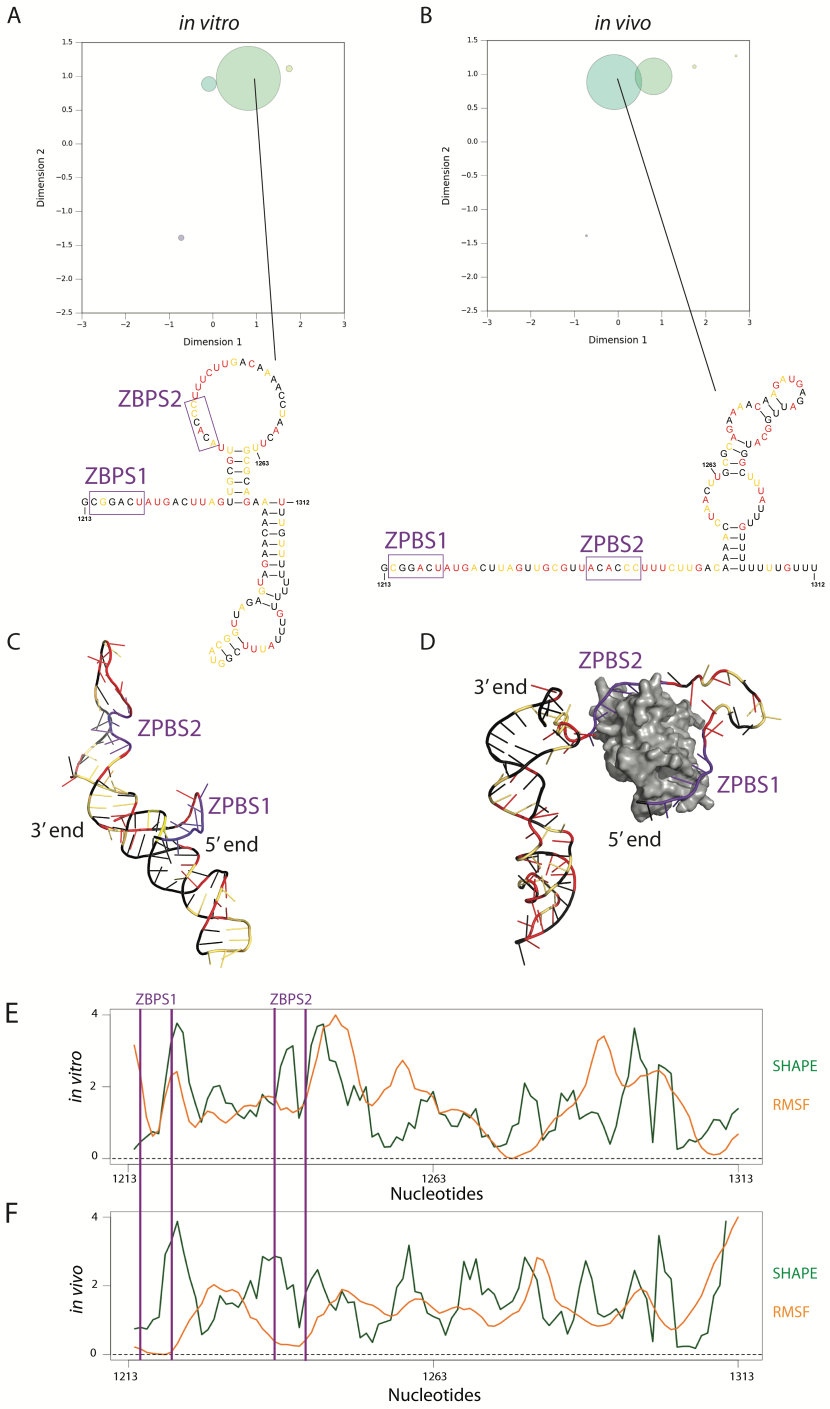


Figure 3.5. Ensemble visualization for in vitro and in vivo human *β-actin* mRNA.

Generation of structures for the *β-actin* mRNA ensemble was guided by the *in vitro* and *in vivo* SHAPE data. We compared the A) *in vitro* and B) *in vivo* ensembles for the region where SHAPE reactivities were expected to be different. The ensemble visualization reveals a large shift away from the dominant

structure *in vitro* toward a second structure *in vivo*. We visualized the second structure for the medoid in each of the largest structure clusters. These nucleotides form different structures *in vitro* and *in vivo*. The region that differs includes the zipcode region with the two ZBP1 binding sites (purple). C) The 3D structure for *β*-actin *in vitro* was modeled using molecular dynamics simulations without ZBP1. D) The 3D structure for *β*-actin *in vivo* was modeled with the ZBP1 (in gray). For both 3D models, the ZBP1 binding regions are highlighted in purple. Red nucleotides correspond to high SHAPE reactivity, yellow correspond to medium reactivity, and black corresponds to low reactivity in Figures A-D and E) Comparison of SHAPE reactivity (green) and normalized Root Mean Square Fluctuation (RMSF; orange) for *β*-actin *in vitro* largely follow the same pattern. F) Comparison of SHAPE reactivity and RMSF for *β*-actin *in vivo* also largely follow the same pattern. The SHAPE reactivities and RMSF values are averaged across a 3-nucleotide moving window. The RMSF is calculated from the 3D structural models. ZPB1 binding sites for Figures E and F are boxed in purple. Figure 3.11 includes further comparisons between *in vitro* and *in vivo* SHAPE reactivity and RMSF.

3.4 Discussion

RNA structure is key component of cellular function in highly specific instances; the ribosome's unique catalytic core is a prime example of the role of a specific RNA structure in performing protein synthesis (Ban *et al.*, 2000; Harms *et al.*, 2001; Ramakrishnan, 2002; Wimberly *et al.*, 2000; Yonath, 2010). Generally, however, the functions of structures in messenger RNAs are poorly understood, except for a few cases, such as the Iron Responsive Element (Halvorsen *et al.*, 2010; Ritz *et al.*, 2012) and the Histone Stem Loop (Harris *et al.*, 1991; Pandey and Marzluff, 1987; Sun *et al.*, 1992), in which single structures are essential for function. Other than ribosomal RNA, no RNA larger than 1 kb, including mRNAs, is known to fold into a unique, well-defined conformation (Ban *et al.*, 2000; Berman *et al.*, 1992; Narayanan

et al., 2013; Ramakrishnan, 2002; Wimberly *et al.*, 2000; Yonath, 2010). Still, although large RNAs do not adopt single conformations, specific regions do fold into complex 3-dimensional structures. One example is riboswitches in bacteria (Figure 3.1). Although riboswitches are considered to be structured (*i.e.*, they can be crystallized), riboswitches adopt multiple conformations that lead to different functions (Ritz *et al.*, 2013). Because RNAs such as riboswitches have evolved to form multiple conformations in order to function, it is essential to consider the suboptimal ensemble when considering structure in messenger RNAs (Ritz *et al.*, 2013).

Our approach to visualizing the suboptimal ensemble is designed to resolve some of the longstanding problems with obtaining a stable projection that allows comparisons of ensembles. *A priori,* this visualization approach requires sampling the entire suboptimal space to identify good principal components. For any biologically relevant RNA, such sampling rapidly becomes computationally intractable because the number of suboptimal conformations increases exponentially with length (Giegerich *et al.*, 2004; Sachs *et al.*, 1997). Thus, our approach is empirical (Figure 3.2) and relies on rapid sampling of suboptimal ensembles for single point mutants of the RNA (Ding *et al.*, 2005). Combined with multi-dimensional scaling and a "shape" based abstraction (Giegerich *et al.*, 2004; Steffen *et al.*, 2006), our maps have the desired properties of stability and they enable comparison of different ensembles.

The main biological motivation for our approach is the need to visualize changes in the ensemble caused by environment. Our results on the *vibrio vulnificus add* riboswitch leverage the empirical relationship between SHAPE reactivity and the free-energy of folding to recapitulate the apo and bound RNA ensembles (Figure 3.3). Importantly, the goal of these visualizations is to facilitate the understanding of a complex process by approximating the

specific abundance of each conformation in an ensemble. Moreover, we aim to extract biological insight from the ensemble calculation; for the *vibrio vulnificus add* riboswitch, our visualization of the ensemble model recapitulates the understanding of this system in an easily interpreted diagram. The system is relatively straightforward, which is not the case for complex eukaryotic mRNAs that tend to be much more highly regulated and structurally sensitive to their environments (Kutchko and Laederach, 2016; Solem *et al.*, 2015).

Our analyses of a full-length human mRNA *in vivo* and *in vitro* revealed some of the complexities associated with interpreting structures in large RNAs. We observed, in both conditions, regions of high (unstructured) and low (structured) median SHAPE (Smola *et al.*, 2016), results consistent with locally structured regions. Overall, the high similarity between *in vivo* and *in vitro* SHAPE data suggests that the mRNA is not globally affected by its environment, but, instead, specific regions are affected by endogenous molecule binding. Local structure is the case for the ZPB1 binding region in the 3' UTR of *β-actin*, which we visualized using our ensemble approach (Figure 3.5).

A significant result of this analysis is the median windowed SHAPE, which overall appeared higher *in vivo* relative to *in vitro* for the ZPB1-binding region. This result may seem counterintuitive, as the ZBP1 would be expected to protect the RNA from the 1M7 reagent. Although protein binding is detectable by SHAPE comparisons *in vitro* to *in vivo* (McGinnis *et al.*, 2015; Smola *et al.*, 2016), SHAPE chemistry is not a traditional "footprinting" technique (Brenowitz *et al.*, 2002; McGinnis *et al.*, 2012; Quarrier *et al.*, 2010; Shcherbakova *et al.*, 2006). Thus, it is likely that the majority of the differences in the SHAPE reactivity in this region are due to a conformational rearrangement due to protein binding, and not the footprint of the protein.

Our model (Figures 3.5A and 3.5B) successfully reports a shift in the ensemble, but the model does not suggest a totally dominant alternative *in vivo* conformation. This restriction is in contrast to the *add* riboswitch, in which ligand excess shifts the ensemble to almost completely the "On" conformation (Figure 3.3C). It is important not to over interpret the relative ratios of the two dominant conformations proposed for the ZBP1-binding region modeled in Figure 3.5B. However, the model is consistent with our expectation of a mixed population of ZBP1-bound and unbound *β-actin* mRNA. Also, the fact that the ZBP1 has two binding sites and these sites are not always simultaneously occupied (Chao *et al.*, 2010; Kim *et al.*, 2015) is an additional aspect that our model cannot currently describe. Thus, our visualization accurately represents the likely state of the population of *β-actin* mRNAs in the cell but still requires biological knowledge to be fully interpretable.

We performed constrained molecular dynamics simulations of the two proposed structural models of *β-actin* mRNA to determine if the models agreed qualitatively with the SHAPE data. Because SHAPE chemistry measures backbone flexibility (Diegan *et al.*, 2009; McGinnis *et al.*, 2012), we report root mean square fluctuations for both models in Figure 3.5E and 3.5F. For the ZBP1-binding region between ZBPS1 and ZBPS1, the agreement between the simulation and SHAPE data is better for the *in vitro* model compared with the *in vivo* simulation. For the *in vivo* model, we constrained both ZBPS1 and ZBPS2 to the binding pockets, which explains the low flexibility of ZBPS1 and ZBPS2. The higher SHAPE data for these two binding sites *in vivo* are consistent with a significant subset of mRNAs being unbound, which agrees with our ensemble model that suggested a further opening of the structure.

In summary, we have developed a computationally-based visualization approach that faithfully represents ensemble mRNA populations and the effects of environment on the

ensembles. The *β-actin* mRNA and the *vibrio vulnificus add* riboswitch are two well characterized systems in which ensemble visualization improves the interpretation of environmentally imposed structural differences. By releasing a software package to create these visualizations easily, we encourage the RNA folding community to simulate more than just minimum free energy structures and to explore the suboptimal ensemble for all mRNAs existing in a cell. It is not clear whether suboptimal alternative conformations are a necessary component of RNA function in the cell or a by-product of the rules that govern RNA folding (Gracia *et al.*, 2016; Herschlag *et al.*, 2015; Russell *et al.*, 2002; Solomatin *et al.*, 2010). Regardless, structure ensembles are a thermodynamic reality of RNAs and are accommodated as a feature of their function.

3.5 Supplementary Materials

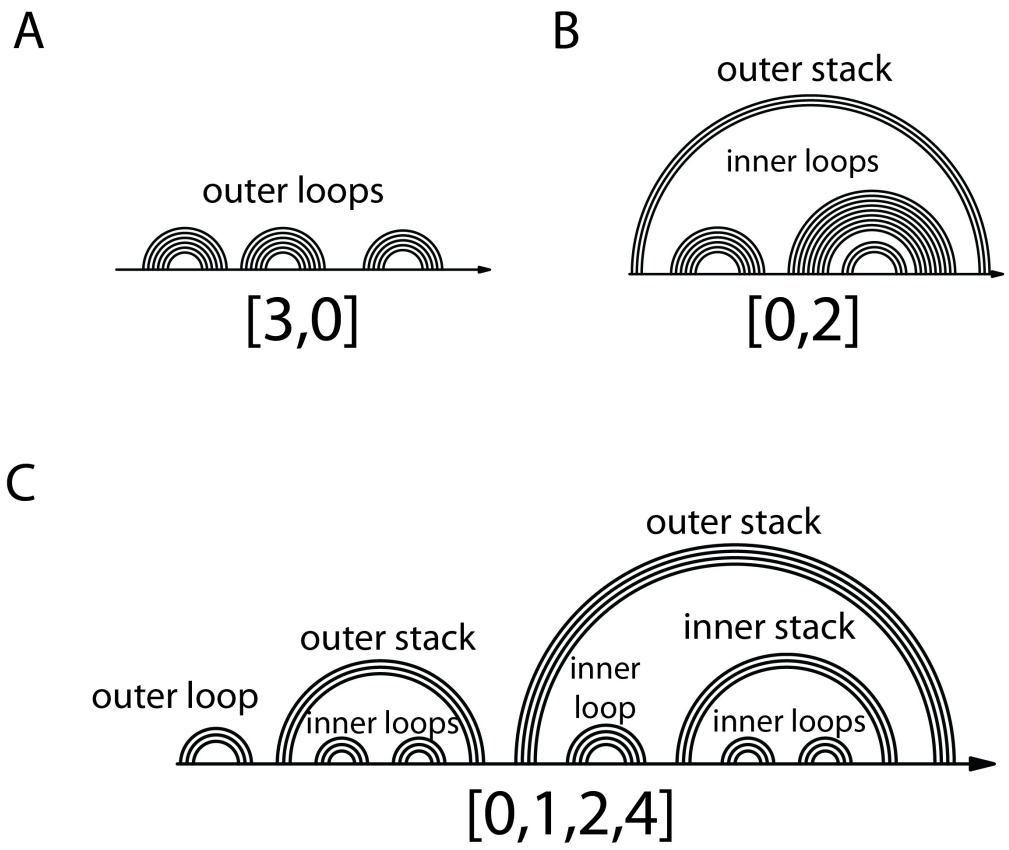


Figure 3.6. RNA structure abstraction and nestedness.

We utilized RNAshapes abstraction to identify unique structure clusters (1). This abstraction assumes that the sizes of structural elements are less important than whether they are present. We used this method to create a numeric vector representation. This representation is based on the nestedness of RNA stacks and loops. A) If only outer loops are present in a structure, we place the number of outer loops in the first column. B) If an outer stack is present, we place the number of loops inside that stack in the n+1 column. C) For each outer loop or outer stack, we place the number of inside loops into the n+1 column

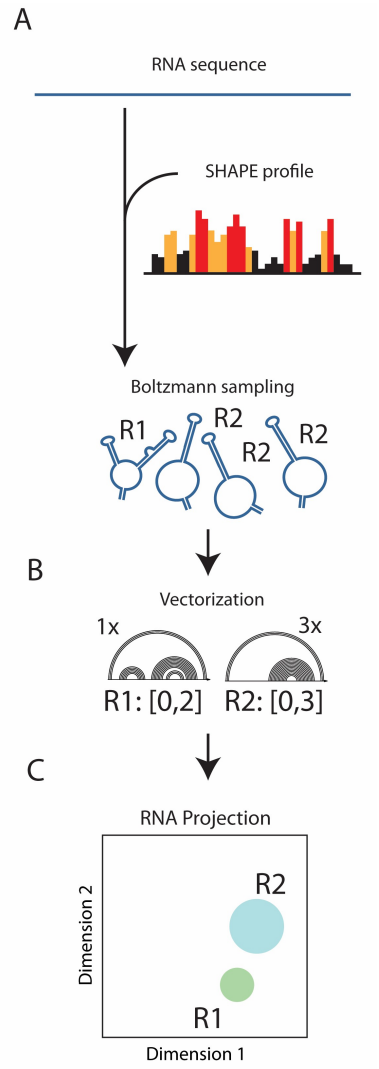


Figure 3.7. Projection of the reference RNA.

A) We generated 1000 structures using Boltzmann-weighted suboptimal sampling from the reference RNA sequence (2). We used experimental structure data collected from selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) to guide the ensemble prediction (3, 4). Each structure was converted into our nestedness representation (Figure 3.2E). We retained the orientation of the points from the map of conformational space. The size of each bubble was varied based on the frequency for that structure cluster in the wild type ensemble.

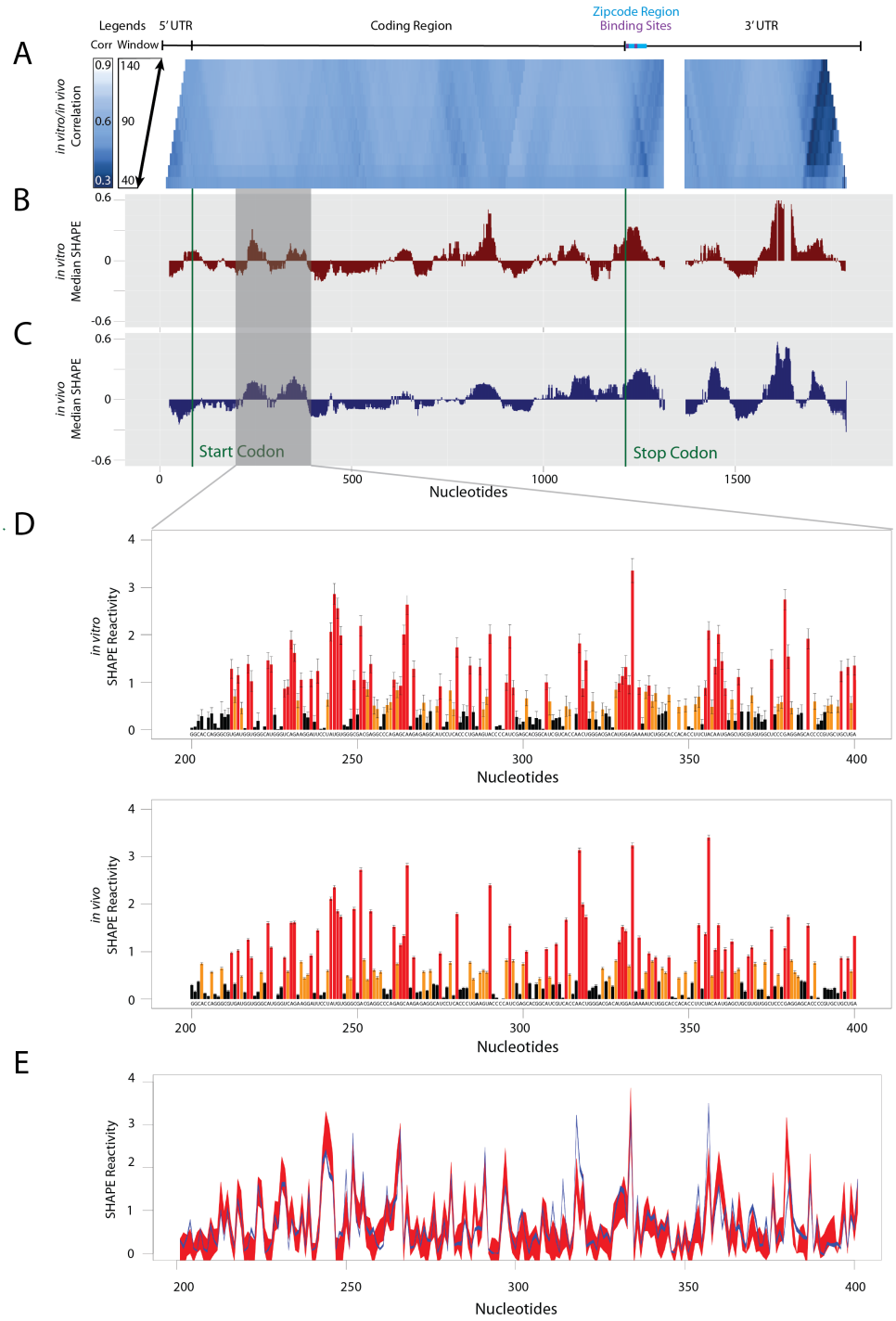


Figure 3.8. Comparison of similar *in vitro* and *in vivo* structure for the *β-actin* mRNA.

A) We calculated the Pearson correlation in windows between the SHAPE reactivities collected *in vitro* and *in vivo* for the *β-actin* mRNA. For each step of the trapezoid from bottom to top, the window size increases by five nucleotides from 40 to 140. High correlation (white) corresponds

to areas that are similar in structure and low correlation (blue) corresponds to areas that are different in structure. The distances from the median SHAPE value for B) *in vitro* and C) *in vivo* *β-actin* were calculated in 50 nucleotide windows. Segments with reactivities above the median are less structured than segments with reactivities below the median. The gray panel highlights a region in which the SHAPE reactivity is the same *in vitro* and *in vivo*. D) This similarity is reflected in the *in vitro* (top) and *in vivo* (bottom) SHAPE traces. E) The *in vitro* (red) and *in vivo* (blue) SHAPE traces were overlaid for this region. The thickness of the line corresponds to the error. Structure probing was performed using the high throughput SHAPE-MaP technique. The zipcode region (bright blue) and two zipcode protein-binding sites (purple) are labeled above the windowed correlation and at the bottom of the SHAPE traces.

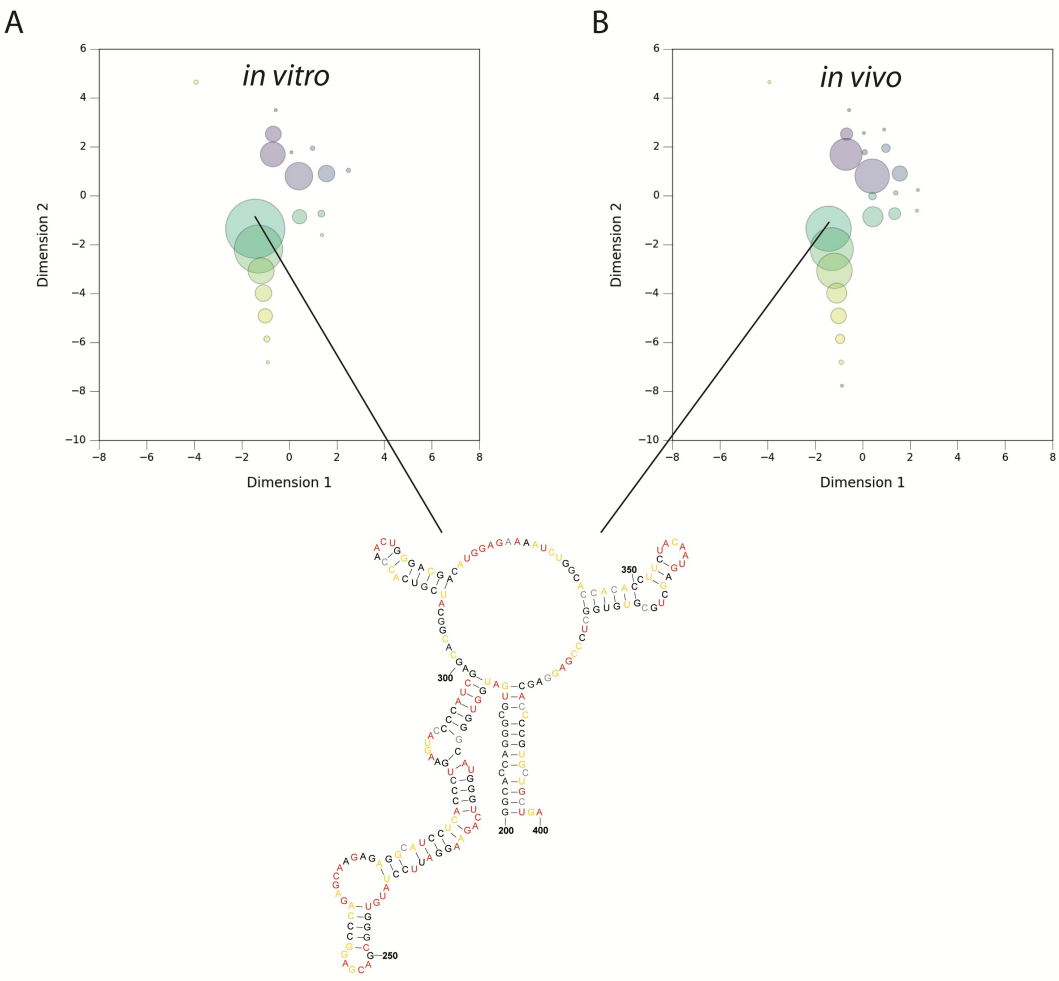


Figure 3.9. Similar *in vitro* and *in vivo* ensembles for human *β-actin* mRNA.

Generation of structures for the *β-actin* mRNA ensemble was guided by the *in vitro* and *in vivo* SHAPE data. The 200-nucleotide regions were folded separately. We compared the visualizations for A) *in vitro* and B) *in vivo* SHAPE-guided ensembles for a region where SHAPE reactivities were expected to be the same. The visualization confirms that the *in vitro* and *in vivo* ensembles are the same. The medoid structure for the most common cluster is shown (center).

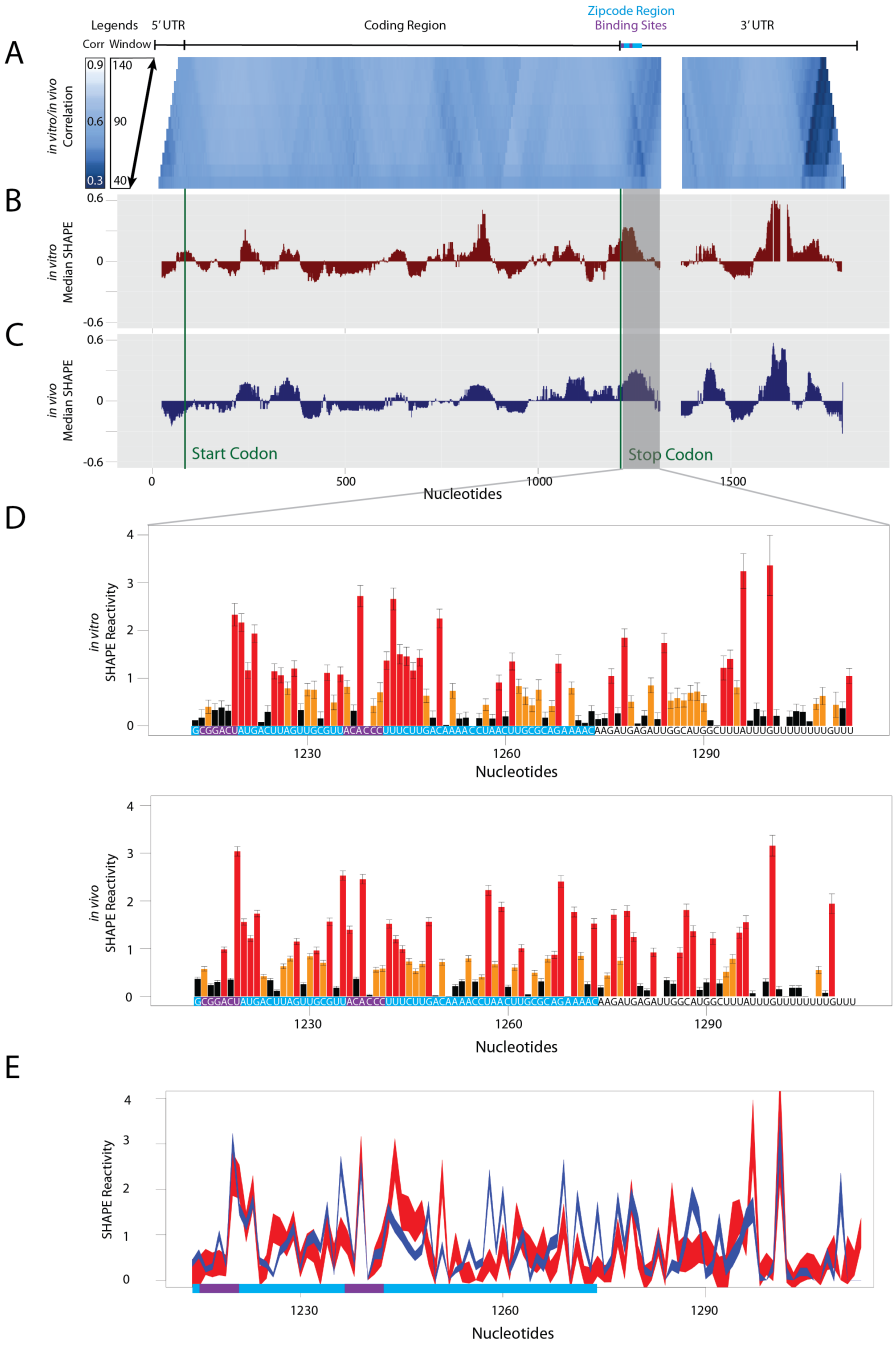


Figure 3.10. Comparison of different *in vitro* and *in vivo* structure for the *β-actin* mRNA.

A) We calculated the Pearson correlation in windows between the SHAPE reactivities collected *in vitro* and *in vivo* for the *β-actin* mRNA. For each step of the trapezoid from bottom to top, the window size increases by five nucleotides from 40 to 140. High correlation (white) corresponds

to areas that are similar in structure and low correlation (blue) corresponds to areas that are different in structure. The distances from the median SHAPE value for B) *in vitro* and C) *in vivo* *β-actin* were calculated in 50 nucleotide windows. Segments with reactivities above the median are less structured than segments with reactivities below the median. The gray panel highlights a region in which the SHAPE reactivity is different *in vitro* and *in vivo*. D) This difference is reflected in the *in vitro* (top) and *in vivo* (bottom) SHAPE traces. E) The *in vitro* (red) and *in vivo* (blue) SHAPE traces were overlaid for this region. The thickness of the line corresponds to the error. Structure probing was conducted using the high throughput SHAPE-MaP technique. The zipcode region (bright blue) and two zipcode protein binding sites (purple) are labeled above the windowed correlation and at the bottom of the SHAPE traces.

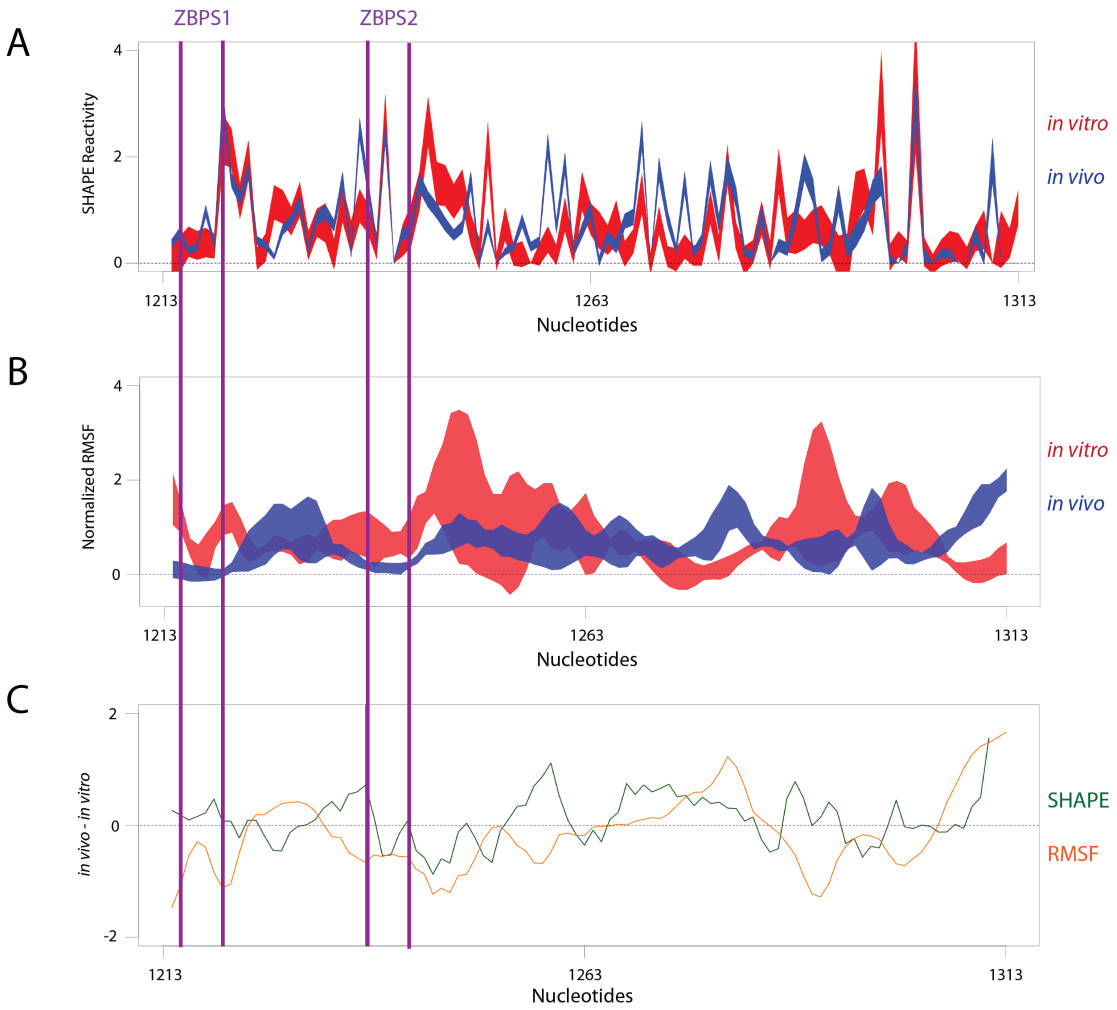


Figure 3.11. Comparison of different *in vitro* and *in vivo* flexibility for the *β-actin* mRNA.

A) The *in vitro* (red) and *in vivo* (blue) SHAPE traces were overlaid for the zipcode region of *β*-actin . The thickness of the line corresponds to the error. Structure probing was performed using the high throughput SHAPE-MaP technique. B) The normalized Root Mean Square Fluctuation (RMSF) for *in vitro* (red) and *in vivo* (blue) *β*-actin. The fluctuation is calculated from the 3D structural models shown in Figure 3.5. RMSF values were averaged over a 3-nucleotide moving window. Line thickness corresponds to standard error over three molecular dynamics simulations for each scenario. C) Comparison of the difference between in vivo and in vitro for SHAPE (green) and RMSF (orange). Values above zero indicate higher reactivities or RMSF for the *in*

vivo sample. These values were averaged over a 3-nucleotide moving window. Values below zero indicate higher reactivities or RMSF for the *in vitro* sample. The zipcode binding sites are labeled with purple vertical lines for Figures A-C.

CHAPTER 4: CONCLUSION

Mutations in RNA will create a riboSNitch, if important structural elements are disrupted (Ritz *et al.*, 2012). Some of these structural changes will result in a functional consequence. Recent ultra-high throughput techniques, such as SHAPE-MaP, enable the collection of structural RNA information on a genome-wide scale (Siegfried *et al.*, 2014). With the ability to gather genome-wide structural information on RNA, it is important that to accurately classify these structural data in order to identify those structural changes that result in a phenotypic outcome. Furthermore, there are significant differences in signal-to-noise in transcriptome-wide data sets, such that false-discovery rates of riboSNitches can be significant (Siegfried *et al.*, 2014). We therefore developed an approach to determine whether a putative structure change is supported by the data. We set out to develop an automated approach to detect structure change in SHAPE data. We used a training set of 167 RNA mutations to detect riboSNitches (Cordero and Das, 2015; Cordero *et al.*, 2012; Kladwang *et al.*, 2011). Comparison of our autonomous classification system against a crowd-sourced manual classification system for these putative riboSNitches gave insight into how well our autonomous system performs in comparison to the manual system. These data and analyses will also allowed us to define the expected minimal change of a mutation in an RNA and thus identify interesting and therefore functional riboSNitches.

We were not only interested in identifying mutations that lead to riboSNitches, but also in understanding what elements of those structures are changing and how those relate to function

and disease. A majority of ribonucleic acid (RNA) molecules adopt multiple conformations (Ritz *et al.*, 2012). The ability to form multiple conformations is often important to the function of an RNA, as in the case of riboswitches. These RNAs control translation and/or transcription of messenger RNAs through a conformational change induced by the binding of a ligand, and consequently must adopt at least two different conformations (Serganov *et al.*, 2004). The ensemble of possible RNA conformations can be sampled using Boltzmann suboptimal sampling (Ding *et al.*, 2005; Ding and Lawrence, 2003). The results of these predictions are highly complex, and there exists no standard approach for visualizing an RNA suboptimal ensemble. Moreover, it is biologically important to capture functionally relevant structural variation in a given visualization. RNA molecules are sensitive to mutations and other factors that cause shifts in the ensemble toward conformations that can disrupt important structural elements (Kutchko *et al.*, 2015). Visualization of the relationships between structures in an ensemble is key to understanding the effects of mutation or environment on RNA folding, stability and function. Current visualization methods rely on single best representations, incomplete structural information, or unstable structural space for comparison (Ding *et al.*, 2005; Ritz *et al.*, 2012). These methods are often useful for visualizing RNA structure and base pairing probability, but are not sufficient to explore the functional consequence of structure and structure change in an ensemble. Therefore, we have developed a method that creates a stable map of conformational space for a given RNA and its mutants. We explore the most diverse conformational space for this map and generate the structures using established Boltzmann-weighted suboptimal sampling algorithms (Ding and Lawrence, 2003; Ding and Lawrence, 1999). Using vector representation based on arc diagram nested loop patterns, we project clusters of structures from the map into two dimensions using metric multi-dimensional scaling. Individual RNA ensembles are

visualized in this space by fluctuating the size of the structure clusters in a bubble plot. In combination with ultra-high throughput experimental methods for structural determination, we used this visualization method to explore differences in RNA ensembles. We visualized how mutations and changes in environment could lead to shifts in the structural ensemble that may alter function. With the inclusion of selective 2'-hydroxyl acylation and primer extension (SHAPE) data, we modeled how experimental data restricts the number of conformations predicted with sampling algorithms. Using RNAs for which we have structural information from nuclear magnetic resonance (NMR) and crystallography or functional information from experimental assays, we can validate these observations. Ultimately we aimed to determine how changing structural elements lead to differences in RNA function, and to establish what are biologically important features for structure change.

4.1 Important Findings

Using the classSNitch classifier, we were better able to recapitulate expert classification of experimental structure probing data into global, local and non-changers. For this classifier, we combined existing metrics for describing changes in experimental data, with metrics we created specifically for this problem. We also re-tooled metrics that have been previously used in other fields for this purpose. These features can be used to better describe the behavior of experimental structure probing data. We identified 2019 SHAPE traces for eleven different RNAs. Classifying these traces we found that some RNAs are resistant to structure changing mutations, this is particularly true of the highly structured RNAs such as the phenylalanine tRNA, 16S four-way junction and 5S ribosomal RNA. Other RNAs are more sensitive to mutations that alter structure, like the synthetic Tebowned aptamer. We hypothesized that the seemingly poor performance of

structure change prediction algorithms may be partly attributable to inaccurate metrics used to establish the ground truth. We instead used classSNitch to assess the performance of these algorithms, and we found that previous metrics may have been underestimating their performance. We also found that by including experimental data from the wild-type RNA (whether the SHAPE data indicated it was paired at the mutation site, or a C/G nucleotide at the mutation site), we could improve the performance of the SNPfold structure change prediction algorithm.

To validate our visualization approach we established a conformational map of the *vibrio vunificus add* adenine riboswitch that reveals five classes of structures. In the presence of adenine, SHAPE-directed sampling and projection onto the map correctly identifies the correct "on" conformation, while in the absence of the ligand, only "off" conformations are present and visualized. We also collected high-accuracy whole-transcript *in vitro* and *in vivo* SHAPE-Map data (Selective 2' Hydroxyl Acylation by Primer Extension- Mutational Profiling) on human *β-actin* mRNA revealing similar folds in both conditions. Nonetheless, specific regions in the mRNA, including near the stop codon, yield significantly different SHAPE-MaP profiles, consistent with a structural rearrangement. Our visualization strategy identifies a specific structural rearrangement in a 54-nucleotide region downstream of the stop codon consistent with zipcode protein binding.

4.2 Approach Weaknesses

Despite the success of these approaches there are apparent drawbacks. The classSNitch classifier was originally trained on structure probing data collected using capillary electrophoresis. However, we eventually wish to apply this technique to genome-wide structure

probing data. Our lab is currently exploring this application of classSNitch and performance of the algorithm seems similar to that found in our previous study. However, as more genome-wide data becomes available, it will be necessary to continue evaluating classSNitch performance. Also this approach is limited to the structural information that can be gleaned from the flexibility of the 2'-hydroxyl of the RNA backbone. The relationship between this information on tertiary structure is complex and is beyond the scope of our study. However, incorporating wild type SHAPE data into structure change prediction helped to improve the performance of the SNPfold algorithm, therefore such experimental data can still be useful in identifying important structure change in RNA. Our approach was limited to 200 experimental traces, due to the laborious task of gathering expert validation. This may limit how generalizable the classifier is for new data. Our data still suggests that this approach is feasible for developing a standard, and can be further expanded in future work.

As the length of an RNA increases, the possible structural ensemble space increases exponentially (Giegerich *et al.*, 2004). Despite the use of data abstraction and dimensionality reduction, for longer RNAs it may be necessary to look at larger groups such as "clusters of clusters" in order to better the long-range structural differences. Additionally, the range of flexibilities among RNAs means that the required number of mutants to create the consistent map of conformational space may be difficult to determine. Changes that occur on a smaller scale, such as the pairing or unpairing of a few bases within a stem loop, may not be caught using our method. Future work may be aimed at overcoming these remaining challenges.

4.3 Future directions

The machine learning tools created in this work are potentially useful for a variety of future applications. The features used in the classifier may be useful for describing structure change in experimental data even when used individually. In particular, the dynamic time-warping metric may be useful in identifying shifts in SHAPE profiles that can be corrected. In future studies, we intend to use classSNitch to identify structure change in genome-wide structure probing data. More work must be done to determine the efficacy of classSNitch on this kind of experiment. Additionally, benchmarking studies comparing SNPfold and other structure change prediction algorithms may be improved by using classSNitch to establish the ground truth. Currently, this tool is being used to classify variants in human mRNAs that are associated with cancer. These variants are located in the coding region of the RNA but do not change the protein sequence, and so we believe that structure may be playing a role. By identifying the structure changing variants, we can choose RNAs for further experiments like functional assays.

EnsembleRNA is currently being used to visualize structural ensembles in RNAs associated with cancer, infertility and viruses. This visualization tool may be particularly useful for understanding structure change *in vivo,* where RNA exists in many possible states. Structure probing methods collect data on an ensemble of structures, and using this visualization tool better reflects the reality of the data. By releasing EnsembleRNA as open-source package, we hope to encourage the RNA community to explore the important structural nuances found in the structural ensemble that will not be reflected in the minimum free energy structure.

These machine learning tools have provided us with the ability to better understand how RNA structure relates to disease. We have shown how useful these analysis tools can be in studying complex data. And shown their potential for use on data sets that are too large for

analysis by individuals. In recent years, technology has allowed us to collect more information than ever before on RNA and other biologically relevant systems. With this burgeoning data in mind, we believe that analysis tools such as those discussed in this work will be more important than ever before.

REFERENCES

Abdi, H. (2007) Metric Multidimensional Scaling (MDS): Analyzing Distance Matrices. Journal.

Abdi, H. and Williams, L. (2010) Pincipal Components Analysis. *WIley Interdisciplinary Teviews: Computational Statistics.* 2, 433–459.

Alexander, T. *et al.* (2009) Hox genes and segmentation of the hindbrain and axial skeleton. *Annu Rev Cell Dev Biol.* 25, 431-456.

Bai, Y. *et al.* (2005) Probing counterion modulated repulsion and attraction between nucleic acid duplexes in solution. *Proceedings of the National Academy of Sciences USA.* 102, 1035-1040.

Ban, N. *et al.* (2000) The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science.* 289, 905-920.

Banerjee, A.R. *et al.* (1993) Thermal unfolding of a group I ribozyme: the low temperature transition is primarily a disruption of tertiary structure. *Biochemistry.* 32, 153-163.

Berman, H.M. *et al.* (1992) The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.* 63, 751-759.

Bernhart, S.H. *et al.* (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol.* 1, 3.

Bokinsky, G. and Zhuang, X. (2005) Single-molecule RNA folding. *Acc Chem Res.* 38, 566-73.

Breiman, L. (2001) Random forests. *Machine Learning.* 45, 5-32.

Brenowitz, M. *et al.* (2002) Probing the structural dynamics of nucleic acids by quantitative time-resolved and equilibrium hydroxyl radical 'footprinting'. *Current Opinion in Structural Biology.* 12, 648-653.

Brenowitz, M. *et al.* (1986) "Footprint" titrations yield valid thermodynamic isotherms. *Proc Natl Acad Sci U S A.* 83, 8462-6.

Brenowitz, M. *et al.* (1986) Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol.* 130, 132-81.

Brun, M. *et al.* (2008) Which Is Better: Holdout or Full-Sample Classifier Design? *EURASIP J Bioinform Syst Biol.* 1, 297945.

Butler, E. *et al.* (2011) Structural basis of cooperative ligand binding by the glycine riboswitch. *Chem Biol.* 18, 293-298.

Chao, J.A. *et al.* (2010) ZBP1 recognition of b-actin zipcode induces RNA looping. *GENES & DEVELOPMENT.* 24, 148-158.

Chauhan, S. and Woodson, S.A. (2008) Tertiary interactions determine the accuracy of RNA folding. *J Am Chem Soc.* 130, 1296-303.

Chen, C. *et al.* (2004) Using random forest to learn imbalanced data. Journal.

Chen, G. *et al.* (2004) Factors affecting thermodynamic stabilities of RNA 3x3 internal loops. *Biochemistry.* 43, 12865-12876.

Chen, X. and Ishwaran, H. (2012) Random forests for genomic data analysis. *Genomic.* 99, 323-329.

Chen, X. *et al.* (2011) The use of classification trees for bioinformatics. *WIREs Data Mining and Knowledge Discovery.* 1, 55-63.

Cheng, C.Y. *et al.* (2015) Consistent global structures of complex RNA states through multidimensional chemical mapping. *Elife.* 4, e07600.

Cheng, Z. *et al.* (2005) Crystal structure and functional analysis of DEAD-box protein Dhh1p. *Rna.* 11, 1258-70.

Churkin, A. *et al.* (2011) The RNAmute web server for the mutational analysis of RNA secondary structures. *Nucleic Acids Res.* 39, W92-9.

Cleary, J. and Trigg, L. (1995) K*: an instance-based learner using an entropic distance measure. *Proceedings of the 12th International Conference on Machine Learning.*

Consortium, G.P. *et al.* (2015) A global

reference for human genetic variation. *Nature.* 526, 68-74.

Cordero, P. and Das, R. (2015) Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis. *PLoS Comput Biol.* 11, e1004473.

Cordero, P. and Das, R. (2015) Rich RNA Structure Landscpes Revealed by Mutate-and-Map Analysis. *PLOS Computational Biology.* 11, e1004473.

Cordero, P. *et al.* (2012) An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics.* 28, 3006-8.

Corley, M. *et al.* (2015) Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic Acids Res.* 43, 1859-68.

Cruz, J.A. *et al.* (2012) RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA.* 18, 610-25.

Darty, K. *et al.* (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics.* 25, 1974-5.

Das, R. *et al.* (2008) Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci U S A.* 105, 4144-4149.

Das, R. *et al.* (2003) The fastest global events in RNA folding: electrostatic relaxation and tertiary collapse of the Tetrahymena ribozyme. *J Mol Biol.* 332, 311-9.

Das, R. *et al.* (2005) SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *Rna.* 11, 344-54.

Defays, D. (1977) An efficient algorithm for a complete link method. *The Computer Journal British Computer Society.* 20, 364-366.

Deigan, K.E. *et al.* (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A.* 106, 97-102.

Delfosse, V. *et al.* (2010) Riboswitch structure: an internal residue mimicking the purine ligand. *Nucleic Acids Res.* 38, 2057-68.

Deras, M.L. *et al.* (2000) Folding mechanism of the Tetrahymena ribozyme P4-P6 domain. *Biochemistry.* 39, 10975-85.

Dever, T.E. and Green, R. (2012) The Elongation, Termination and Recycling Phases of Translation in Eukaryotes. *Cold Spring Harbor Persepectives in Biology.* 4, a013706.

Diegan, K. *et al.* (2008) Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences, USA.* 106, 97-102.

Diegan, K.E. *et al.* (2009) Acccurate SHAPE-directed RNA Structure Determination. *Proceedings of the National Academy of Sciences USA.* 106, 97-102.

Ding, F. *et al.* (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA.* 14, 1164-1173.

Ding, Y. *et al.* (2005) A secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA.* 11, 1157-1166.

Ding, Y. *et al.* (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* 32, W135-41.

Ding, Y. *et al.* (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna.* 11, 1157-66.

Ding, Y. *et al.* (2006) Clustering of RNA Secondary Structures with Application to Messenger RNAs. *Journal of Molecular Biology.* 359, 554-571.

Ding, Y. and Lawrence, C. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research.* 31, 7280-7301.

Ding, Y. and Lawrence, C.E. (1999) A Bayesian statistical algorithm for RNA secondary structure prediction. *Computational Chemistry.* 23, 387-400.

Diri, B. and Albayrak, S. (2008) Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications.* 34, 628-634.

Dokholyan, N.V. *et al.* (1998) Discrete molecular dynamics studies of the folding of a protein-like mode. *Fold Des.* 3, 577-587.

Dokholyan, N.V. *et al.* (2011) Discrete molecular dynamics. *WIREs Comput Mol Sci.* 1, 80-92.

Dujardin, G. *et al.* (2014) How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Molecular Cell.* 54, 683-690.

Eddy, S. (2014) Computational Analysis of Conserved RNA Secondary Structure in Transcriptomes and Genomes. *Annual Review of Biophysics.* 43, 433-456.

Eddy, S.R. (2004) How do RNA folding algorithms work. *Nature Biotechnology.* 22, 1457-1458.

Eddy, S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 23, 205-11.

Eddy, S.R. (2014) Computational Analysis of Conserved RNA Secondary Structure in Transcriptomes and Genomes. *Annual Review of Biophysics.* 43, 433-456.

Ehresmann, C. *et al.* (1987) Probing the structure of RNAs in solution. *Nucleic Acids Research.* 15, 9109-9128.

Few, S. Data Visualization for Human Perception. Journal.

Frederiksen, J. *et al.* (2012) Metal-ion rescue revisited: biochemical detection of site-bound metal ions important for RNA folding.

. *RNA.* 18, 1123-1141.

Giegerich, R. *et al.* (2004) Abstract shapes of RNA. *Nucleic Acids Research.* 32, 4843-4851.

Gracia, B. *et al.* (2016) RNA Structural Modules Control the Rate and Pathway of RNA Folding and Assembly. *J Mol Biol.* 428, 3972-3985.

Guo, C. *et al.* (2013) ACTB in cancer. *Clinica Chimica Acta.* 417, 39-44.

Gutell, R.R. *et al.* (2002) The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology.* 12, 301-310.

Hall, M. *et al.* (2009) The WEKA Data Mining Software:An Update. *SIGKDD Explorations.* 11,

Halvorsen, M. *et al.* (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLOS Genetics.* 6, e1001074.

Halvorsen, M. *et al.* (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* 6, e1001074.

Hamada, M. *et al.* (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics.* 25, 465-73.

Harms, J. *et al.* (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell.* 107, 679-688.

Harris, M.E. *et al.* (1991) Regulation of histone mRNA in the unperturbed cell cycle: evidence suggesting control at two posttranscriptional steps. *Mol Cell Biol.* 11, 2416-2424.

Heilman-Miller, S.L. and Woodson, S.A. (2003) Effect of transcription on folding of the *Tetrahymena* ribozyme. *RNA.* 9, 722-733.

Herschlag, D. *et al.* (2015) From static to dynamic: the need for structural ensembles and a predictive model of RNA folding and function. *Curr Opin Struct Biol.* 30, 125-133.

Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Research.* 31, 3429-3431.

Hofacker, I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie.* 125, 167-188.

Holbrook, S. and Kim, S. (1997) RNA crystallography. *Biopolymers.* 44, 3-21.

Hua, W. *et al.* (2009) A Brief Review of Machine Learning and Its Application. Journal.

Hüttelmaier, S. *et al.* (2005) Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. *Nature.* 438, 512-515.

Ingola, N. *et al.* (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science.* 324, 255-258.

Jackson, R.J. *et al.* (2010) The mechanism of eukaryotic translation initation and principles of its regulation. *Nature Reviews Molecular Cell Biology.* 11, 113-127.

Karabiber, F. *et al.* (2013) QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA.* 19, 63-73.

Kertesz, M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature.* 467, 103-107.

Kertesz, M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature.* 467, 103-7.

Kierzek, R. *et al.* (1999) Thermodynamics of single mismatches in RNA duplexes. *Biochemistry.* 28, 14214-14223.

Kim, H.H. *et al.* (2015) Different motif requirements for the localization zipcode element of β-actin mRNA binding by HuD and ZBP1. *Nucleic Acids Research.* 43, 7432-7446.

Kislauskis, E. *et al.* (1994) Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *J Cell Biol.* 127, 441-451.

Kladwang, W. *et al.* (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA.* 17, 522-34.

Kladwang, W. *et al.* (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA.* 17, 522-534.

Kladwang, W. *et al.* (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature Chemistry.* 3, 952-962.

Kladwang, W. *et al.* (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem.* 3, 954-62.

Kladwang, W. *et al.* (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry.* 50, 8049-56.

Kohn, M.S. *et al.* (2014) IBM's Health Analytics and Clinical Decision Support. *Yearb Med Inform.* 9, 154-162.

Komar, A. (2009) A pause for thought alon gthe co-translational folding pathway. *Trends in Biochemical Sciences.* 34, 16-24.

Kubota, M. *et al.* (2015) Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology.* 11, 933-941.

Kuhn, M. (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software.* 28,

Kutchko, K.M. and Laederach, A. (2016) Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *WIRES RNA.*

Kutchko, K.M. and Laederach, A. (2016) Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *WIREs RNA.* 8, e1374.

Kutchko, K.M. *et al.* (2015) Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR. *RNA.* 21, 1274-85.

Lawrence, J. and Singer, R. (1986) Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell.* 45, 407-415.

Lee, J. *et al.* (2014) RNA design rules from a massive open laboratory. *Proc Natl Acad Sci U S A.* 111, 2122-7.

Lee, T.I. and Young, R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics.* 34, 77-137.

Lemay, J.F. *et al.* (2011) Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet.* 7, e1001278.

Lemay, J.F. and Lafontaine, D.A. (2007) Core requirements of the adenine riboswitch aptamer for ligand binding. *Rna.* 13, 339-50.

Lemay, J.F. *et al.* (2009) Molecular basis of RNA mediated gene regulation on the adenine riboswitch by single-molecule approaches. *Methods Molecular Biology.* 540, 65-76.

Lemay, J.F. *et al.* (2009) Molecular basis of RNA-mediated gene regulation on the adenine riboswitch by single-molecule approaches. *Methods Mol Biol.* 540, 65-76.

Lemay, J.F. *et al.* (2006) Folding of the adenine riboswitch. *Chemistry and Biology.* 13, 857-868.

Lemay, J.F. *et al.* (2006) Folding of the adenine riboswitch. *Chem Biol.* 13, 857-68.

Li, F. *et al.* (2012) Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell.* 24, 4346-4359.

Liaw, A. and Weiner, M. (2002) Classification and Regression by randomForest. *R News.* 2, 18-22.

Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News.* 2, 18-22.

Libbrecht, M. and Noble, W. (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics.* 16, 321-322.

Lipert, J. *et al.* (2007) Structural transitions and themodynamics of glycine-dependent riboswitch from *Vibrio* chlolerae. *Journal of Molecular Biology.* 365, 1393-1406.

Liu, Y. *et al.* (2015) Synthesis and applications of RNAs with position-selective labelling and mosaic composition. *Nature.* 522, 368-372.

Lokody, I. (2014) RNA: riboSNitches reveal heredity in RNA secondary structure. *Nat Rev Genet.* 15, 219.

Longfellow, C.E. *et al.* (1990) Themodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry.* 29, 278-285.

Lucks, J.B. *et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences, USA.* 108, 11063-11068.

Macke, T.J. *et al.* (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* 29, 4724-4735.

Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol.* 453, 3-31.

Martin, J.S. *et al.* (2012) Structural effects of linkage disequilibrium on the transcriptome. *Rna.* 18, 77-87.

Martin, J.S. *et al.* (2012) Structural effects of linkage disequilibrium on the transcriptome. *RNA.* 18, 77-87.

Mathews, D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna.* 10, 1178-90.

Matthews, D. (2006) Review: Revolutions in RNA Secondary Structure Prediction. *Journal of Molecular Biology.* 359, 526-532.

Matthews, D. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology.* 288, 911-940.

Matthews, D.H. (2004) Using an RNA Secondary Structure Partition Function to Determine Confidence in Base Pairs Predicted by Free Energy Minimization. *RNA.* 10, 1178-1190.

Matthews, D.H. *et al.* (1997) Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA.* 3, 1-16.

Matthews, D.H. *et al.* (2004) Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure. *Proceedings of the National Academy of Sciences USA.* 101, 7287-7293.

Matthews, D.H. *et al.* (2004) Incorporating chemical modification constraitns into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences, USA.* 101, 7287-7292.

Matthews, D.H. and Turner, D.H. (2002) Experimentally derived nearest neighbor parameters for the stability of RNA three and four-way multibranch loops. *Biochemistry.* 41, 869-880.

McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers.* 29, 1105-19.

McCaskill, J.S. (1990) The equilibrium partition function and base pair probabilities for RNA secondary structure. *BIopolymers.* 29, 1105-1119.

McGinnis, J.L. *et al.* (2012) The Mechanisms of RNA SHAPE Chemistry. *J. Am. Chem. Soc.* 134, 6617-6624.

McGinnis, J.L. *et al.* (2015) In-cell SHAPE reveals that free 30S ribosome subunits are in the inactive state. *Proceedings of the National Academy of Sciences USA.* 112, 2425-2430.

Merino, E.J. *et al.* (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc.* 127, 4223-4231.

Meyer, I. (2005) Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Research.* 33,

Miao, Z. *et al.* (2015) RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA.* 21, 1066-84.

Michel, F. *et al.* (2000) Modeling RNA tertiary structure from patterns of sequence variation. *Methods in Enzymology.* 317, 491-510.

Michel, F. and Westhof, E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol.* 216, 585-610.

Mitra, S. *et al.* (2011) RNA molecules with conserved catalytic cores but variable peripheries fold along unique energetically optimized pathways. *Rna.* 17, 1589-603.

Mitra, S. *et al.* (2008) High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Research.* 36, e63.

Mitra, S. *et al.* (2008) High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res.* 36, e63.

Molinaro, A. *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics.* 21, 3301-3307.

Moreno, N.N. *et al.* (2015) Chromatin, DNA structure and alternative splicing. *FEBS Letters.* 589, 3370-3378.

Mortimer, S. *et al.* (2012) SHAPE-Seq: High-Throughput RNA Structure Analysis. *Curr Protoc Chem Biol.* 4, 275-297.

Mortimer, S.A. *et al.* (2014) Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics.* 15, 469-479.

Mortimer, S.A. and Weeks, K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc.* 129, 4144-5.

Narayanan, B.C. *et al.* (2013) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Research.* 42, D114-122.

Noller, H.F. (2005) RNA structure: reading the ribosome. *Science.* 309, 1508-14.

Nussinov, R. and Jacobson, A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences, USA.* 77, 6309-6313.

Nussinov, R. *et al.* (1978) Algorithm for loop matchings. *SIAM Journal on Applied Mathematics.* 35, 68-82.

Pandey, N.B. and Marzluff, W.F. (1987) The stem-loop structure at the 3′ end of histone mRNA is necessary and sufficient for regulation of histone mRNA stability. *Mol Cell Biol.* 7, 4557-4559.

Patel, V.L. *et al.* (2012) Spatial arrangement of an RNA zipcode identifies mRNAs under post-transcriptional control. *GENES & DEVELOPMENT.* 26, 43-53.

Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine.* 2, 559-572.

Peattie, D. and Gilbert, W. (1980) Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences, USA.* 77, 4679-4682.

Petri, V. and Brenowitz, M. (1997) Quantitative nucleic acids footprinting: thermodynamic and kinetic approaches. *Curr Opin Biotechnol.* 8, 36-44.

Ponty, Y. (2008) Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: the boustrophedon method. *J Math Biol.* 56, 107-27.

Qi, Y. (2012) Random Forest for Bioinformatics. Journal. 307-323.

Quarrier, S. *et al.* (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *Rna.* 16, 1108-17.

Ramakrishnan, V. (2002) Ribosome structure and the mechanism of translation. *Cell.* 108, 557-572.

Raychaudhuri, S. *et al.* (2009) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium Biocomputing.* 455-466.

Regulski, E. and Breaker, R. (2008) In-line probing analysis of riboswitches. *methods in Molecular Biology.* 419, 53-67.

Rice, G.M. *et al.* (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA.* 20, 846-54.

Riddick, G. *et al.* (2011) Predicting *in vitro* drug sensitivity using Random Forests. *Bioinformatics.* 27, 220-224.

Ritz, J. *et al.* (2012) Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics.* 13, S6.

Ritz, J. *et al.* (2012) Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics.* 13 Suppl 4, S6.

Ritz, J. *et al.* (2013) Evolutionary evidence for alternative structure in RNA sequence co-variation. *PLoS Comput Biol.* 9, e1003152.

Rivas, E. and Eddy, S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics.* 2,

Rocca-Serra, P. *et al.* (2011) Sharing and archiving nucleic acid structure mapping data. *Rna.* 17, 1204-12.

Roh, J.H. *et al.* (2010) Multistage collapse of a bacterial ribozyme observed by time-resolved small-angle X-ray scattering. *J Am Chem Soc.* 132, 10148-54.

Ross, A. *et al.* (1997) Characterization of a beta-actin mRNA zipcode-binding protein. *Mol Cell Biol.* 17, 2158-2165.

Rouskin, S. *et al.* (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature.* 505, 701-5.

Rowles, T.A. (2013) Power to the people: does Eterna signal the arrival of a new wave of crowd-sourced projects? *BMC Biochem.* 14, 26.

Russell, R. *et al.* (2002) Rapid compaction during RNA folding. *Proc Natl Acad Sci U S A.* 99, 4266-71.

Russell, R. *et al.* (2002) Exploring the folding landscape of a structured RNA. *Proceedings of the National Academy of Sciences USA.* 99, 155-160.

Russell, R. *et al.* (2002) Exploring the folding landscape of a structured RNA. *Proc Natl Acad Sci U S A.* 99, 155-60.

Sabarinathan, R. *et al.* (2013) RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. *Hum Mutat.*

Sachs, A.B. *et al.* (1997) Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell.* 89, 831-838.

Saeys, Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007, 19.

Sakoe, H. and Chibe, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoust., Speech and Signal Process. .* 26, 43-49.

Salari, R. *et al.* (2013) Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. *Nucleic Acids Res.* 41, 44-53.

Sansone, S.A. *et al.* (2012) Toward interoperable bioscience data. *Nat Genet.* 44, 121-6.

Schroeder, S.J. *et al.* (1999) The energetics of small internal loops in RNA. *Biopolymers.* 52, 157-167.

Schultes, E.A. and Bartel, D.P. (2000) One sequence, two ribozymes: Implications for emergence of new ribozyme folds. *Science.* 289, 448-452.

Sclavi, B. *et al.* (1997) Time-resolved synchrotron X-ray "footprinting", a new approach to the study of nucleic acid structure and function: application to protein-DNA interactions and RNA folding. *J Mol Biol.* 266, 144-59.

Sclavi, B. *et al.* (2005) Real-time characterization of intermediates in the pathway to open complex formation by Escherichia coli RNA polymerase at the T7A1 promoter. *Proc Natl Acad Sci U S A.* 102, 4706-11.

Seeber, M. *et al.* (2011) Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J Comput Chem.* 32, 1183-1194.

Serganov, A. *et al.* (2015) Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing mRNAs. *Chem. Biol.* 11, 1729-1741.

Serganov, A. *et al.* (2004) Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem Biol.* 11, 1729-41.

Shabalina, S.A. *et al.* (2006) A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Research.* 8, 2428-2437.

Shannon, C.E. (1951) Prediction and Entropy of Printed English. *Bell System Technical Journal.* 30, 50-64.

Shapiro, B.A. (1988) An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.* 4, 387-393.

Shapiro, B.A. *et al.* (2001) RNA folding pathway functional intermediates: their prediction and analysis. *J Mol Biol.* 312, 27-44.

Shatkin, A.J. and Manley, J.L. (2000) The ends of the affair: Capping and polyadenylation. *Nature Structural Biology.* 7, 838-842.

Shcherbakova, I. *et al.* (2006) Fast Fenton footprinting: a laboratory-based method for the time-resolved analysis of DNA, RNA and proteins. *Nucleic Acids Research.* 34, e48.

Shcherbakova, I. *et al.* (2008) Energy barriers, pathways, and dynamics during folding of large, multidomain RNAs. *Curr Opin Chem Biol.* 12, 655-66.

Shirvanyants, D. *et al.* (2012) 116. *29.* 8375-8382,

Siegfred, N.A. *et al.* (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods.* 11, 959-965.

Siegfried, N.A. *et al.* (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods.* 11, 959-65.

Sigurdsson, S. *et al.* (1995) Probing RNA tertiary structure: interhelical crosslinking of the hammerhead ribozyme. *RNA.* 1, 575-583.

Silva, L. *et al.* (2008) Data classification with multilayer perceptrons using a generalized error function. *Neural Networks.* 21, 9.

Sinan, S. *et al.* (2011) The Azoarcus group I intron ribozyme misfolds and is accelerated for refolding by ATP-dependent RNA chaperone proteins. *J Biol Chem.* 286, 37304-12.

Smola, M.J. *et al.* (2015) Detection of RNA–Protein Interactions in Living Cells with SHAPE. *Biochemistry.* 54, 6867-6875.

Smola, M.J. *et al.* (2016) SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proceedings of the National Academy of Sciences USA.*

Solem, A.C. *et al.* (2015) The potential of the riboSNitch in personalized medicine. *Wiley Interdiscip Rev RNA.* 6, 517-32.

Solem, A.C. *et al.* (2015) The potential of the riboSNitch in personalized medicine. *WIREs RNA.* 6, 517-532.

Solomatin, S.V. *et al.* (2010) Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature.* 463, 681-684.

Spitale, R.C. *et al.* (2013) RNA SHAPE analysis in living cells. *Nature Chemical Biology.* 9, 18-22.

Steffen, P. *et al.* (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics.* 22, 500-503.

Steger, G. *et al.* (1984) Conformational transitions in viroids and virusoids: comparison of results from energy minimization algorithm and from experimental data. *Journal of Biomolecular Structure and Dynamics.* 2, 543-571.

Strambio-De-Castillia, C. *et al.* (2010) The nuclear pore complex: bridging nuclear transport and gene regulation. *Nature Reviews Molecular Cell Biology.* 11, 490-501.

Sun, J. *et al.* (1992) The histone mRNA 3' end is required for localization of histone mRNA to polyribosomes. *Nucleic Acids Research.* 25, 6057-6066.

Takamoto, K. *et al.* (2004) Principles of RNA compaction: insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations. *J Mol Biol.* 343, 1195-206.

Thirumalai, D. and Hyeon, C. (2005) RNA and protein folding: common themes and variations. *Biochemistry.* 44, 4957-70.

Tian, S. *et al.* (2014) High-throughput mutate-and-map rescute elevates SHAPE-directed RNA structure and uncovers excited states. *RNA.* 20, 1815-1826.

Tijerina, P. *et al.* (2007) DMS Footprinting of Structured RNAs and RNA-Protein Complexes. *Nature Protocols.* 2, 2608-2623.

Torgerson, W. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika.* 17, 401-419.

Touw, W. *et al.* (2013) Data mining in the LIfe Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics.* 14, 315-326.

Treuille, A. and Das, R. (2014) Scientific rigor through videogames. *Trends Biochem Sci.* 39, 507-9.

Tucker, B.J. and Breaker, R.R. (2005) Riboswitches as versatile gene control elements. *Current Opinion in Structural Biology.* 15,

Tullius, T. and Greenbaum, J. (2005) Mapping nucleic acid structure by hydroxyl radical cleavage. *Current Opinion in Structural Biology.* 9, 127-134.

Waldispuhl, J. and Clote, P. (2007) Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model. *J Comput Biol.* 14, 190-215.

Waldispuhl, J. and Reinharz, V. (2015) Modeling and predicting RNA three-dimensional structures. *Methods Mol Biol.* 1269, 101-21.

Wan, Y. *et al.* (2012) Genome-wide measurement of RNA folding energies. *Molecular Cell.* 48, 169-181.

Wan, Y. *et al.* (2012) Genome-wide measurement of RNA folding energies. *Mol Cell.* 48, 169-81.

Wan, Y. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature.* 505, 706-9.

Wan, Y. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature.* 505, 706-709.

Weeks, K. (2010) Advances in RNA structure analysis by chemical probing *Current Opinion in Structural Biology.* 20, 295-304.

Wilkinson, K.A. *et al.* (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols.* 1, 1610-1616.

Wilkinson, K.A. *et al.* (2009) Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *Rna.*

Williams, A.L. and Tinoco, I. (1986) Dynamic programming algorithm for finding alternative RNA secondary structures. *Nucleic Acids Research.* 14, 299-315.

Wilson, R.C. and Doudna, J.A. (2013) Molecular Mechanisms of RNA Interference. *Annual Review of Biophysics* 42, 217-239.

Wimberly, B.T. *et al.* (2000) Structure of the 30S ribosomal subunit. *Nature.* 407, 327-339.

Wolin, S. and Walter, P. (1988) Ribosome pausing nad stacking during translation of a eukaryotic mRNA. *The EMBO Journal.* 7, 3559-3569.

Woodson, S.A. (2000) Recent insights on RNA folding mechanisms from catalytic RNA. *Cellular and Molecular Life Sciences.* 57, 796-808.

Wu, B. *et al.* (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics.* 19, 1636-1643.

Wuchty, S. *et al.* (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers.* 49, 145-165.

Xia, T. *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson Crick paris. *Biochemistry.* 37,

Xue, S. *et al.* (2015) RNA regulons in Hox 5'UTRs confer ribosome specificity to gene regulation. *Nature.* 517, 33-38.

Yang, P. *et al.* (2010) A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics.* 5, 296-308.

Yonath, A. (2010) Hibernating Bears, Antibiotics and the Evolving Ribosome (Nobel Lecture). *Angewandte Chemie.* 49, 4340-4354.

Yoon, S. *et al.* (2011) HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics.* 27, 1798-805.

Zhang, W. *et al.* (2009) Structures of the ribosome in intermediate states of ratcheting. *Science.* 325, 1014-1017.

Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science.* 244,

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research.* 31, 3406-3415.

Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research.* 9, 133-148.