Insights into RNA structure by melding experiment and computation

Christine Elizabeth Hajdin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Department of Chemistry.

Chapel Hill
2013

Approved by:

Dr. Kevin Weeks

Dr. Nikolay Dokholyan

Dr. Dorothy Erie

Dr. Howard Fried

Dr. Alain Laederach

Dr. Ron Swanstrom

ABSTRACT

CHRISTINE ELIZABETH HAJDIN: Insights into RNA structure by melding experiment and computation
(Under the direction of Kevin Weeks)

The ability of RNA to perform diverse cellular functions depends on its capability to form complex structures. Therefore, determining RNA structure is critical to understanding RNA function. Computational methods allow for quick determination of RNA structures, but are often prone to inaccuracies in their predictions. A newly developed technology, known as SHAPE, can be used to probe RNA structure and identify nucleotides that are likely to be single stranded and base paired[7]. This SHAPE data can be inputted into an RNA structure program to refine predictions. Previous studies have shown that the incorporation of SHAPE data can increase the accuracy of prediction by over 30% compared to traditional mFold class algorithms[26]. In this work, I utilize SHAPE technology to refine RNA predictions and solve new challenges. First, I create an algorithm, ShapeKnots, which incorporates SHAPE data and the prediction of pseudoknots. Pseudoknots are relatively rare RNA structural motifs that have a tendency of occurring in functional regions, but, due to their complexity, are often eliminated from structural prediction. Second, I utilize the ShapeKnots algorithm to identify pseudoknots in HIV-1 and test their role in viral replication. Third, I develop a modified partition function calculation to identify the *de novo* accuracy of secondary structure predictions.

This allows end users to not only obtain a predicted structure, but also, to know the confidence of that prediction. Fourth, I utilize SHAPE-directed folding to identify potential alternative structures in the ribosome. Finally, I create a method to identify the accuracy of tertiary structure predictions. This allows for a quantitative measurement of accuracy when comparing predicted tertiary structures with previously determined conventional structures.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

1M6 - 1-methyl-6-nitroisatoic anhydride

1M7 - 1-methyl-7-nitroisatoic anhydride

di-GMP –diguanylate

DIS – Dimer initiation site

DMD - Discrete molecular dynamics

DNA – Deoxyribonucleic acid

GDT – Global distance test

HCV - Hepatitis C virus

HIV-1 – Human immunodeficiency virus type 1

INF – Interaction network fidelity

IRES – Internal ribosome entry site

LNA – Locked nucleic acid

NE – Nested helices

NL – Inline helices

NMIA - N-methylisatoic anhydride

NMR - Nuclear magnetic resonance

MFE – Minimum free energy

mRNA – Messenger ribo-nucleic acid

PBS – Primer binding site

PreQ1- Pre-queuosine

RCSB - Research Collaboratory for Structural Bioinformatics

RMSD – Root mean square deviation

RNA – Ribo-nucleic acid

RNase P – Ribonuclease P

RRE – Rev response element

rRNA – Ribosomal ribo-nucleic acid

SAM-I – S'adenosyl methionine

SARS – Severe acute respiratory syndrome

SHAPE - Selective 2'OH acylation analyzed by primer extension

SRP – Signal recognition particle

SS – Single stranded

TAR - Trans-activation response

tat- Trans-activator of transcription

tRNA – Transfer ribo-nucleic acid

TPP – Thiamine pyrophosphate

UTR – Untranslated region

Chapter 1: Introduction

1.1 RNA structure and function

Although ribonucleic acid (RNA) is often dismissed as a passive component of translation, RNA plays keys roles in viruses and cells. For example, viral RNAs, like the Human immunodeficiency virus type 1 (HIV-1) RNA, form important elements of structure essential for viral replication and transcription[1]. Riboswitch RNAs, like the thiamine pyrophosphate (TPP) riboswitch, regulate cellular function[2-4] and ribosomal RNAs are critical to translation. For instance the 16S and 23S rRNA found in *E. coli* form the main structural component of the ribosome and coordinate between the mRNA, the tRNA and the elongating amino acid chain[5, 6].

The ability of RNA to perform these multiple diverse functions depends on its capacity to form distinct structures. Identifying these distinct RNA structures is critical to understanding and characterizing the role of RNA in cells and viruses.

To date, the most accurate way of determining RNA structure is to use high-resolution three-dimensional structural probing techniques like X-ray crystallography[7-9]. X-ray crystallography works by irradiating a crystalized RNA with beams of X-rays creating a series of diffraction patterns. These diffraction patterns can be mathematically transformed into an electron density map that can be used to model RNA structures. X-ray crystallography has been successfully used to determine the structure of many RNAs,

like the signal recognition particle (SRP)[10], lysine riboswitch[11], and even the 16S rRNA [12, 13].

Despite the advancements of X-ray crystallography, it is not suitable for every RNA. To accurately identify a tertiary structure, the RNA must be able to form a well-ordered crystal. Without a well-ordered crystal, the diffraction data becomes 'fuzzy' and it is hard to model the electron density. Crystallization, or the act of making an RNA crystal, is a difficult and time-consuming process. Success depends on the specific RNA being tested, the concentration of different components in the solution, the pH, and the flexibility of the RNA. To aid in the crystallization process, and increase the stability of the RNA, high concentrations of proteins, ligands and other stabilizing elements are added. These conditions can perturb the RNA structure and shift it away from its lowest energy state which is adopted in solution[13-15].

1.2 Using computational algorithms to determine RNA structure

Due to the disadvantages of traditional structure probing techniques, computational algorithms have been developed as a useful alternative for RNA structure determination. However, predicting the tertiary structure of an RNA directly from a linear sequence is a difficult challenge. An important first step in this process is to determine the secondary structure.

Minimization of free energy (MFE) is one of the most popular methods for secondary structure prediction. Using previously established Turner energy rules; the algorithm folds a linear sequence of the RNA into potential structures[16]. For RNA secondary structure prediction, free energy parameters for basic structural motifs are estimated or extrapolated from chemical melting experiments[17]. The energy associated

with all motifs in a structure is summed to identify the energy of a potential structure. The potential structures are then sorted by their energy and the "correct" structure is identified to be the lowest energy structure. These MFE algorithms, classically called the mFold class of algorithms[18], tend to work well for small RNAs, but suffer from inaccuracies due to incomplete energy rules and an inability to correctly rank structures with similar energies, (see Chapter 2) so do not work well for long or complicated RNAs.

Heuristic algorithms work in a similar fashion to MFE algorithms, but attempt to redefine energy rules by supplementing additional constraints from known structures[19, 20]. These constraints are implemented into the program using a series of fit equations that are optimized against known RNAs. These algorithms tend to work well for small RNAs that can be accurately fit with a small number of equations and parameters. However, the complexity of large RNAs requires additional equations and constraints. Since the number of known large RNA structures is small and biased toward those that are stable enough for crystallography, it results in over-optimized fits to a few RNAs.

Partition function algorithms use statistical characterizations of the equilibrium ensemble of RNA to determine secondary structures[21, 22]. They function by calculating the base pairing probability of each base pair combination in an RNA and then use this information to rank order potential structures. Like traditional MFE algorithms, partition function algorithms employ classic rules from thermodynamics to assign probabilities. As in the MFE algorithm, this tends to work well for small RNAs, but is not accurate for larger RNAs.

Co-variation algorithms provide yet another popular way of determining structure[23]. Co-variation measures the number of instances that base pairing ability is

maintained when bases in the pair are mutated. For example, if one serotype of a virus has a predicted CG base pair, co-variation would be observed if in another serotype there is an AU pair in the same relative position. When a large number of homologous sequences are known and there is great variability in the sequences, this method, can provide accurate RNA secondary structures. Many structures have been solved by this method such as the 16S rRNA[24, 25]. However, if there are not a large number of sequences known, few base pairs can be correctly identified.

1.3 Using SHAPE data to refine structural predictions

Previously, work demonstrated how incorporating experimental SHAPE data can help to improve the accuracy of computational algorithms[26]. SHAPE allows for the differentiation of single-stranded and base-paired nucleotides[27] (Figure 1.1). The technique works by chemically foot-printing an RNA using a SHAPE reagent. This molecule preferentially reacts with single-stranded nucleotides, creating bulky 2'-O-adducts. These adducts can be probed using a reverse transcriptase, which dissociates when it encounters the 2'-O-adduct. This creates a series of cDNAs whose lengths correspond to the position of modification and whose abundance corresponds to the degree of modification. These cDNAs are resolved using capillary electrophoresis and aligned to a sequencing ladder[27]. After integration, background subtraction from a no-reagent control reaction, and further data processing the end result is a SHAPE reactivity profile (Figure 1.1).

Since the SHAPE reagents preferentially react at single-stranded positions, the SHAPE reactivity can be used to inform base pairing. This can severely limit the

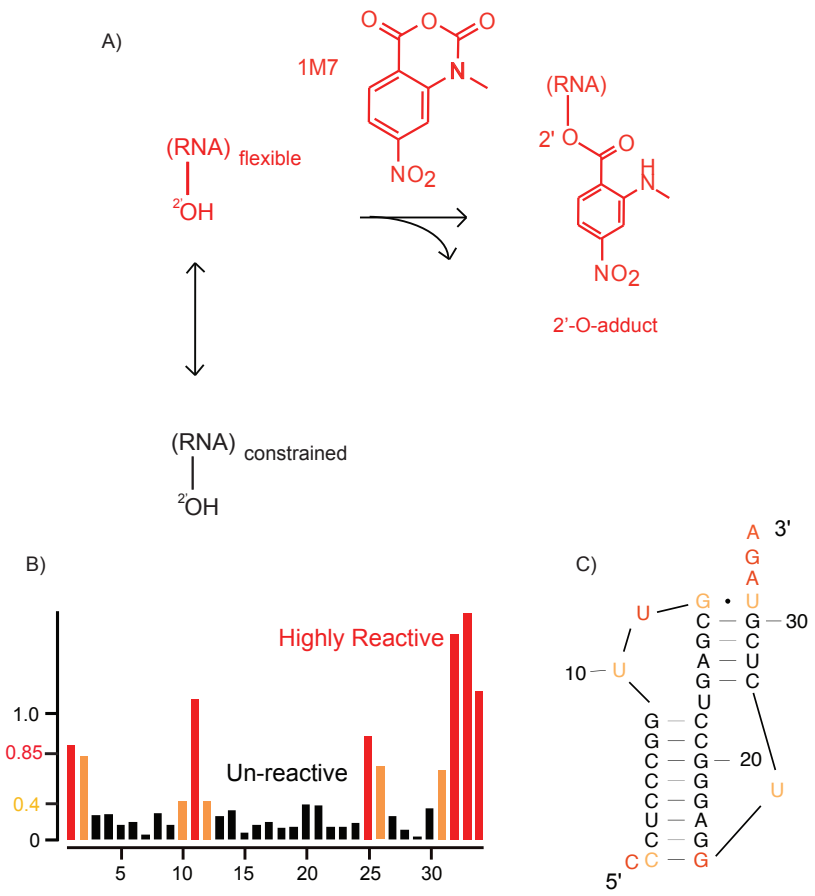

Figure 1.1: Overview of SHAPE mechanism

A) Schematic of SHAPE reagent (1M7) reacting with RNA. The SHAPE reagent preferentially reacts with single stranded nucleotides forming bulky 2'-O-adducts. These adducts can be detected using capillary electrophoresis.
B) After data processing, the SHAPE data can be viewed as a SHAPE reactivity profile. The profile plots the nucleotide sequence along the x-axis and the SHAPE reactivities along the y-axis. The reactivates are colored so that very highly reactive nucleotides (> 0.85) are colored red, highly reactivity orange (between 0.4 and 0.85), and lowly reactive (< 0.4) are colored black.
C) When the SHAPE data is superimposed on an RNA structure, single stranded nucleotides are usually red or orange indicating highly flexible and base paired nucleotides are colored back, indicating highly constrained.

sampling space and increase the accuracy of prediction. When this method was incorporated into a secondary structure prediction algorithm, RNAstructure, and applied to the 16S rRNA and 23S rRNA, the accuracy of the prediction of structure for these RNAs increased by more than 20% over a traditional mFold class algorithm [26].

1.4 Challenges of prediction algorithms

The RNAstructure algorithm provided a critical first step in refining traditional dynamic programming, but despite its advances, RNAstructure still had several deficiencies. For instance, none of the traditional RNAstructure predictions allows for the prediction of pseudoknots. Pseudoknots form when the loop region of a helix base pairs to another place in a RNA structure. Figure 1.2 shows the secondary structure of a simple pseudoknot on a traditional secondary structure plot and on a circleplot. A circleplot plots the sequence of the RNA around the outside of the circle and base pairs as lines running through the circle. Pseudoknots are easily identified on circleplots because they form a cross-hatching pattern. Pseudoknot motifs are relatively rare, but often occur in key functional areas, such as the pseudoknot near the 5' end of HIV-1 that allows for frame shifting[28], and the central pseudoknot in the SAMI riboswitch[29] necessary for ligand binding. Because of their biological importance, there is a need to confidently identify pseudoknots in RNA secondary structures[30, 31].

Furthermore, although predictions with SHAPE data tend to be highly accurate, mistakes in the structure are hard to distinguish and identify. It is not clear from the traditional version of the program which parts of the structures are most likely to be correct or incorrect. Being able to identify mistakes is structural predictions can be as critical as the prediction. Knowing that a structure prediction is highly accurate enables

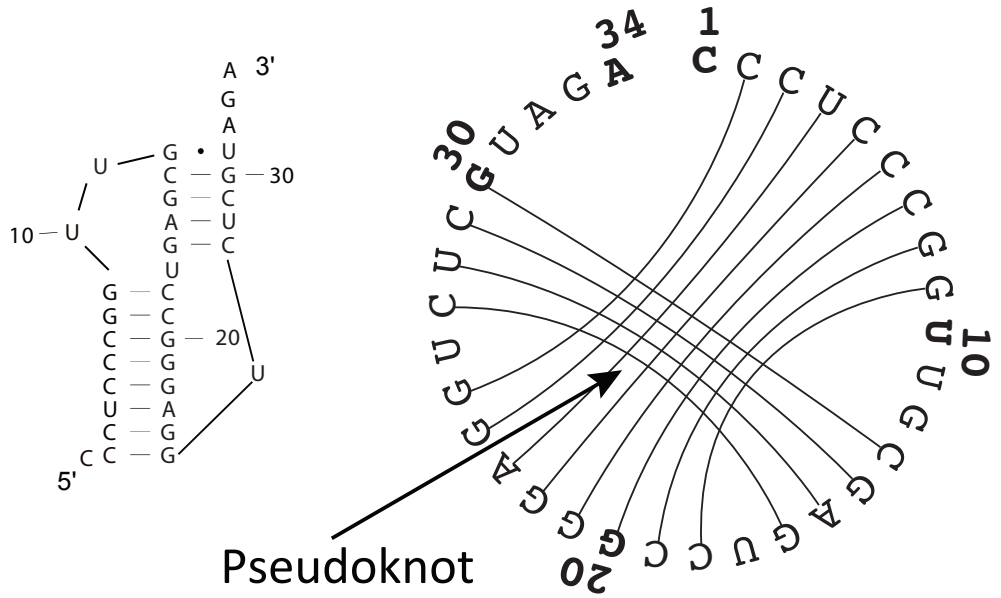

Figure 1.2: Simple pseudoknot motif in RNA.

On the left, the pseudoknot is shown on a traditional secondary structure plot. On the right, the pseudoknot is shown on a circleplot. Circleplots plot the sequence of the RNA around the circumference of the circle; lines running through the circle represent base pairs. Pseudoknots are easy to identify because they create a cross hatching pattern.

key hypotheses to be made. Conversely, identifying a poorly supported structure allows incorrect hypotheses to be avoided. Lastly, previous work has generally focused on using SHAPE data to understand secondary structure predictions. To obtain a full understanding of RNA structure, we must eventually target tertiary structures.

1.5 Research Overview

In this work, I sought to modify current algorithms to increase prediction accuracy and better understand RNA structure. My focus has been on finding techniques to include pseudoknots using SHAPE data in secondary structure predictions, and identifying methods to calculate the accuracy of structure predictions.

In Chapter 2, I discuss how we created a secondary structure prediction algorithm, ShapeKnots, which more accurately incorporates SHAPE reactivities and predicts pseudoknots. By incorporating these features, I show that the accuracy of prediction increases 30% over classic mFold class algorithms to reach 94% accuracy. In Chapter 3, I discuss how I can use the ShapeKnots algorithm to identify pseudoknots in HIV-1. I show that when tested *in virio*, these pseudoknots are critical to the replication of the HIV-1. In Chapter 4, I discuss how I can determine the accuracy of secondary structure prediction by calculating the Shannon entropy from a modified partition function.

Chapters 5 and 6 are devoted to tertiary structure modeling. In Chapter 5, I examine an alternative structure of the 3' minor domain of the 16S rRNA. In particular, I use Discrete Molecule Dynamics (DMD)[32] to identify whether or not the alternative structure is topologically consistent with the conventional 16S rRNA structure[33] and with the X-ray crystallography electron density[12]. Finally in Chapter 6, I discuss how we can

use the central limit theorem to identify a useful metric for categorizing the success of tertiary structure predictions. This metric, denoted the "q value", can be used to evaluate tertiary structure prediction quality.

1.6 Perspective

In this work, I utilize experimental and computational principles to refine RNA structure. I show that with these refined technologies, I am able to increase the accuracy of structure prediction, identify unique motifs in HIV-1, better understand the accuracy of secondary structure predictions, apply information identified from SHAPE-directed secondary structure predictions to identify potential alternative structures and create a useful metric for determining the accuracy of tertiary structures. It is my hope that the methods I present will be widely useful in refining highly accurate RNA structure models in the future.

1.7 References

1. Frankel, A.D. & Young, J.A. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem* **67**, 1-25 (1998).

2. Serganov, A., Polonskaia, A., Phan, A.T., Breaker, R.R. & Patel, D.J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167-1171 (2006).

3. Bocobza, S. et al. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes Dev* **21**, 2874-2879 (2007).

4. Kubodera, T. et al. Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR. *FEBS Lett* **555**, 516-520 (2003).

5. Traub, P. & Nomura, M. Structure and function of E. coli ribosomes. V. Reconstitution of functionally active 30S ribosomal particles from RNA and proteins. *Proc Natl Acad Sci U S A* **59**, 777-784 (1968).

6. Shajani, Z., Sykes, M.T. & Williamson, J.R. Assembly of bacterial ribosomes. *Annu Rev Biochem* **80**, 501-526 (2011).

7. Wilkinson, K.A. et al. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol* **6**, e96 (2008).

8. Mooers, B.H. Crystallographic studies of DNA and RNA. *Methods* **47**, 168-176 (2009).

9. Smyth, M.S. & Martin, J.H. x ray crystallography. *Mol Pathol* **53**, 8-14 (2000).

10. Hainzl, T., Huang, S. & Sauer-Eriksson, A.E. Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. *Nature* **417**, 767-771 (2002).

11. Garst, A.D., Heroux, A., Rambo, R.P. & Batey, R.T. Crystal structure of the lysine riboswitch regulatory mRNA element. *Journal of Biological Chemistry* **283**, 22347-22351 (2008).

12. Dunkle, J.A., Xiong, L., Mankin, A.S. & Cate, J.H. Structures of the Escherichia coli ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. *Proc Natl Acad Sci U S A* **107**, 17152-17157 (2010).

13. Sagi, I. et al. Crystallography of ribosomes: attempts at decorating the ribosomal surface. *Biophys Chem* **55**, 31-41 (1995).

14. Lu, J., Li, N.S., Sengupta, R.N. & Piccirilli, J.A. Synthesis and biochemical application of 2'-O-methyl-3'-thioguanosine as a probe to explore group I intron catalysis. *Bioorg Med Chem* **16**, 5754-5760 (2008).

15. Egli, M. & Pallan, P.S. Insights from crystallographic studies into the structural and pairing properties of nucleic acid analogs and chemically modified DNA and RNA oligonucleotides. *Annu Rev Biophys Biomol Struct* **36**, 281-305 (2007).

16. Zhang, S.J. et al. RhesusBase: a knowledgebase for the monkey research community. *Nucleic Acids Res* **41**, D892-905 (2013).

17. Tian, B. et al. Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *Rna* **6**, 79-87 (2000).

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

19. Westhof, E., Masquida, B. & Jaeger, L. RNA tectonics: towards RNA design. *Fold Des* **1**, R78-88 (1996).

20. Ren, J., Rastegari, B., Condon, A. & Hoos, H.H. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *Rna* **11**, 1494-1504 (2005).

21. Vreede, F.T., Chan, A.Y., Sharps, J. & Fodor, E. Mechanisms and functional implications of the degradation of host RNA polymerase II in influenza virus infected cells. *Virology* **396**, 125-134 (2010).

22. Dong, H., Ding, L., Yan, F., Ji, H. & Ju, H. The use of polyethylenimine-grafted graphene nanoribbon for cellular delivery of locked nucleic acid modified molecular beacon for recognition of microRNA. *Biomaterials* **32**, 3875-3882 (2011).

23. Eddy, S.R. & Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res* **22**, 2079-2088 (1994).

24. Tocilj, A. et al. The small ribosomal subunit from Thermus thermophilus at 4.5 A resolution: pattern fittings and the identification of a functional site. *Proc Natl Acad Sci U S A* **96**, 14252-14257 (1999).

25. Schluenzen, F. et al. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* **102**, 615-623 (2000).

26. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).

27. Wilkinson, K.A. et al. Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *Rna* **15**, 1314-1321 (2009).

28. Gaudin, C. et al. Structure of the RNA signal essential for translational frameshifting in HIV-1. *J Mol Biol* **349**, 1024-1035 (2005).

29. Montange, R.K. & Batey, R.T. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441**, 1172-1175 (2006).

30. Brierley, I., Gilbert, R.J. & Pennell, S. RNA pseudoknots and the regulation of protein synthesis. *Biochemical Society transactions* **36**, 684-689 (2008).

31. Brierley, I., Pennell, S. & Gilbert, R.J. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol* **5**, 598-610 (2007).

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

33. Woese, C.R., Gutell, R., Gupta, R. & Noller, H.F. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiological reviews* **47**, 621-669 (1983).

Chapter 2: ShapeKnots: accurate RNA secondary structure predictions, including pseudoknots

2.1 Introduction

2.1.1 RNA structure and function.

RNA constitutes the central information conduit in biology[1]. Information is encoded in an RNA molecule at two levels: in its primary sequence and in its ability to form higher-order secondary and tertiary structures. Nearly all RNAs can fold to form some secondary structure and, in many RNAs, highly structured regions encode important regulatory motifs. Such structured regulatory elements can be comprised of canonical base pairs but may also feature specialized and distinctive RNA structures. Among the best characterized of these specialized structures are RNA pseudoknots. Pseudoknots are relatively rare but occur overwhelmingly in functionally important regions of RNA[2-4]. For example, all of the large catalytic RNAs contain pseudoknots[5, 6]; roughly two-thirds of the known classes of riboswitches contain pseudoknots that appear to be essential for ligand binding and gene regulatory functions[7]; and pseudoknots occur prominently in the regulatory elements that viruses use to usurp cellular metabolism[3]. Pseudoknots are thus harbingers of biological function. An important and challenging goal is to identify these structures reliably.

2.1.2 Pseudoknots in RNA structure predictions.

Pseudoknots are excluded from the most widely employed algorithms used to model RNA secondary structure [8]. This exclusion is based on the challenge of incorporating the pseudoknot structure into the efficient dynamic programming algorithm used in the most popular secondary structure prediction approaches and because of the additional computational effort required. The prediction of lowest free energy structures with pseudoknots is NP-complete[9], which means that lowest free energy structure cannot be solved as a function of sequence length in polynomial time. In addition, allowing pseudoknots greatly increases the number of (incorrect) helices possible and tends to reduce secondary structure prediction accuracies, even for RNAs that include pseudoknots. Current algorithms also have high false positive rates for pseudoknot prediction, necessitating extensive follow-up testing and analysis of proposed structures.

Pseudoknot prediction is challenging, in part, for the same reasons that RNA secondary structure prediction is difficult. First, energy models for loops are incomplete because they extrapolate from a limited set of experiments. Second, folding can be affected by kinetic, ligand-mediated, tertiary, and transient interactions that are difficult or impossible to glean from the sequence. Prediction is also difficult for a third reason unique to pseudoknots: Energy models for pseudoknot formation are generally incomplete because the factors governing their stability are not fully understood[10-12]. The result is that current algorithms that model pseudoknots predict the base pairs in the simplest pseudoknots (termed H-type, formed when bases in a loop region bind to a single-stranded region), when the beginning and end of the pseudoknotted structure is known, with accuracies of only about 75% [10]. Secondary structure prediction is much less

accurate for full-length biological RNA sequences, with as few as 5% of known pseudoknotted pairs predicted correctly and with more false positive than correct pseudoknot predictions in some benchmarks [13].

2.1.3 Using SHAPE data to probe RNA structure.

The accuracy of secondary structure prediction is improved dramatically by including experimental information as restraints[14, 15]. SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) probing data has proven especially useful in yielding robust working models for RNA secondary structure [15, 16]. In essence, inclusion of SHAPE information provides an experimental adjustment to the well-established, nearest neighbor model parameters[17] for RNA folding. This adjustment is implemented as a simple pseudo-free energy change term, $\Delta G^{\circ}_{\mathrm{SHAPE}}$. SHAPE reactivities are approximately inversely proportional to the probability that a given nucleotide is base paired (high reactivities correspond to a low likelihood of being paired and vice versa) and the logarithm of a probability corresponds to an energy, in this case $\Delta G^{\circ}_{\mathrm{SHAPE}}$, which has the form:

$$\Delta G^{\circ}_{\mathrm{SHAPE}} = m\, ln\, [\mathrm{SHAPE} + 1] + b \quad (1)$$

The slope, m, corresponds to a penalty for base pairing that increases with the experimental SHAPE reactivity, and the intercept, b, reflects a favorable pseudo-free energy change term for base pairing at nucleotides with low SHAPE reactivities. These two parameters must be determined empirically. This pseudo-free energy change approach yields high-quality secondary structure models for both short RNAs and those that are kilobases long[15, 16].

Our original SHAPE-directed algorithm did not allow for pseudoknotted base pairs[15]. Given the strong relationship between pseudoknots and functionally critical regions in RNA and the fact that it is impossible to know *a priori* whether an RNA contains a pseudoknot, this limitation severely restricts the accuracy and generality of experimentally-directed RNA structure analysis. Here, I describe a concise approach for applying SHAPE-directed RNA secondary structure modeling to include pseudoknots, in an algorithm I call ShapeKnots, and I show that the algorithm yields high quality structures for diverse RNA sequences.

2.2 Results

2.2.1 A challenging RNA test set.

We developed the ShapeKnots algorithm using a test set of 16 non-pseudoknotted and pseudoknot-containing RNAs that were selected for their complex, and generally difficult to predict, structures (Table 2.1, top). These RNAs included (*i*) five RNAs with lengths >300 nucleotides, both with and without pseudoknots; (*ii*) five riboswitch RNAs whose structures only form upon binding by specific ligands, for which thermodynamic rules are obligatorily incomplete; (*iii*) four RNAs with structures that are predicted especially poorly, with accuracies <60% using nearest-neighbor thermodynamic parameters; and (*iv*) three RNAs whose structures are probably modulated by protein binding. SHAPE experiments were performed on each of the RNAs in the presence of ligand if applicable but in the absence of any protein. Each of the training set RNAs had SHAPE probing patterns that suggested these RNAs folded in solution into structures generally consistent with accepted secondary structure models based on either X-ray crystallography or comparative sequence analyses.

Allow pseudoknots				–				–				+				+			
SHAPE data				–				+				–				+			
	Length	Features	PKs	sens	ppv	geo	PK	sens	ppv	geo	PK	sens	ppv	geo	PK	sens	ppv	geo	PK
Training set																			
Pre-Q1 riboswitch, *B. subtilis*	34	L	1	62.5	100	79.1	x	62.5	100	79.1	x	62.5	100	79.1	x	100	100	100	✓
Telomerase pseudoknot, human	47	P	1	40.0	75.0	54.8	x	60.0	75.0	67.1	x	100	100	100	✓	100	100	100	✓
tRNA(asp), yeast	75	–	0	95.2	95.2	95.2	✓	95.2	95.2	95.2	✓	95.2	95.2	95.2	✓	95.2	95.2	95.2	✓
TPP riboswitch, *E. coli*	79	L	0	77.3	85.0	81.0	✓	95.5	87.5	91.4	✓	77.3	85.0	81.0	✓	95.5	87.5	91.4	✓
SARS corona virus pseudoknot	82	–	1	69.2	90.0	78.9	x	69.2	75.0	72.1	x	65.4	68.0	66.7	x	84.6	88.0	86.3	✓
cyclic-di-GMP riboswitch, *V. cholerae*	97	L	0	75.0	77.8	76.4	✓	89.3	86.2	87.7	✓	75.0	77.8	76.4	✓	89.3	86.2	87.7	✓
SAM I riboswitch, *T. tengcongensis*	118	L	1	74.4	80.6	77.4	x	76.9	85.7	81.2	x	76.9	81.1	79.0	x	92.3	97.3	94.8	✓
M-Box riboswitch, *B. subtilis*	154	L	0	87.5	91.3	89.4	✓	87.5	91.3	89.4	✓	87.5	91.3	89.4	✓	87.5	91.3	89.4	✓
P546 domain, bI3 group I intron	155	–	0	42.9	44.4	43.6	✓	94.6	96.4	95.5	✓	42.9	44.4	43.6	✓	94.6	96.4	95.5	✓
Lysine riboswitch, *T. maritima*	174	L	1	77.8	83.1	80.4	x	79.4	89.3	84.2	x	84.1	82.8	83.5	✓	87.3	88.7	88.0	✓
Group I intron, *Azoarcus* sp.	214	–	1	73.0	75.4	74.2	x	81.0	85.0	83.0	x	73.0	75.4	74.2	x	92.1	95.1	93.6	✓
Hepatitis C virus IRES domain	336	–	1	39.4	38.0	38.7	x	79.8	86.5	83.1	x	39.4	36.3	37.8	x	92.3	96.0	94.1	✓
Group II intron, *O. iheyensis*	412	–	1	88.6	97.5	92.9	x	74.2	84.5	79.2	x	88.6	97.5	92.9	x	93.2	97.6	95.4	✓
Group I Intron, *T. thermophila*	425	–	1	83.2	74.3	78.6	x	87.8	88.6	88.2	x	83.2	74.3	78.6	x	93.9	91.2	92.5	✓
5' domain of 23S rRNA, *E. coli* †	511	L,P	0	97.2	73.8	84.7	✓	97.2	76.8	86.4	✓	97.2	73.8	84.7	✓	97.2	76.8	86.4	✓
5' domain of 16S rRNA, *E. coli* †	530	L,P	0	63.6	59.1	61.3	✓	93.0	83.6	88.2	✓	63.6	59.1	61.3	✓	93.0	83.6	88.2	✓
Average				**71.7**	**77.5**	**74.2**		**82.7**	**86.7**	**84.4**		**75.7**	**77.6**	**76.5**		**93.0**	**91.9**	**92.4**	
Test set																			
Fluoride riboswitch, *P. syringae*	66	L	1	56.3	64.3	60.1	x	62.5	71.4	66.8	x	93.8	93.8	93.8	✓	93.8	93.8	93.8	✓
Adenine riboswitch, *V. vulnificus*	71	L	0	100	100	100	✓	100	100	100	✓	100	100	100	✓	100	100	100	✓
tRNA(phe), *E. coli*	76	–	0	95.2	100	97.6	✓	100	100	100	✓	95.2	100	97.6	✓	100	84.0	91.7	x
5S rRNA, *E. coli*	120	L,P	0	28.6	25.0	26.7	✓	85.7	76.9	81.2	✓	28.6	25.0	26.7	✓	85.7	76.9	81.2	✓
5' domain of 16S rRNA, *H. volcanii* †	473	L,P	0	85.6	71.9	78.5	✓	96.2	83.2	89.5	✓	85.6	71.9	78.5	✓	96.2	83.2	89.5	✓
HIV-1 5' pseudoknot domain §	500	–	1				x				x				x				✓
Average				**73.1**	**72.2**	**72.6**		**88.9**	**86.3**	**87.5**		**80.6**	**78.1**	**79.3**		**95.1**	**87.6**	**91.2**	
Overall				**72.0**	**76.3**	**73.8**		**84.2**	**86.6**	**85.2**		**76.9**	**77.7**	**77.1**		**93.5**	**90.9**	**92.1**	
Reactivities incompatible with accepted structures																			
Signal recognition particle RNA, human	301	L,P	0	0	0	0	✓	59.0	59.0	59.0	✓	0	0	0	✓	55.0	53.9	54.4	x
RNase P, *B. subtilis*	405	L,P	1	57.4	55.0	56.2	x	76.5	81.5	79.0	x	57.4	55.0	56.2	x	75.7	79.8	77.7	x

Accuracy (%): 100 High; 90; 80; 70; 60; 50; 40; 30; 20; 10; 0 Low

Table 2.1: Prediction accuracies as a function of algorithm and SHAPE information.

Sensitivities (sens), positive predictive value (ppv), and their geometric average (geo) are shown for four test cases: no pseudoknots allowed and no SHAPE data; no pseudoknots allowed and with SHAPE data (both by free energy minimization); pseudoknots allowed and no SHAPE data; and pseudoknots allowed and with SHAPE data (both using ShapeKnots). Complicating features are ligand (L) and protein (P) binding that are not accounted for in nearest-neighbor thermodynamic parameters. Pseudoknot (PK) predictions are indicated with a checkmark (✓) or X; a checkmark indicates that pseudoknots were predicted correctly and that there were no false-positive pseudoknot predictions. For the ribosomal RNAs (†), regions in which the SHAPE reactivities were clearly incompatible with the accepted structure, as described [18], were omitted from the sensitivity and ppv calculations; for the *E. coli* 16 rRNA, this included nucleotides 143-220. The HIV-1 5' leader domain (§) was included as an example of pseudoknot prediction in a large RNA. Because the accepted structure for this RNA is based on SHAPE-directed prediction [19], we did not include sensitivity and ppv for this RNA in the overall Average values; however, the pseudoknot was proven independently [20] and is included.

The structures of the 16 RNAs in the test set are predicted poorly by a conventional algorithm based on their sequences alone: The average sensitivity (sens; fraction of base pairs in the accepted structure predicted correctly), positive predictive value (ppv, the fraction of predicted pairs that occur in the accepted structure), and geometric average of these metrics are 72, 78, and 74%, respectively (Table 2.1).

In the process of developing this training set, we also analyzed two RNAs – RNase P RNA and the human signal recognition particle RNA – whose *in vitro* SHAPE reactivities were incompatible with the accepted structures for these RNAs. I include prediction statistics for these RNAs at the bottom of Table 2.1, but do not use these to evaluate our SHAPE-directed modeling algorithm.

2.2.2 A simple, robust model for pseudoknot formation.

The favorable energetic contributions for forming the helices that comprise a pseudoknot are likely to be predicted accurately by the Turner nearest-neighbor model [17, 21] when modified by the experimental $\Delta G°_{SHAPE}$ term (Eqn. 1). In addition, pseudoknot formation must overcome an entropic penalty; these energetics are difficult to estimate. The most widely used models are complex and include a large number of constituent parameters[11, 12]. We adopted a simple approach to estimate the entropies based on three primary insights. First, any secondary structure prediction must ultimately be compatible with a specific, energetically favorable, three-dimensional fold in the RNA in which nucleotides that base pair in the pseudoknot are close in three-dimensional space. This fundamental close-in-space feature must also be recapitulated in secondary structure prediction.

We modeled RNA pseudoknots as the sum of simple distance features, or beads. There are exactly three possibilities for the structures that comprise a pseudoknot: single-stranded nucleotides, nested helices, and in-line helices (Figure 2.1). Duplexes containing single nucleotide bulges are counted as a single helix. This model emphasizes structures rather than topologies and appears to be compatible with the vast majority of known pseudoknots. In essence, energetically favorable pseudoknots feature a small number of the single-stranded, nested helix, and in-line helix "beads." To account for the number of constituent single-stranded (*SS*) nucleotides and nested (*NE*) helices (Figure 2.1), we adopted a simple polymer physics-based model [22]. The energetic penalty associated with each of these features is weighted by distances of e = 6.5 Å and f = 15 Å, the mean lengths of a single-stranded nucleotide and a nested helix element, respectively[22] (Figure 1). Finally, we created a penalty for in-line (*IL*) helices (Figure 2.1). The potential to form these structures is weighted by their end-to-end length (n) in the context of A-form helix geometry and the distribution of in-line helices in RNAs of known structure. The model for the entropic cost of pseudoknot formation, ΔG°_{PK}, has two adjustable parameters, *P1* and *P2*:

$$\Delta G^\circ_{PK} = P1\ ln\ (e^2\ SS + f^2\ NE) + P2\ ln\ \Sigma_{IL(n)}(\lambda_n{}^2) \qquad (2)$$

where λ_n is the penalty constant for in-line helices of length n (see Table 2.2). The first term penalizes formation of pseudoknots with long single-stranded regions and many nested helices, whereas the second term enforces an optimal geometry for in-line helices.

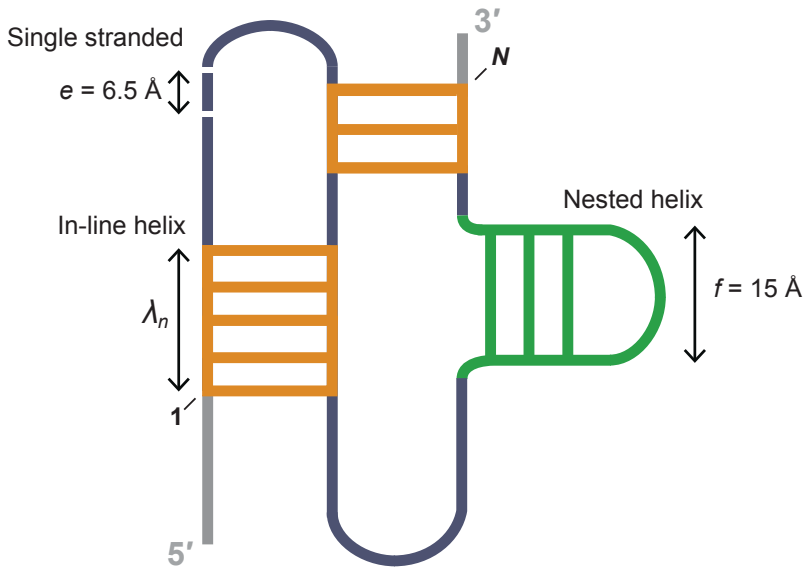

Figure 2.1: Overview of pseudoknot structure model and entropic penalty terms.

Length features are incorporated into ΔG°_{PK} as described in Eqn. 2. Energy penalties for single-stranded nucleotides and nested helices are based on a previously developed model [22]; the penalty for in-line helices was developed in this work.

Helix Length (n)	p_n (Å)	q_n	$\lambda_n = p_n / q_n$
2	0	0.0000	0
3	6.1	0.2546	24
4	11.9	0.4975	24
5	16.9	0.4962	34
6	20.9	0.8795	24
7	23.8	0.6869	35
8	25.5	0.4430	58
9	26.4	0.3217	82
10	26.8	0.4104	65
11	27.4	0.0519	527
12	28.6	0.0117	2447
13	30.9	0.0074	4199
14	34.1	0.0052	6564
15	38.0	0.0030	12540

Table 2.2: Energy penalty per in-line pseudoknotted helix of length *n*.

p_n is the end-to-end distance (in Å) between the C4' of the first and last nucleotide of an (in-line) helix of length n. The value q_n was calculated in two steps. First, for five classes of RNA – group I introns [24, 25], RNase P [26], SRP [27], tmRNA [28] and telomerase[29] – we calculated the fraction of in-line helices of length n over the total number of pseudoknotted structures in each class of RNA. Second, we averaged the fractions of length n across the five RNA classes. λ_n, the penalty constant for an in-line helix of length n, is the quotient of p_n and q_n.

2.2.3 RNA structure interrogation by SHAPE.

Most RNAs were transcribed *in vitro* and contained short hairpin-containing structure cassettes at their 5' and 3' ends [23]. The 16S and 23S ribosomal RNAs were isolated from total *E. coli* RNA[15]. The transcribed RNAs were folded in a standard buffer with physiologically relevant ion concentrations (and saturating ligand concentrations for riboswitches) and treated with 1-methyl-7-nitroisatoic anhydride (1M7) [30]. Sites of 2'-*O*-adduct formation were detected by primer extension using a previously described high-throughput SHAPE approach[31]. SHAPE reactivities were normalized to place them on a scale from zero (unreactive) to ~1.5 (highly reactive). In this work, we illustrate modeling results in the form of circle plots, which provide an unbiased way to visualize correct and incorrect base pairs. The nucleotide sequence is arrayed on the outer circle: unreactive nucleotides (SHAPE reactivities < 0.4) are colored black, moderately reactive nucleotides (0.4 – 0.85) are yellow, and highly reactive nucleotides (> 0.85) are red. Base pairs are shown as arcs, colored by whether they are predicted correctly or not (Figure 2.2). Pseudoknots correspond to helices whose arcs cross in the circle plot. In general, there was a strong correspondence between SHAPE reactivities and the pattern of base pairing in the accepted structures. Nucleotides that participate in canonical base pairs were generally unreactive; whereas nucleotides in loops, bulges, and other connecting regions were reactive (Figure 2.2).

2.2.4 Algorithm and Parameter Determination.

Our ShapeKnots algorithm has four underlying parameters: m and b used in calculation of $\Delta G°_{SHAPE}$ and $P1$ and $P2$ used to calculate $\Delta G°_{PK}$ from Eqns. 1 and 2, respectively. The $\Delta G°_{SHAPE}$ parameters, m and b, penalize or favor base pairs with high

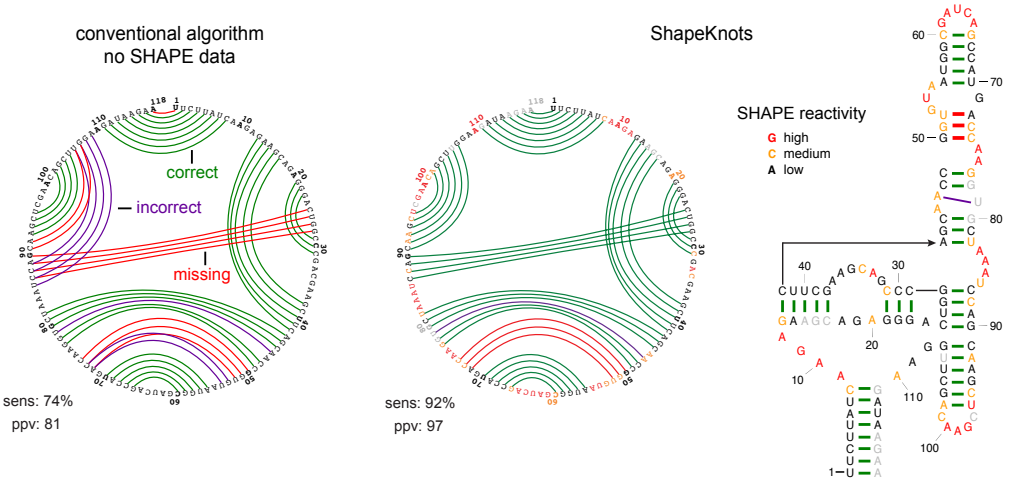

Figure 2.2: Representative ShapeKnots structure prediction for the SAM I riboswitch.

In all panels, base pair predictions are illustrated with colored lines: green, correctly predicted; red, missed base pair relative to the accepted [32] structure; purple, prediction of a pair not in the accepted structure. Left-hand panel shows predictions without SHAPE data. Center and right-hand panels show predictions made when SHAPE data were included, using circle plot and conventional representations, respectively. Sensitivity (sens) and ppv are listed for each structure. SHAPE data are shown as colored nucleotide letters on a black, yellow, red scale for low, medium and high SHAPE reactivities, respectively.

and low SHAPE reactivities, respectively, are universal to all RNAs, and do not directly contribute to the entropic penalty for pseudoknot formation. These parameters can thus be fit independently of the $\Delta G°_{PK}$ terms, *P1* and *P2*. *m* and *b* were optimized using the seven RNAs in our dataset that do not contain pseudoknots. To reduce over-optimization of these parameters, we used a leave-one-out jackknife approach[33] to assess prediction sensitivities, ppv, and the geometric mean of these parameters at each grid point for seven quasi-independent data sets each containing six of the seven RNAs.

Our algorithm for identification of pseudoknots follows the approach implemented in HotKnots[10]. A two-stage refinement first finds stable helices using a dynamic programming algorithm that does not allow pseudoknots. The second stage uses the same dynamic programming algorithm to predict structures for each stable helix found in stage one. In stage two, structures are predicted such that nucleotides in the stable helix are forced to not pair. These pairs are subsequently added back to the structure, and these helices can therefore be pseudoknotted. This allows the prediction of up to one pseudoknot per run. Run times for the final ShapeKnots algorithm were less than 1 min for RNAs of fewer than 150 nts and ~90 min for the longest (530 nt) RNA (Table 2.3).

The pseudoknot-specific parameters, *P1* and *P2*, were fit using a jackknife approach incorporating data from all 16 RNAs in the training set. Parameters were optimized in three stages (see Methods). In this analysis, m = 1.8 and b = -0.6 kcal/mol yielded the most accurate secondary structure predictions (Figure 2.3). These parameters differ slightly from the values (m = 2.6 and b = -0.8 kcal/mol) determined previously using only *E. coli* 23S rRNA[18].

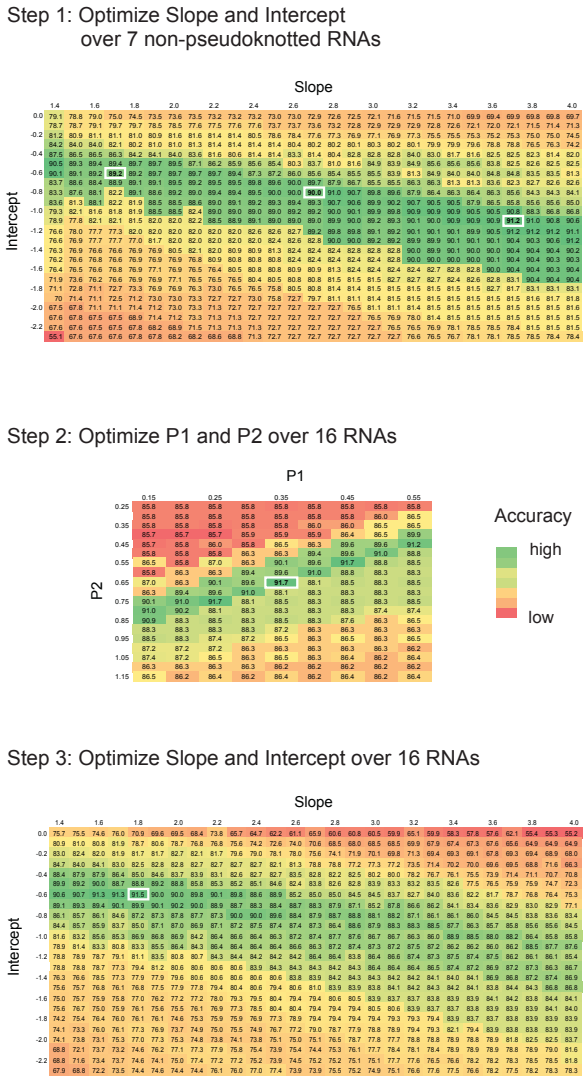

Figure 2.3: Optimization of the $\Delta G°_{SHAPE}$ and $\Delta G°_{PK}$ parameters (in kcal/mol) by jackknifing.

Each of the three panels shows a representative grid in which the M-box RNA was left out. Optimal parameters in each case are emphasized with a white box. Each box in the grid represents the accuracy (calculated as the geometric mean of sens and ppv) for the test set at each slope and intercept for Steps 1 and 3, and each P1 and P2 value for Step 2. For clarity, only a subset of parameter optimizations is shown.

Table 2.3: ShapeKnots run times as a function of RNA length.

RNA	Length	Folding time (sec)	Folding time (min)
Pre-Q1 riboswitch, *B. subtilis*	34	0.05	< 1
Telomerase pseudoknot, human	47	0.31	< 1
$tRNA^{Asp}$, yeast	75	2.48	< 1
TPP riboswitch, *E. coli*	79	3.60	< 1
SARS corona virus pseudoknot	82	2.50	< 1
cyclic-di-GMP riboswitch, *V. cholerae*	97	4.70	< 1
SAM I riboswitch, *T. tengcongensis*	118	7.63	< 1
M-Box riboswitch, *B. subtilis*	154	24.2	< 1
P546 domain, bI3 group I intron	155	14.2	< 1
Lysine riboswitch, *T. maritime*	174	117	1.9
Group I intron, *Azoarcus* sp.	214	212	3.5
Hepatitis C virus IRES domain	336	900	15.0
Group II intron, *O. iheyensis*	412	1840	30.7
Group I Intron, *T. thermophila*	425	2530	42.2
5' domain of 23S rRNA, *E. col*	511	4620	77.0
5' domain of 16S rRNA, *E. coli*	530	5480	91.4

Run times for test set RNAs based upon single processor (non-parallel) calculation using a Linux Server with a 2.93 GHz Intel Xeon (model X5679) processor and 48 GB memory per node.

We recommend use of these new values for RNA structure prediction both with and without pseudoknots. Applying ShapeKnots using these ΔG°_{SHAPE} and ΔG°_{PK} parameters yielded an average sensitivity for secondary structure prediction of 93% for the sixteen RNAs in the test set (Table 2.1).

2.2.5 Extension to additional RNAs.

We used ShapeKnots to model secondary structures for six RNAs that were not used to optimize the final algorithm. Three RNAs – the adenine riboswitch, $tRNA^{Phe}$, and *E. coli* 5S rRNA – were chosen because prior approaches using non-standard data analysis had suggested that they folded poorly with SHAPE data[16]. The other three RNAs – the fluoride riboswitch pseudoknot, 5' domain of the *H. volcanii* 16S rRNA, and the 5' pseudoknot leader of the HIV-1 RNA genome – adopt structures that are predicted poorly by conventional approaches. Overall prediction sensitivities for these six RNAs were ~95% (Table 1), and the pseudoknots in the HIV-1 and fluoride riboswitch RNAs [34-36] were identified correctly

2.3 Discussion

Pseudoknots are relatively rare in large RNAs but are highly overrepresented in important functional regions[2, 3, 6, 7]. Despite their importance, the most commonly used RNA structure prediction algorithms do not permit pseudoknots because allowing pseudoknots both increases algorithmic complexity and the number of possible structures. Current algorithms that allow pseudoknots recover only ~70% of the total accepted base pairs.

A – Overall prediction accuracies

	ShapeKnots				ProbKnot				DotKnot+KL		DotKnot-KL		ipknot		pknotsRG-mfe		HotKnots		Fold			
SHAPE data	+		–		+		–		–		–		–		–		–		+		–	
	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv
With Pseudoknot																						
Pre-Q1 riboswitch, *B. subtilis*	100	100	62.5	100	62.5	83.3	62.5	83.3	62.5	100	62.5	100	62.5	100	62.5	100	62.5	100	62.5	100	62.5	100
Telomerase pseudoknot, human	100	100	100	100	80.0	80.0	86.7	100	100	100	100	100	100	100	100	100	100	100	60.0	75.0	40.0	75.0
SARS corona virus pseudoknot	84.6	88.0	65.4	68.0	65.4	65.4	65.4	63.0	88.5	95.8	88.5	95.8	69.2	72.0	88.5	88.5	69.2	62.1	69.2	75.0	69.2	90.0
SAM I riboswitch, *T. tengcongensis*	92.3	97.3	76.9	81.1	76.9	85.7	76.9	85.7	94.9	97.4	94.9	97.4	89.7	97.2	84.6	97.1	79.5	91.2	76.9	85.7	74.4	80.6
Lysine riboswitch, *T. maritima*	87.3	88.7	84.1	82.8	84.1	86.9	82.5	89.7	71.4	75.0	65.1	74.6	82.5	91.2	77.8	83.1	77.8	83.1	79.4	89.3	77.8	83.1
Group I intron, *Azoarcus* sp.	92.1	95.1	73.0	75.4	76.0	82.5	60.3	66.7	44.4	45.2	41.3	44.1	66.7	75.0	42.9	45.0	46.0	49.2	81.0	85.0	73.0	75.4
Hepatitis C virus IRES domain	92.3	96.0	39.4	36.3	74.6	83.2	52.9	54.5	26.0	27.3	29.8	30.7	56.7	71.1	30.8	29.6	29.8	30.1	79.8	86.5	39.4	38.0
Group II intron, *O. iheyensis*	93.2	97.6	88.6	97.5	79.0	92.1	54.9	64.0	60.2	64.5	51.1	58.6	80.5	89.9	88.0	93.6	89.5	95.2	74.2	84.5	88.6	97.5
Group I Intron, *T. thermophila*	93.9	91.2	83.2	74.3	85.6	83.8	83.3	78.2	66.7	64.5	70.5	69.6	90.2	89.6	79.6	76.3	72.7	72.4	87.8	88.6	83.2	74.3
Average	**92.9**	**94.9**	**74.8**	**79.5**	**76.0**	**82.5**	**69.5**	**76.1**	**68.3**	**74.4**	**67.1**	**74.5**	**77.6**	**87.3**	**72.7**	**79.2**	**69.7**	**75.9**	**74.5**	**85.5**	**67.6**	**79.3**
Without Pseudoknot																						
tRNA(asp), yeast	95.2	95.2	95.2	95.2	95.2	80.0	76.2	61.5	71.4	62.5	76.2	61.5	81.0	100	66.7	56.0	61.9	56.5	95.2	95.2	95.2	95.2
TPP riboswitch, *E. coli*	95.5	87.5	77.3	85.0	100	81.5	77.3	85.0	72.7	80.0	72.7	80.0	68.2	88.2	77.3	85.0	22.7	20.8	95.5	87.5	77.3	85.0
cyclic-di-GMP riboswitch, *V. cholerae*	89.3	86.2	75.0	77.8	89.3	92.6	85.7	92.3	42.9	48.0	42.9	48.0	82.1	79.3	96.4	93.1	92.9	89.7	89.3	86.2	75.0	77.8
M-Box riboswitch, *B. subtilis*	87.5	91.3	87.5	91.3	87.5	85.7	87.5	82.4	87.5	91.3	87.5	91.3	87.5	91.3	87.5	91.3	87.5	91.3	87.5	91.3	87.5	91.3
P546 domain, bI3 group I intron	94.6	96.4	42.9	44.4	98.2	94.8	76.8	89.6	44.6	45.5	44.6	45.5	55.4	88.6	42.9	44.4	42.9	44.4	94.6	96.4	42.9	44.4
5' domain of 23S rRNA, *E. coli* †	97.2	76.8	97.2	73.8	96.3	80.8	93.6	76.3	78.9	59.3	70.6	53.8	93.6	89.5	90.8	72.3	90.8	70.7	97.2	76.8	97.2	73.8
5' domain of 16S rRNA, *E. coli* †	93.0	83.6	63.6	59.1	93.0	78.7	64.3	59.0	61.5	55.0	71.3	68.0	80.4	90.6	76.9	71.4	76.9	70.1	93.0	83.6	63.6	59.1
Average	**93.2**	**88.2**	**77.0**	**75.2**	**94.2**	**84.9**	**80.2**	**78.0**	**65.7**	**63.1**	**66.6**	**64.0**	**78.3**	**89.6**	**76.9**	**73.4**	**67.9**	**63.4**	**93.2**	**88.2**	**77.0**	**75.2**

B – Prediction accuracies for pseudoknotted base pairs only

	ShapeKnots				ProbKnot				DotKnot+KL		DotKnot-KL		ipknot		pknotsRG-mfe		HotKnots	
SHAPE data	+		–		+		–		–		–		–		–		–	
	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv	sens	ppv
With Pseudoknot																		
Pre-Q1 riboswitch, *B. subtilis*	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Telomerase pseudoknot, human	100	100	100	100	0	0	86.7	100	100	100	100	100	100	100	100	100	100	100
SARS corona virus pseudoknot	77.8	87.5	0	0	0	0	0	0	83.3	100	83.3	100	0	0	83.3	100	0	0
SAM I riboswitch, *T. tengcongensis*	100	100	0	0	0	0	0	0	100	100	100	100	0	0	0	0	0	0
Lysine riboswitch, *T. maritima*	81.1	83.3	78.4	78.4	0	0	0	0	59.5	62.9	0	0	0	0	0	0	0	0
Group I intron, *Azoarcus* sp.	91.7	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hepatitis C virus IRES domain	94.1	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Group II intron, *O. iheyensis*	97.6	95.2	0	0	0	0	0	0	48.8	41.7	0	0	75.6	70.5	0	0	0	0
Group I Intron, *T. thermophila*	68.8	68.8	0	0	0	0	0	0	0	0	0	0	62.5	83.3	0	0	0	0
Average	**90.1**	**92.8**	**19.8**	**19.8**	**0.0**	**0.0**	**9.6**	**11.1**	**43.5**	**44.9**	**31.5**	**33.3**	**26.5**	**28.2**	**20.4**	**22.2**	**11.1**	**11.1**
Without Pseudoknot																		
tRNA(asp), yeast	100	100	100	100	100	100	0	0	0	0	0	0	100	100	0	0	100	100
TPP riboswitch, *E. coli*	100	100	100	100	100	100	100	100	0	0	0	0	0	0	100	100	0	0
cyclic-di-GMP riboswitch, *V. cholerae*	100	100	100	100	100	100	100	100	0	0	0	0	0	0	100	100	100	100
M-Box riboswitch, *B. subtilis*	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
P546 domain, bI3 group I intron	100	100	100	100	100	100	100	100	100	100	100	100	0	0	100	100	100	100
5' domain of 23S rRNA, *E. coli* †	100	100	100	100	0	0	100	100	0	0	0	0	100	100	100	100	100	100
5' domain of 16S rRNA, *E. coli* †	100	100	100	100	0	0	0	0	0	0	0	0	100	100	100	100	100	100
Average	**100**	**100**	**100**	**100**	**71.4**	**71.4**	**71.4**	**71.4**	**28.6**	**28.6**	**28.6**	**28.6**	**57.1**	**57.1**	**85.7**	**85.7**	**85.7**	**85.7**

C – Reference structure statistics

	Length	PK Min	Total PK
With Pseudoknot			
Pre-Q1 riboswitch, *B. subtilis*	34	3	8
Telomerase pseudoknot, human	47	6	15
SARS corona virus pseudoknot	82	7	18
SAM I riboswitch, *T. tengcongensis*	118	4	12
Lysine riboswitch, *T. maritima*	174	6	39
Group I intron, *Azoarcus* sp.	214	6	12
Hepatitis C virus IRES domain	336	6	17
Group II intron, *O. iheyensis*	412	7	44
Group I Intron, *T. thermophila*	425	6	14

	Length
Without Pseudoknot	
tRNA(asp), yeast	75
TPP riboswitch, *E. coli*	79
cyclic-di-GMP riboswitch, *V. cholerae*	97
M-Box riboswitch, *B. subtilis*	154
P546 domain, bI3 group I intron	155
5' domain of 23S rRNA, *E. coli* †	511
5' domain of 16S rRNA, *E. coli* †	530

Table 2.4: Prediction accuracies for seven RNA folding algorithms (following page).

A) Overall prediction accuracies. Accuracies are shown as percent sensitivity (sens) and positive predictive value (ppv), and allow pairing to be shifted by one position on one side of a pair [37]. ShapeKnots, ProbKnot [13], and Fold (the standard RNAstructure algorithm) [38] were run both with and without SHAPE data. Other included algorithms are DotKnot+KL and DotKnot-KL (KL indicates kissing loops) [39, 40], ipknot [41], pknotsRG-mfe [42], and HotKnots [12]. Note that Fold does not allow pseudoknots. All algorithms were run using their default parameters. ShapeKnots, Fold and ProbKnot used *m* and *b* parameters (Eqn. 1) of 1.8 and -0.6 kcal/mol, respectively.
B) Prediction accuracies for pseudoknotted base pairs only. Accuracies are evaluated using sensitivity and ppv, allowing for mis-pairing by one position on one side of a pair [37]. If both accepted and predicted structures contain no pseudoknot, sens and ppv are defined as 100%. If only the predicted structure contains a pseudoknot, sens and ppv are set to 0. A pseudoknotted pair is scored as correctly predicted only if there is at least one other correctly predicted pair with which it forms a pseudoknot. Fold is excluded because it does not allow prediction of pseudoknots.
C) Reference structure statistics. PK Min is the minimum number of pairs required to break a pseudoknot [43]. Total PK is the total number of pseudoknotted pairs in the accepted structure.

The prediction sensitivity for base pairs that specifically form pseudoknots varies by algorithm and benchmark RNAs but averages only 5-20%, with many false-positive predictions[13] (Table 2.4). Thus, the current generation of pseudoknot prediction algorithms is poorly suited for designing testable biological hypotheses.

ShapeKnots combines an iterative pseudoknot discovery algorithm with experimental SHAPE information and a simple energy model for the entropic cost of pseudoknot formation. The pseudoknot penalty in ShapeKnots has only two adjustable parameters (Figure 2.1 and Eqn. 2) that limit formation of pseudoknots with long single-stranded regions and many nested helices and that enforce an optimal geometry for in-line helices. ShapeKnots also allows incorporation of an experimental correction to standard free energy terms. Including SHAPE data both limits the number of possible structures and provides information that accounts for hidden features that stabilize RNA folding, including the significant effects of metal ion and ligand binding.

Our set of training structures was comprised of sixteen RNAs of known structure that ranged in length from 34 to 530 nucleotides; pseudoknots occur in nine of the sixteen RNAs. Prediction accuracies were consistently high (Table 2.1). ShapeKnots significantly outperformed currently available pseudoknot prediction algorithms and is the only algorithm to achieve >90% overall and pseudoknot-specific sensitivities with this test set (Table 2.4; see Methods for additional discussion). Both the specific pseudoknot energy penalty and use of SHAPE data contribute to the accuracy of the ShapeKnots approach. It is likely that inclusion of SHAPE data will generally improve accuracies for pseudoknot prediction algorithms.

We summarize our modeling results by emphasizing four classes of RNA: (*i*) short pseudoknotted RNAs with structures that ShapeKnots predicts very accurately, (*ii*) large, challenging RNAs that ShapeKnots predicts with good accuracy, (*iii*) RNAs with high likelihood of being mischaracterized with false-positive or missed pseudoknots that ShapeKnots predicts accurately, and (*iv*) RNAs that interact with other molecules such as ligands, proteins, and metal ions that pose unique challenges. For most RNAs analyzed here, differences between models generated by ShapeKnots and currently accepted structures were minor and typically involved short-range interactions or base pairs at the ends of helices. In some cases, differences likely reflect thermodynamically accessible states at equilibrium in solution.

2.3.2 Short pseudoknotted RNAs.

The first class includes small RNAs that contain H-type pseudoknots: the pre-Q1 riboswitch, human telomerase, the fluoride riboswitch, and a SARS corona virus domain. Because the most commonly used dynamic programming algorithms cannot predict base pairs in an H-type pseudoknot, prediction sensitivities using a conventional algorithm[38] were quite poor; in contrast, ShapeKnots yielded perfect or near-perfect predictions in each case (Figure 2.4). The only ShapeKnots-predicted base pairs that do not occur in the accepted structures involve sets of two or fewer base pairs located at the ends of individual helices in the fluoride riboswitch and SARS domain. These results suggest that ShapeKnots prediction of H-type pseudoknots in short RNAs is robust.

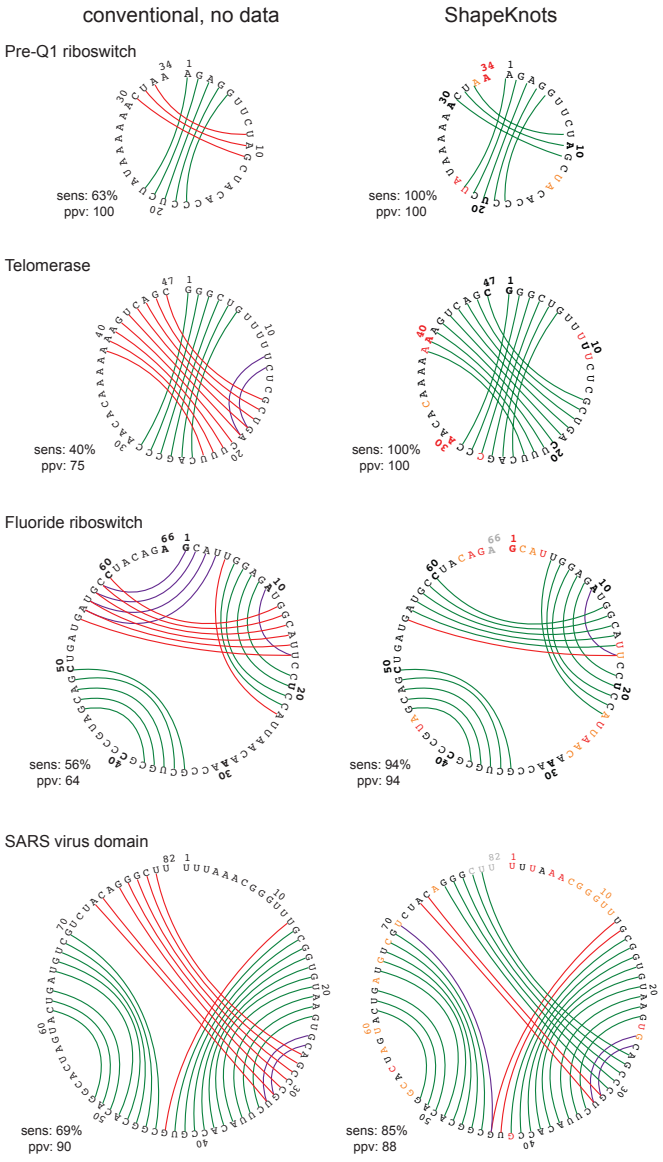

Figure 2.4: Summary of predictions for four H-type pseudoknots.

Base pair predictions are illustrated as outlined in Figure 2.2; sensitivity (sens) and ppv are listed for each structure. Left- and right-hand columns show predictions for a conventional mfold-class algorithm versus ShapeKnots (with experimental SHAPE restraints).

2.3.1 Large, complex RNAs.

The second class includes large RNAs that do not require ligands or protein co-factors for correct folding. Large RNAs pose a challenge to modeling algorithms due to the vast number of possible structures and due to the large number of structures with similar folding free energies changes. For example, in the absence of experimental structure probing data, two representative RNAs, the *Azoarcus* group I intron and the hepatitis C virus IRES domain are predicted with sensitivities of 73 and 39%, respectively. Mis-predictions occur primarily in two hairpin motifs in the *Azoarcus* RNA but span essentially the entire HCV IRES RNA (Figure 2.5). Inclusion of SHAPE data yielded near-perfect predictions in each case, including correct identification of the pseudoknot in each RNA (Figure 2.5).

2.3.2 RNAs with difficult to predict pseudoknots.

Within a given RNA sequence, several physically reasonable pseudoknots are often possible; for example, Figure 2.6 shows the SARS virus domain with two potential pseudoknotted helices are identified in purple and red. Conversely, as exemplified by the SAM I riboswitch, pseudoknots can be missed because the energy function does not distinguish small differences in stabilities of a pseudoknot-forming versus a more local helix (Figure 2.6). The experimental SHAPE-based correction correctly re-ranked the stabilities for the two possible helices located close to one another in topological space in the SARS and riboswitch RNAs, ultimately avoiding both false-positive and false-negative pseudoknot predictions (Figure 2.6).

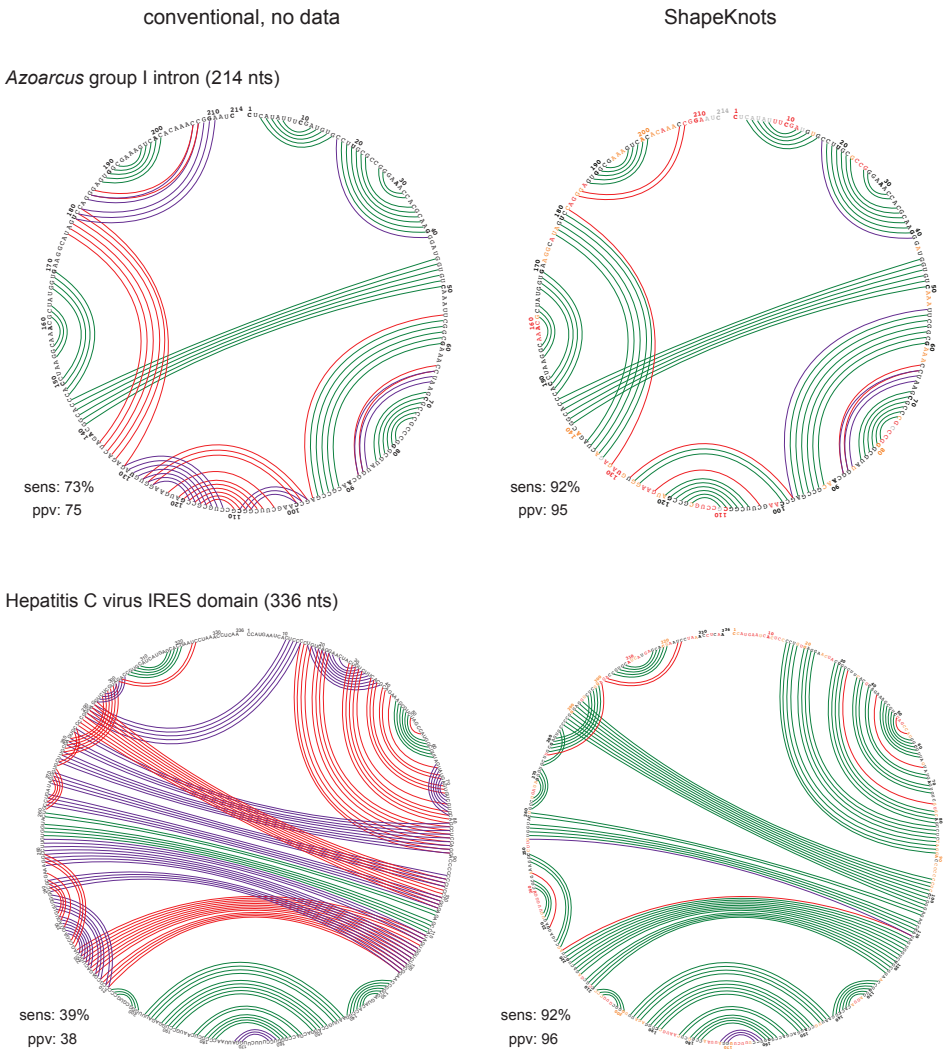

Figure 2.5: Prediction summaries for two large, pseudoknot-containing RNAs. Structural annotations are as described in Figure 2.2.

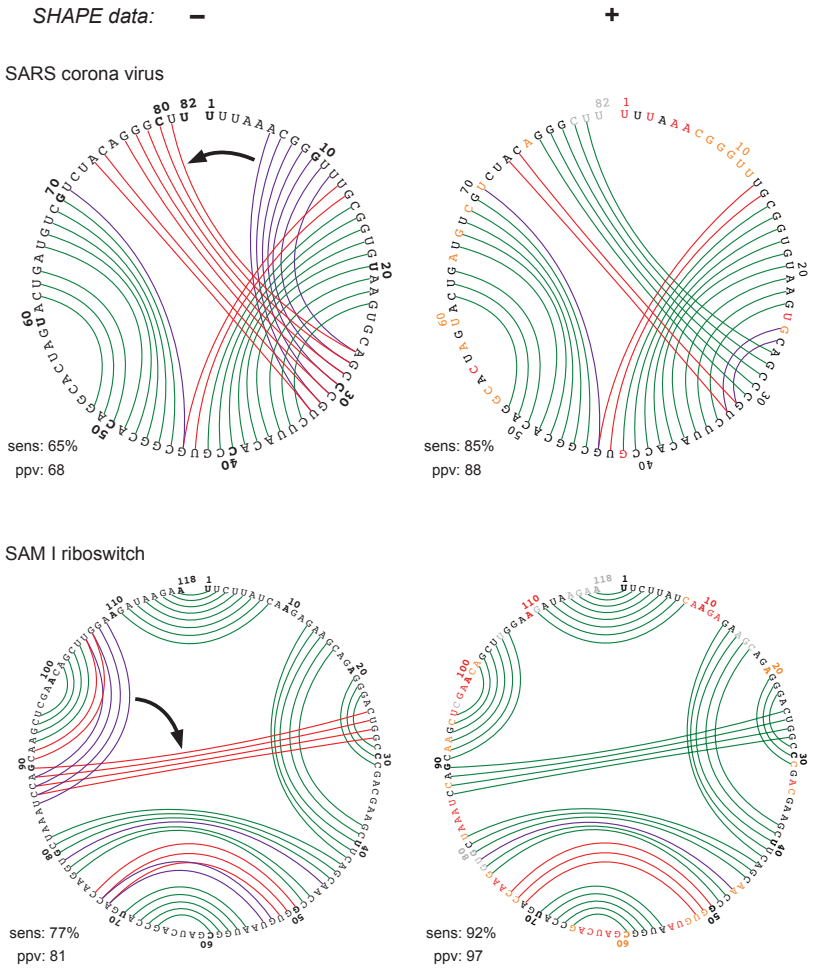

Figure 2.6: Representative examples in which ShapeKnots avoids false-positive (top) or false-negative (bottom) pseudoknot predictions.

Left- and right-hand panels show the results of ShapeKnots predictions without and with SHAPE data, respectively. The bold arrow in the left-hand panels emphasizes the replacement of an accepted (red) with an incorrect (purple) helix in the absence of data. Other structural annotations are as described in Figure 2.2.

2.3.1 RNAs that do not adopt their accepted structures.

During our analysis of experimentally directed structure modeling, we examined two RNAs for which the *in vitro* SHAPE data were clearly incompatible with the accepted structure. These RNAs were the signal recognition particle RNA and RNase P. In each case, the SHAPE-directed model using ShapeKnots provided a significant improvement relative to the pseudoknot-free lowest free energy predicted structure (Table 2.1). Nonetheless, a large part of each structure was mis-predicted relative to the accepted structure. In each case, nucleotides in some helices in the accepted structural model were reactive by SHAPE, suggesting that these helices do not form under the solution conditions used here for *in vitro* structure probing (Figure 2.7). There are several possible explanations for the observed discrepancies. First, the conditions under which these RNAs were crystallized are different from the roughly physiological ion conditions used in SHAPE probing experiments. The differences in conditions could cause the crystallographic structure to be different from that in solution or there may be structural inhomogeneity in solution. Second, both the RNase P and signal recognition particle RNAs function as RNA-protein complexes. These proteins were not present during *in vitro* SHAPE experiments.

2.3.2 Perspective.

It is difficult to account for many factors that impact RNA secondary structure – including effects of metal ions, ligands, and protein binding – using a system based on thermodynamic or structural parameters. For example, the M-Box and fluoride riboswitch RNAs undergo large conformational changes upon binding by Mg^{2+} or F^- ions, respectively[36, 44], and binding of ligands to the pre-Q1, TPP, cyclic-di-GMP, SAM,

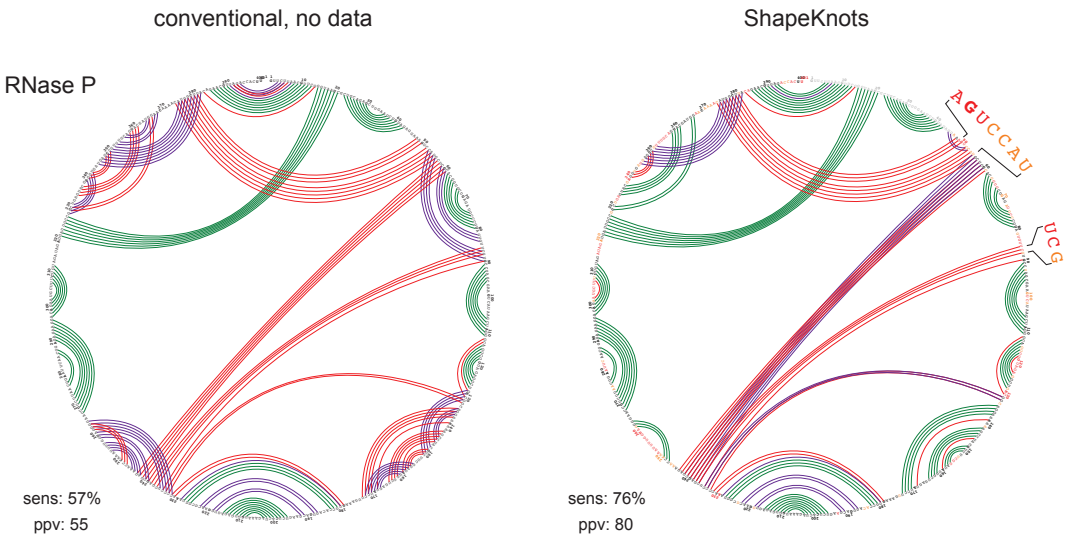

Figure 2.7: Prediction summary for RNase P RNA.

This RNA, along with the signal recognition particle RNA, does not appear to fold into its conventionally accepted structure based on in-solution SHAPE data. Regions of strongest disagreement are highlighted as magnified letters.

and adenine riboswitches provides a large fraction of the total interactions that ultimately stabilize the accepted structure[7]. In addition, many of the RNA in our dataset contain base triple interactions, which are common in pseudoknots [45]. With the inclusion of SHAPE data, the ShapeKnots approach does a good job of modeling these interactions (Table 2.1). Other challenges to structure prediction are that some base pairs may only be stable in the presence of bound proteins and some RNAs, especially as exemplified by riboswitches [7], sample multiple conformations. Finally, *in vitro* refolding and probing protocols may not fully recapitulate the functional or *in vivo* structure. Our analyses of the signal recognition particle RNA and RNase P illustrate these challenges: Neither of these RNAs appears to fold stably to the accepted structure under solution conditions used in this work (Figure 2.7). These two RNAs are widely used to benchmark folding algorithms, even though they may only fold robustly to their accepted structures in the context of their native RNA-protein complexes. In this case, for the specific solution environment used here, the SHAPE-directed structures appear to be roughly "correct" but just not the expected ones.

In the context of the diverse RNAs examined in this work, the ShapeKnots algorithm recovered 93% of accepted base pairs in well-folded RNAs (Table 2.1), significantly out-performing current algorithms. Nonetheless, evaluation of ShapeKnots is currently restricted by challenges that impact the entire RNA structure modeling field[16]. There exist relatively few RNAs with non-trivial structures that are known at a high level of confidence. The ShapeKnots energy penalty and search algorithm may require adjustment as new pseudoknot topologies are discovered. RNAs that have been solved by crystallography have features that make them simultaneously both more and less

difficult to predict than more typical structures: They tend to contain a relatively high level of non-canonical and complex tertiary interactions (difficult to predict features), and they fold into structures with many stable base-paired regions (more readily predicted using thermodynamics-based algorithms). In addition, the structures inferred from high-resolution data may not represent the solution conformation of the purified RNAs. For RNAs in which the accepted structure is based on phylogenetic and in-solution evidence – as exemplified by the SARS virus and HCV IRES domains – ShapeKnots predictions may identify correct features missed in current accepted structures. The approaches outlined in this work – use of simple models for base pairing and pseudoknot formation, including experimental corrections to thermodynamic parameters, and nuanced interpretation of differences between current accepted and modeled structures – represents a critical departure point for future accurate RNA secondary structure modeling.

2.4 Experimental

2.4.1 ShapeKnots algorithm.

ShapeKnots predicts and ranks a set of low free energy, potentially pseudoknot-containing structures. Two steps use dynamic programming algorithm calculations, using pseudoknot-free predictions, to first identify possible pseudoknotted helices and then fold the remaining sequence, possibly creating a pseudoknot. This approach is closely related to the HotKnots algorithm[10]. The following steps are performed:

The dynamic programming algorithm is used to generate the pseudoknot-free minimum free energy structure, S_{mfe}. S_{mfe} along with up to 99 low energy suboptimal structures are included in the final list of candidate structures, S. The folding free energy

change of a suboptimal structure must be within 20% of the $\Delta G°$ of S_{mfe}, with no restrictions on how different suboptimal structures are from each other (a window size of zero). The algorithm is also used to generate an energy dot plot, indicating, for all nucleotides *i* and *j*, the lowest folding free energy possible for a structure containing the *i*-*j* base pair. The $\Delta G°$ values are calculated using the current Turner nearest neighbor parameters [17, 21] but with the multi-branch loop per helix parameter value of -0.6 kcal/mol[46, 47]. The SHAPE pseudo-free energy terms are incorporated into the dynamic programming algorithm for each paired nucleotide per base pair stack of an adjacent paired nucleotide[15].

A candidate pseudoknot helix list, *H*, along with the corresponding helix energies is generated from the energy dot plot. Helix H_i is accepted into *H* if it spans at least three base-pairs. For sequences longer than 100 nucleotides, H_i also has to occur in a structure with a $\Delta G°$ within 25% of the free energy of S_{mfe}. The $\Delta G°$ of H_i is calculated as the sum of the nearest neighbor stacks and terminal AU/GU pair penalties[21].

The set of helices, *H,* is filtered in two steps. First, helices are compared to those in the minimum free energy structure. Helices are discarded if more than 50% of their nucleotides are base paired in S_{mfe}. Second, to increase computational efficiency, *H* is trimmed to include a maximum of 100 of the most thermodynamically stable helices.

For each H_i, a new set of structures, composed of the lowest free energy structure and up to 100 suboptimal structures, is generated by the dynamic programming algorithm, where all nucleotides in H_i are prohibited from pairing (forced single-stranded) [48]. Suboptimal structures are chosen in the same way as in step 1. After these structures have been generated, base pairs from H_i are restored to the structures. The $\Delta G°$ of each structure is

incremented by the free energy of the corresponding helix H_i. All unique structures are added to *S*.

For each structure in *S* that contains a pseudoknot, the entropic cost of pseudoknot formation is penalized by $\Delta G°_{PK}$ (Eqn. 2). All pseudoknots require at least two helices, arranged such that at least part of the loop defined by one-helix base pairs to form a second helix. We define the nucleotides involved in a given pseudoknot as starting with the 5'-most nucleotide of the first helix and ending with the last nucleotide of any helix participating in the pseudoknot (nucleotides *1* and *N* in Figure 2.1). There are three possible classes of intervening structures that can be formed in a pseudoknotted structure. *SS* is the number of single-stranded nucleotides inside the pseudoknot, *NE* is the number of nested helices inside the pseudoknot, and *IL*(*n*) is the number of in-line helixes of length *n* base pairs. Before the intervening structures are calculated, the pseudoknot is preprocessed by filling single and tandem mismatches with base pairs and removing isolated pairs. Helices containing a single bulged nucleotide are counted as a single helix. The penalty for single-stranded and nested helices results from a simplified version of a polymer-theory model[22], and the in-line penalty is unique to this work. The terms *e*, *f*, and λ_n (Eqn. 2) are penalty constants per single-stranded nucleotide, nested helix, and in-line helix of length *n*, respectively. Terms *e* and *f* scale the entropic penalty by the distance between the 4' carbons of neighboring unpaired nucleotides and across a single base pair, respectively [22]. We penalize each in-line helix (which, by definition, includes the two that define the pseudoknot plus any other helices with this connectivity; Figure 2.1) by λ_n, an empirical parameter related to the likelihood that an in-line helix, of length *n*, comprises a pseudoknot. λ_n is calculated as the C4'-to-C4' helix length, p_n, divided by a frequency

factor, q_n (Table 2.2). The frequency factors were tabulated in two steps. First, by counting the number of in-line helices of length n from five pseudoknot-containing structure classes – group I introns [49, 50], RNase P RNA [51], SRP RNA [52], tmRNA [53], and telomerase RNA [29] – and dividing by the total number of structures in each class. Second, by averaging the frequencies across the five RNA classes. In-line helix frequencies *P1* and *P2* are constant energy parameters that include Boltzmann constants and temperature terms and must be determined empirically. $\Delta G°_{PK}$ is added to the total $\Delta G°$ of each pseudoknot-containing structure.

S is sorted based on total energy of structures. By default, the 20 lowest free energy structures are reported; the outputted structures are constrained using a Window parameter to ensure that they are sufficiently different from each other[54]. To be included, a structure must contain at least Window base pairs that are more than Window nucleotides distant from pairs in lower free energy structures. The default Window parameter is selected based on the sequence length, where a larger value is used on longer sequences. Finally, a maximum percent energy difference parameter is used to ensure that no structures are included that are higher in folding free energy change than the specified percent difference from the lowest free energy structure; the default value is 10%.

Coaxial stacking of helices stabilizes pseudoknot formation and is included indirectly in the energy function. First, the choice of helices for assembling pseudoknots from the initial dot plot is guided by inclusion of coaxial stacking in the dynamic programming algorithm. Second, separations between the helices enter the pseudoknot calculation as an increase in the number of single stranded nucleotides (*SS*, Eqn. 2), and

thus penalize the absence of coaxial stacking. The pseudo-free energy change approach developed here is broadly applicable and terms for incorporating additional structural information could readily be added.

2.4.2 Parameterization of $\Delta G°_{SHAPE}$ and $\Delta G°_{PK}$.

Two pseudo-free energy change terms are used to direct folding. The first, $\Delta G°_{SHAPE}$, functions to bias predictions toward helices supported by SHAPE data as described previously [15]. The second, $\Delta G°_{PK}$, is the entropic cost of forming a pseudoknot. Four parameters (*m*, *b*, *P1* and *P2*; Eqns. 1 and 2) are involved. The values for these parameters were optimized using a set of RNAs selected for their complex, and generally difficult to predict, structures. RNAs and literature references to their accepted secondary structures are: Pre-Q1 riboswitch [55, 56], human telomerase RNA [57], $tRNA^{Asp}$ [58], TPP riboswitch [59], and SARS corona virus pseudoknot [60], di-cyclic-GMP riboswitch [61], M-Box riboswitch [44], bI3 group I intron P546 domain [62], SAM I riboswitch [63], *Azoarcus* group I intron [64], lysine riboswitch [65], HCV IRES domain [66], *O. iheyensis* group II intron [67], *Tetrahymena* group I intron [68], and 16S and 23S *E. coli* rRNAs [69]. Parameters were fit using a three-step procedure (Figure 2.3). (*i*) *m* and *b* (Eqn. 1) were determined based on data from seven non-pseudoknotted RNAs using the original RNAstructure algorithm for predicting lowest free energy structures [15] that does not allow for pseudoknots. (*ii*) *P1* and *P2* were determined (Eqn. 2) using data from the complete set of 16 non-pseudoknotted and pseudoknot-containing RNAs using the *m* and *b* values determined in step 1 using the ShapeKnots algorithm. (*iii*) *m* and *b* were re-evaluated based on data from all 16 RNAs and the *P1* and *P2* terms identified in step 2 using ShapeKnots. The steps are described in detail below.

In step 1, *m* and *b* (Eqn. 1) were fit to seven non-pseudoknotted RNAs using the original RNAstructure free energy minimization algorithm that does not allow for pseudoknots. The geometric means of the sensitivity and PPV relative to accepted structures for each RNA were calculated over a grid of *m* and *b* values (Figure 2.3). Values for *m* were varied from 0 to 4.0 and for *b* from -2.5 to 0 kcal/mol in increments of 0.1 kcal/mol. Typically, a range of *m* and *b* parameters gave optimal structure predictions for each RNA. We used a jackknifing procedure [70] to identify the best parameters for all RNAs and to avoid over-fitting; in addition, the of RNAs in our dataset are highly diverse, which also reduces over-fitting. In this procedure, one RNA grid was removed from the set and the remaining six grids were averaged together. This process was repeated such that each RNA was left out once. The *m* and *b* parameters resulting in the top 1% highest geometric averages for each averaged grid were recorded. Three sets of *m* and *b* parameters were consistently optimal for each of the seven jackknifed grids: 3.7 and -1.1, 2.7 and -0.8, and 1.7 and -0.6 (in kcal/mol), respectively (Figure 2.3). All three sets of *m* and *b* values were evaluated in the next step.

In the second step, *P1* and *P2* (Eqn. 2) were fit using data from the complete set of 16 non-pseudoknotted and pseudoknot-containing RNAs using the *m* and *b* values determined in Step 1. *P1* and *P2* were varied from 0 to 1.5 kcal/mol in increments of 0.05. Jackknifing was performed as described in Step 1. Seven sets of parameters overlapped at points of highest accuracy for each of the 16 grids. The average of these sets was 0.35 and 0.65 (in kcal/mol) for *P1* and *P2*, respectively (Figure 2.3); these values were used in the following step.

In step 3, *m* and *b* for $\Delta G°_{SHAPE}$ were re-fit using all 16 RNAs and the *P1* and *P2* terms identified in Step 2. Grid searches were performed on all 16 RNAs, varying *m* and *b* in an approach analogous to that outlined in Step 1. The jackknife procedure yielded values of 1.8 and -0.6 kcal/mol for *m* and *b*, respectively (Figure 2.3).

2.4.3 SHAPE structure probing.

RNAs were transcribed from DNA templates (Exiqon or IDT) and purified by denaturing electrophoresis[31], with the exception of the ribosomal RNAs which were obtained from total *E. coli* or *H. volcanii* RNA. The ribosomal RNAs were obtained from *E. coli* or *H. volcanii* cells and were purified under non-denaturing conditions and fully deproteinized by treatment with proteinase K and extraction against phenol/chloroform [15]. The 5' domains of the *E. coli* 16S and 23S rRNA were defined as positions 27-556 and 15-525, respectively; and the *H. volcanii* the 16S rRNA 5' domain was defined as positions 1-473. The pre-Q1, fluoride, adenine, TPP, SAM I, M-Box and lysine riboswitches, Azoarcus group I intron, hepatitis C virus IRES domain, and 5S rRNA were refolded in 100 mM HEPES (pH 8.0), 100 mM NaCl, and 10 mM $MgCl_2$. The telomerase pseudoknot, $tRNA^{Phe}$, SARS corona virus pseudoknot, cyclic-di-GMP riboswitch, HIV-1 5' pseudoknot domain, *T. thermophila* group I intron, *O. iheyensis* group II intron, signal recognition particle RNA, and RNase P RNA were refolded in 50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), and 3 mM $MgCl_2$. Data for the bI3 P546 domain were reported previously [62] [and was refolded in 40 mM MOPS (pH 8.0) 80 mM potassium acetate, and 20 mM $MgCl_2$]. For all riboswitch SHAPE experiments, reactions were supplemented with a concentration of 5 μM ligand, except the pre-Q1 riboswitch (4 μM ligand). After folding at 37 °C for 30 min, RNAs were

treated with 1M7 (in anhydrous DMSO)[30] to a final concentration of 3 mM and allowed to react at 37 °C for 3 min. Concurrently, a no-reagent DMSO reaction was performed omitting 1M7. Frequencies of 2'-hydroxyl modification were identified by primer extension, resolved using capillary electrophoresis, and quantified using custom software [71,72].

2.4.4 Parameterization of SHAPE data.

After determining the inter-quartile range of the data, nucleotides whose reactivities were greater than 1.5 times interquartile range were taken to be outliers[15]; the maximum number of outliers was capped at 10% for RNAs >100 nts and 5% for RNAs <100 nts. SHAPE reactivities were then divided by the mean of the 10% most reactive non-outlier data, which ultimately placed reactivities on a scale spanning zero (no reactivity) to ~1.5.

We now use and recommend a three-color scale for illustrating SHAPE data in which reactivities less than 0.4 are black, between 0.4 and 0.85 are yellow, and greater than 0.85 are red. The 0.4 point represents the value at which the $\Delta G°_{SHAPE}$ term (Eqn. 1) for base pairing transitions from favorable (negative) to unfavorable (positive) and 0.85 represents a net thermodynamic penalty of 0.5 kcal/mol or 1.0 kcal/mol per internal dinucleotide stack.

2.4.5 Comparison with other algorithms.

We evaluated the importance of SHAPE data and the new penalty for pseudoknot formation (Eqn. 2) by performing additional benchmarks with the programs ProbKnot [13], DotKnot+KL and DotKnot-KL (where KL indicates whether kissing loops are included)

[73, 74], ipknot [41], pknotsRG-mfe [75], and HotKnots [10, 12] (Table 2.4). These programs are freely available and can be run locally. ProbKnot is capable of predicting structures restrained by SHAPE data, and it was therefore benchmarked with and without SHAPE data.

The benchmarks demonstrate the importance of both the pseudoknot free energy change function (Eqn. 2) and the use of SHAPE data for accurate structure prediction, including pseudoknots (Table 2.4). The overall accuracy, when SHAPE data are used, is highest for ShapeKnots, which is the only program that achieves greater than 90% average sensitivity and ppv with the RNAs evaluated in this work. Without SHAPE data, Ipknot performs better than ShapeKnots, and both perform better than ProbKnot, DotKnot+KL, DotKnot-KL, pknotsRG-mfe, and HotKnots.

With respect to predicting the specific base pairs involved in pseudoknot formation in our dataset, ShapeKnots with SHAPE data is the only program that obtains >90% sensitivity and ppv. DotKnot+KL performs best in the absence of SHAPE data at predicting known pseudoknots, and ShapeKnots results in the fewest false positive pseudoknots in the absence of SHAPE data (Table 2.4). Interestingly, the overall accuracy of ProbKnot improved with SHAPE data, but the performance at predicting pseudoknots decreased when SHAPE data were included. ProbKnot relies on a partition function calculation over pseudoknot-free structures to identify the two helices that minimally define a pseudoknot. SHAPE data cause the pseudoknot-free partition function to (too strongly) favor one of the two helices that define the pseudoknot.

2.4.6 Data and software availability.

ShapeKnots is freely available as part of the RNAstructure software package at http://rna.urmc.rochester.edu. All SHAPE datasets are available at http://www.chem.unc.edu/rna and at the SNRNASM community structure probing database [76].

2.5 References

1. Sharp, P.A. The centrality of RNA. *Cell* **136**, 577-580 (2009).

2. Staple, D.W. & Butcher, S.E. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* **3**, e213 (2005).

3. Brierley, I., Pennell, S. & Gilbert, R.J. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.* **5**, 598-610 (2007).

4. Pleij, C.W. Pseudoknots: a new motif in the RNA game. *Trends Biochem. Sci.* **15**, 143-147 (1990).

5. Powers, T. & Noller, H.F. A functional pseudoknot in 16S ribosomal RNA. *EMBO J.* **10**, 2203-2214 (1991).

6. Reiter, N.J., Chan, C.W. & Mondragon, A. Emerging structural themes in large RNA molecules. *Curr. Opin. Struct. Biol.* **21**, 319-326 (2011).

7. Roth, A. & Breaker, R.R. The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem.* **78**, 305-334 (2009).

8. Liu, B., Mathews, D.H. & Turner, D.H. RNA pseudoknots: folding and finding. *F100 Biol. Rep.* **2**, 8 (2010).

9. Lyngsø, R.B. & Pederson, C.N. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* **7**, 409-427 (2000).

10. Ren, J., Rastegari, B., Condon, A. & Hoos, H.H. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *Rna* **11**, 1494-1504 (2005).

11. Dirks, R.M. & Pierce, N.A. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem.* **25**, 1295-1304 (2004).

12. Andronescu, M.S., Pop, C. & Condon, A.E. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *Rna* **16**, 26-42 (2010).

13. Bellaousov, S. & Mathews, D.H. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *Rna* **16**, 1870-1880 (2010).

14. Mathews, D.H. et al. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *P Natl Acad Sci USA* **101**, 7287-7292 (2004).

15. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *P Natl Acad Sci USA* **106**, 97-102 (2009).

16. Leonard, C.W. et al. Principles for understanding the accuracy of SHAPE-directed RNA structure modeling. *Biochemistry* **52**, dx.doi.org/10.1021/bi300755u (2013).

17. Turner, D.H. & Mathews, D.H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **38**, D280-282 (2010).

18. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).

19. Cannon, C.P. Update to International Classification of Diseases, 9th Revision codes: distinguishes STEMI from NSTEMI. *Crit Pathw Cardiol* **4**, 185-186 (2005).

20. Dirks, R.M. & Pierce, N.A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of computational chemistry* **24**, 1664-1677 (2003).

21. Xia, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-14735 (1998).

22. Aalberts, D.P. & Nandagopal, N. A two-length-scale polymer theory for RNA loop free energies and helix stacking. *Rna* **16**, 1350-1355 (2010).

23. Wilkinson, K.A., Merino, E.J. & Weeks, K.M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols* **1**, 1610-1616 (2006).

24. Ray, K.K. & Cannon, C.P. Time to benefit: an emerging concept for assessing the efficacy of statin therapy in cardiovascular disease. *Crit Pathw Cardiol* **4**, 43-45 (2005).

25. Lader, E.W. & Cannon, C.P. How to join a clinical trial. *Crit Pathw Cardiol* **4**, 26-29 (2005).

26. Haas, E.S. & Brown, J.W. Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res* **26**, 4093-4099 (1998).

27. Larsen, N., Samuelsson, T. & Zwieb, C. The Signal Recognition Particle Database (SRPDB). *Nucleic Acids Res* **26**, 177-178 (1998).

28. Wower, J. & Zwieb, C. The tmRNA database (tmRDB). *Nucleic Acids Res* **27**, 167 (1999).

29. Chen, J.L., Blasco, M.A. & Greider, C.W. Secondary structure of vertebrate telomerase RNA. *Cell* **100**, 503-514 (2000).

30. Mortimer, S.A. & Weeks, K.M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129**, 4144-4145 (2007).

31. Wilkinson, K.A., Merino, E.J. & Weeks, K.M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.* **1**, 1610-1616 (2006).

32. Milne, R. & Cannell, M.G. Estimating forest and other terrestrial carbon fluxes at a national scale: the U.K. experience. *SEB Exp Biol Ser*, 57-76 (2005).

33. Tukey, J.W. Bias and confidence in not quite large samples. *Ann. Math Stats.* **29**, 614 (1958).

34. Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C. & Marquet, R. In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem.* **277**, 5995-6004 (2002).

35. Wilkinson, K.A. et al. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.* **6**, e96 (2008).

36. Ren, A., Rajashankar, K.R. & Patel, D.J. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature* **486**, 85-89 (2012).

37. Bizzarri, A.R. & Cannistraro, S. SERS and tunneling spectroscopy investigation of iron-protoporphyrin IX adsorbed on a silver tip. *J Phys Chem B* **109**, 16571-16574 (2005).

38. Mathews, D.H. et al. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* **101**, 7287-7292 (2004).

39. Becerra, R. et al. Time-resolved gas-phase kinetic and quantum chemical studies of the reaction of silylene with nitric oxide. *J Phys Chem A* **109**, 1071-1080 (2005).

40. Fitzgerald, M., Canny, M. & O'Flanagan, D. Vaccination catch-up campaign in response to recent increase in Hib infection in Ireland. *Euro Surveill* **10**, E050929 050922 (2005).

41. Sato, K., Kato, Y., Hamada, M., Akutsu, T. & Asai, K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* **27**, I85-I93 (2011).

42. Wilcox, A.B. et al. Use and impact of a computer-generated patient summary worksheet for primary care. *AMIA Annu Symp Proc*, 824-828 (2005).

43. Clayton, P.D. et al. Physician use of electronic medical records: issues and successes with direct data entry and physician productivity. *AMIA Annu Symp Proc*, 141-145 (2005).

44. Dann, C.E., 3rd et al. Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130**, 878-892 (2007).

45. Cao, S., Giedroc, D.P. & Chen, S.J. Predicting loop-helix tertiary structural contacts in RNA pseudoknots. *RNA* **16**, 538-552 (2010).

46. Diamond, J.M., Turner, D.H. & Mathews, D.H. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry* **40**, 6971-6981 (2001).

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

48. Mathews, D.H., Sabina, J., Zuker, M. & Turner, D.H. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911-940 (1999).

49. Waring, R.B. & Davies, R.W. Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing – a review. *Gene* **28**, 277-291 (1984).

50. Damberger, S.H. & Gutell, R.R. A comparative database of group I intron structures. *Nucleic Acids Res.* **22**, 3508-3510 (1994).

51. Haas, E.S. & Brown, J.W. Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res.* **26**, 4093-4099 (1998).

52. Larsen, N., Samuelsson, T. & Zwieb, C. The Signal Recognition Particle Database (SRPDB). *Nucleic Acids Res.* **26**, 177-178 (1998).

53. Wower, J. & Zwieb, C. The tmRNA database (tmRDB). *Nucleic Acids Res.* **27**, 167 (1999).

54. Zuker, M. On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48-52 (1989).

55. Roth, A. et al. A riboswitch selective for the queuosine precursor preQ(1) contains an unusually small aptamer domain. *Nat. Struct. Mol. Biol.* **14**, 308-317 (2007).

56. Klein, D.J., Edwards, T.E. & Ferre-D'Amare, A.R. Cocrystal structure of a class I preQ(1) riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. *Nat. Struct. Mol. Biol.* **16**, 343-344 (2009).

57. Chen, J.L. & Greider, C.W. Functional analysis of the pseudoknot structure in human telomerase RNA. *P Natl Acad Sci USA* **102**, 8080-8085 (2005).

58. Westhof, E., Dumas, P. & Moras, D. Crystallographic refinement of yeast aspartic acid transfer RNA. *J. Mol. Biol.* **184**, 119-145 (1985).

59. Serganov, A., Polonskaia, A., Phan, A.T., Breaker, R.R. & Patel, D.J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167-1171 (2006).

60. van Batenburg, F.H., Gultyaev, A.P., Pleij, C.W., Ng, J. & Oliehoek, J. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.* **28**, 201-204 (2000).

61. Smith, K.D., Lipchock, S.V., Livinston, A.L., Shanahan, C.A. & Strobel, S.A. Structural and Biochemical Determinants of Ligand Binding by the c-di-GMP Riboswitch. *Biochemistry* **49**, 7351-7359 (2010).

62. Duncan, C.D. & Weeks, K.M. SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA. *Biochemistry* **47**, 8504-8513 (2008).

63. Montange, R.K. & Batey, R.T. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441**, 1172-1175 (2006).

64. Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J.M. & Strobel, S.A. Crystal structure of a self-splicing group I intron with both exons. *Nature* **430**, 45-50 (2004).

65. Garst, A.D., Heroux, A., Rambo, R.P. & Batey, R.T. Crystal structure of the lysine riboswitch regulatory mRNA element. *J. Biol. Chem.* **283**, 22347-22351 (2008).

66. Honda, M., Beard, M.R., Ping, L.H. & Lemon, S.M. A phylogenetically conserved stem-loop structure at the 5' border of the internal ribosome entry site of hepatitis C virus is required for cap-independent viral translation. *J. Virol.* **73**, 1165-1174 (1999).

67. Toor, N. et al. Tertiary architecture of the Oceanobacillus iheyensis group II intron. *Rna* **16**, 57-69 (2010).

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

69. Cannone, J.J. et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics* **3**, 2 (2002).

70. Alonzo, E., Allegra, A.D., Cannizzaro, V., Fardella, M. & La Carrubba, R. [Sanitary and nutritional control of the Catania school catering service]. *Ig Sanita Pubbl* **61**, 285-292 (2005).

71. McGinnis, J.L., Duncan, C.D. & Weeks, K.M. High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. *Methods Enzymol.* **468**, 67-89 (2009).

72. Karabiber, F., McGinnis, J.L., Favorov, O.V. & Weeks, K.M. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *Rna* **19**, 63-73 (2013).

73. Sperschneider, J. & Datta, A. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res.* **38**, e103 (2010).

74. Sperschneider, J., Datta, A. & Wise, M.J. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA* **17**, 27-38 (2011).

75. Reeder, J. & Giegerich, R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC bioinformatics* **5**, 104 (2004).

76. Cannon, C.P. What is the optimal timing of clopidogrel in acute coronary syndromes? *Crit Pathw Cardiol* **4**, 46-50 (2005).

Chapter 3: Identifying pseudoknots in HIV-1 genomic RNA

3.1 Introduction

3.1.1 Pseudoknots perform critical functions in viruses

Folded RNAs contain many different structure motifs. These motifs serve as building blocks for complex RNA architectures[1] and allow the RNA molecules to perform multiple diverse functions. One such structure motif is called a pseudoknot. Pseudoknots form when the loop region of a hairpin base pairs to a region in the RNA molecule outside the hairpin. Pseudoknots are relatively rare, but highly overrepresented in functionally critical motifs. This suggests that pseudoknots are often central components of functional RNA structures, making them attractive drug targets[2, 3].

In viruses, pseudoknots are frequently found in the highly structured regions in the 5' and 3' termini of the untranslated regions (UTR) where they carry out important functions[4]. For instance, a pseudoknot in the internal ribosome entry site (IRES) domain of hepatitis C virus (HCV) serves as a tRNA mimic and positions an initiation codon in the binding cleft of the 40S ribosome, allowing this virus to bypass cellular translational regulation[5]. Another pseudoknot, near the 5' end of human immunodeficiency virus type 1 (HIV-1), is part of the frame-shifting motif that allows genomic RNA to be translated in more than one reading frame to create two unique proteins[6].

Pseudoknots tend to form in close proximity to key functional regions. As a result, correct identification of the pseudoknots motif is fundamental to a structural and

functional understanding of RNA. To accurately predict RNA secondary structure, including pseudoknots, we developed an algorithm called ShapeKnots (see Chapter 2). ShapeKnots combines dynamic programming often used for prediction of RNA structures with experimental information and a simple energy model of the entropic cost of pseudoknot formation. Unlike other algorithms that attempt to predict pseudoknots, ShapeKnots has high prediction accuracy across a variety of RNA lengths and types including pseudoknotted and non-pseudoknotted RNAs. This robust performance and high degree of accuracy makes ShapeKnots an ideal tool to identify pseudoknots, and indications of important biological functions, in a broad range of RNAs.

3.1.2 Using ShapeKnots to identify pseudoknots in HIV-1

In this work, I used the ShapeKnots algorithm to predict the secondary structure and pseudoknot formation within the NL4-3 HIV-1 RNA genome. By utilizing the ShapeKnots algorithm, I was able to identify three potential pseudoknots: pseudoknot 1 (which forms over nucleotides 242-253, 257-261, 263-276, 339-343), pseudoknot 2 (977-981, 986-1000, 1003-1007,1009-1014) and pseudoknot 3 (7249-7253, 7256-7260, 7275-7279, 7318-7322, 7324-7328). To determine whether or not these pseudoknots were likely to form and whether these structures were important in the viral life cycle, I tested these pseudoknots using three different techniques. The first, called mutual information analysis[7-9], determined the probability that nucleotides in the pseudoknot stems co-varied. Co-variation measures the number of instances that base pairing ability is maintained when bases in the pair are mutated[10, 11]. For example, if one lineage has a predicted CG base pair, co-variation would be observed if in another lineage there is an AU pair in the same relative position. The second testing method involved binding a locked nucleic acid

(LNA) oligonucleotide[12-14] to one side of the pseudoknotted helix. I then used the nucleotide-resolution chemical probing technique called SHAPE (for selective 2' hydroxyl acylation analyzed by primer extension) to analyze the reactivities of the LNA-bound RNA. It was hypothesized that the bound LNA would disrupt helical interactions[15] and that the SHAPE reactivities would increase on the other side of the helix. Finally, mutations were made in the predicted pseudoknotted helices and effects were probed using SHAPE and *in virio* studies. As in the LNA binding studies, we expected the mutations to disrupt the pseudoknot resulting in changes in SHAPE reactivity. We also expected that disrupting the pseudoknot would cause a decrease in viral infectivity.

Identifying new pseudoknots in viruses can lead to a better understanding of viral structure. Additionally, since pseudoknots tend to form in key functional regions, their identification can also lead to the identification of potential therapeutic targets. In this work we chose to identify pseudoknots in HIV-1 because of its large complex structure and the previous identification of a pseudoknot at the 5' end.

3.2 Results

3.2.1 Three pseudoknots identified by ShapeKnots algorithm

Pseudoknots are important biological motifs that tend to be located in functionally important, structured regions of RNA. Due to this tendency, discovery and characterization of pseudoknots is critical to understanding the function of an RNA. In this work, we utilized the ShapeKnots algorithm to identify three pseudoknots in HIV-1. Two of the pseudoknots are near the 5' end of the RNA, and the third occurs close to the 3' end of the *env*-coding region (Figure 3.1). The circleplots also shown in Figure 3.1

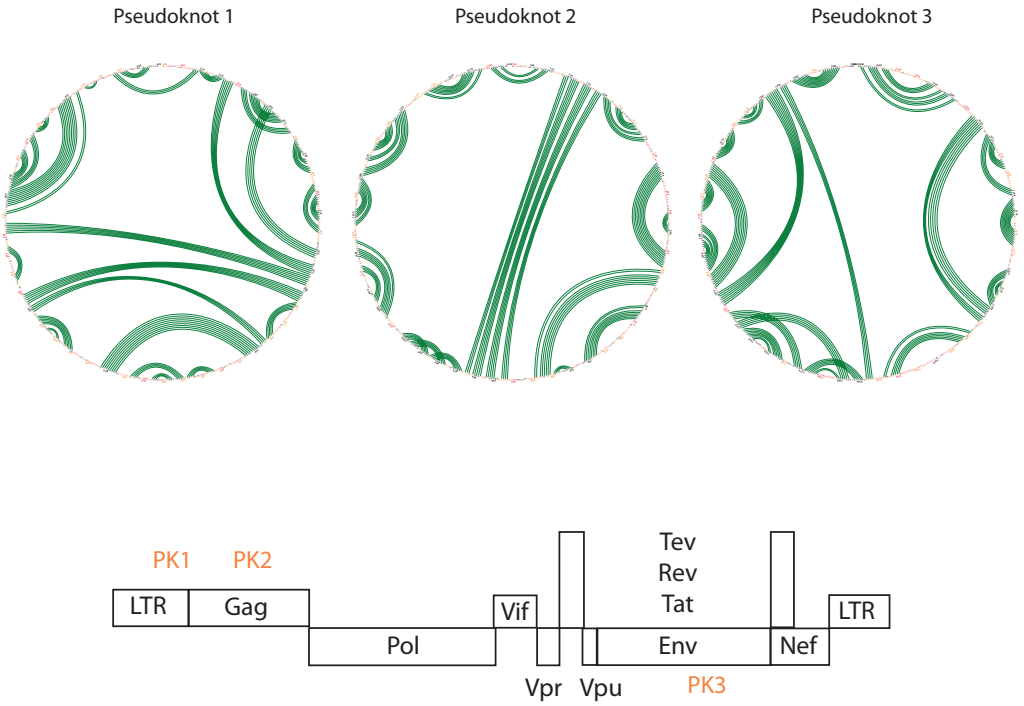

Figure 3.1: Predicted pseudoknots in HIV-1.

Circleplots of three potential pseudoknots in HIV-1 identified using ShapeKnots. The sequence is plotted around the outside of the circle and arcs represent base pairs. Pseudoknots are easy to identify because they form a cross hatching pattern. In the lower part of the figure, a simple schematic of the HIV-1 genome and the locations of the three identified pseudoknots within the genome are shown.

indicate two regions of base pairing predicted to form in each pseudoknot. In a circleplot, the sequence of the RNA is listed around the outside of a circle and the basepairs are indicated as arcs through the circle. Pseudoknots form when the loop region of a helix base pairs to another region in a structure and are easy to identify on circleplots because they form a cross hatching pattern. Color-coded SHAPE reactivities are super-imposed on the nucleotides using a scale from ~0 to 1. The pseudoknots predicted to form in HIV-1 tend to be compact, forming through local RNA-RNA interactions. The regions of predicted pseudoknots have low SHAPE reactivities, indicating that these nucleotides are likely base paired.

3.2.2 Using mutual information to test for evolutionary support for pseudoknots

To test whether the pseudoknots identified with the ShapeKnots algorithm are likely to exist, I looked for evolutionary conservation and co-variation across HIV-1 genomes. Nucleotides are said to co-vary when base-pairing possibility is maintained despite mutation. Mutual information relates co-variation to the probability that a given nucleotide *i* base pairs with a nucleotide *j*. The mutual information is scaled from -1 to 1, where -1 indicates low probability that an *i, j* base pair combination exists and 1 indicates a high probability that the *i, j* base pair combination exists. The distribution of the mutual information for all possible base pair combinations in HIV-1 is given in Figure 3.2. This distribution is fit to a normal curve to identify the 75 and 90 percentiles. These intervals represent the 25% and 10% most highly significant base pair combinations, respectively. The average mutual information for each pseudoknot is indicated by a purple line. The figure demonstrates that all of the averages fall within the top 25% of the possible base

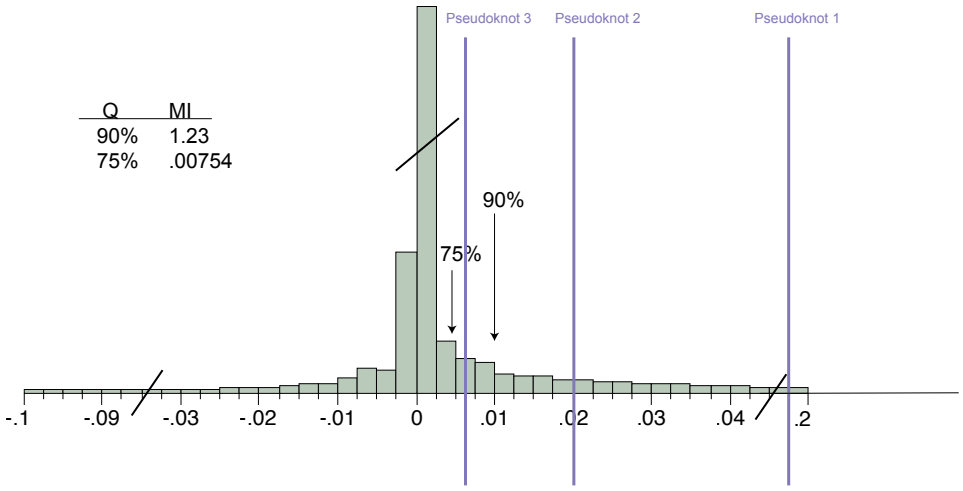

Figure 3.2: Mutual information distribution for all base-pairing conformations in the HIV-1 genome.

The 75% and 90% intervals are highlighted. The averages of mutual information for each pseudoknot are shown in purple.

pair combinations for this RNA, indicating that these base pairs are significant and very likely form in folded HIV-1 RNA.

3.2.3 LNA binding to potential pseudoknot motifs

I then used an LNA binding technique to evaluate formation of two of the pseudoknots (pseudoknots 1 and pseudoknot 2). Due to experimental constraints, the pseudoknot near the 3' end was not tested. LNAs are modified RNA oligonucleotides that contain a sugar-bridging modification that locks the sugar into a 3' endo pucker. LNAs form very stable duplexes with complementary RNA. Due to their low K_d values, LNAs can successfully compete with intramolecular RNA structure. In this technique, I added an LNA complementary to one side of a potential pseudoknotted helix to a sample of HIV-1 RNA and performed SHAPE. Then I compared the bound to unbound reactivities in the predicted helix. I expected that the corresponding side of the LNA-bound helix would increase in SHAPE reactivity indicating that it had gone from a bound to an unbound state. This method produced the expected increase in SHAPE reactivity for the two regions predicted to be part of pseudoknots near the 5' end of the genome. Figure 3.3 shows the results of the LNA experiment evaluating the pseudoknot 2. In the LNA bound case, the nucleotides of the corresponding side of the predicted pseudoknotted helix increased in reactivity.

3.2.4 *In virio* mutants of HIV-1 pseudoknots

To test the role of these pseudoknots *in virio*, we made single nucleotide mutations in our pseudoknotted helices and looked at the viral replication rates.

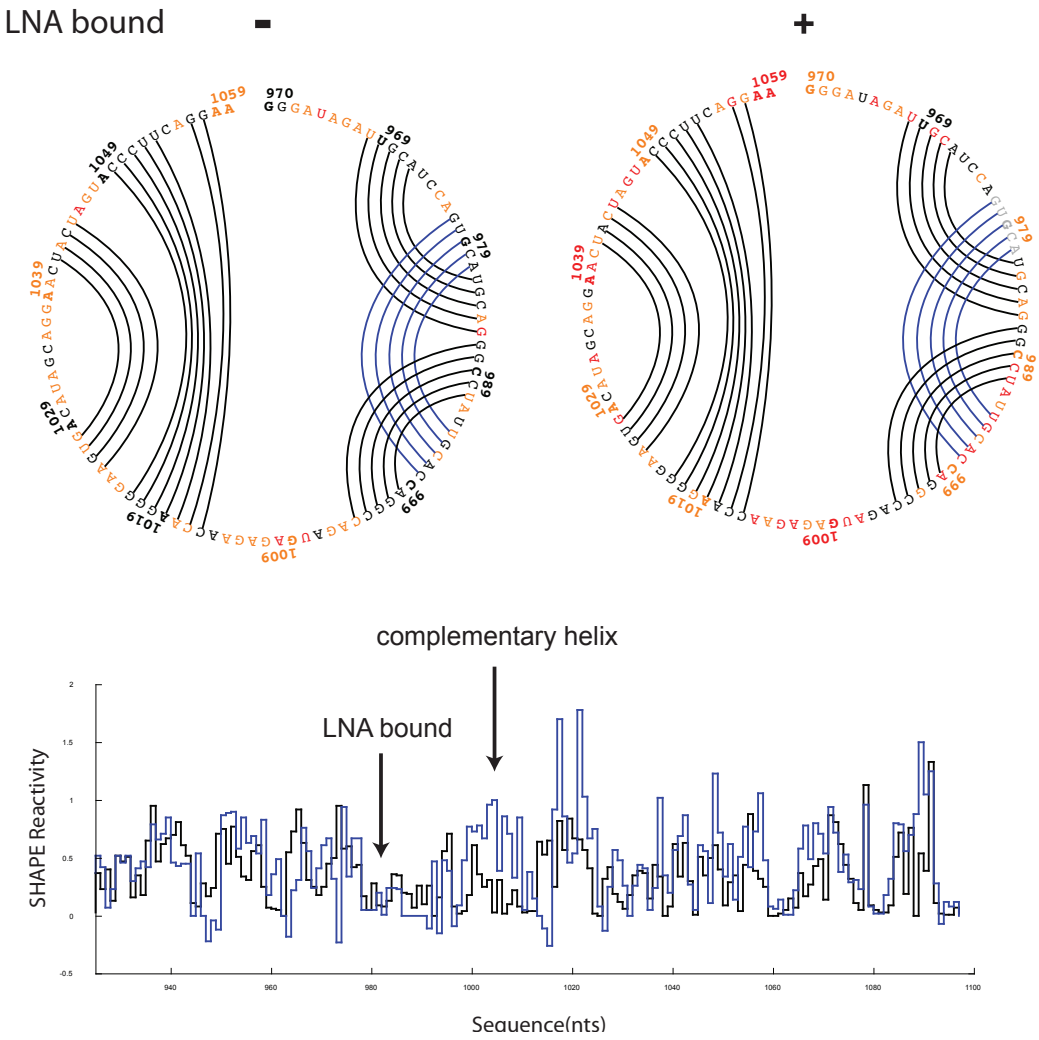

Figure 3.3: LNA binding confirms predicted HIV-1 pseudoknots.

The bound and unbound SHAPE reactivity data is superimposed on the predicted pseudoknotted structure. Highly reactive (>.85) nucleotides are colored in red, medium (.4-.85) orange and low (<.4) black. When the LNA is bound (bound helix colored blue), the SHAPE reactivities in the complementary pseudoknot stem region increase. In the lower part of the figure, the SHAPE reactivity traces for the bound (blue) and unbound (black) SHAPE data are shown.

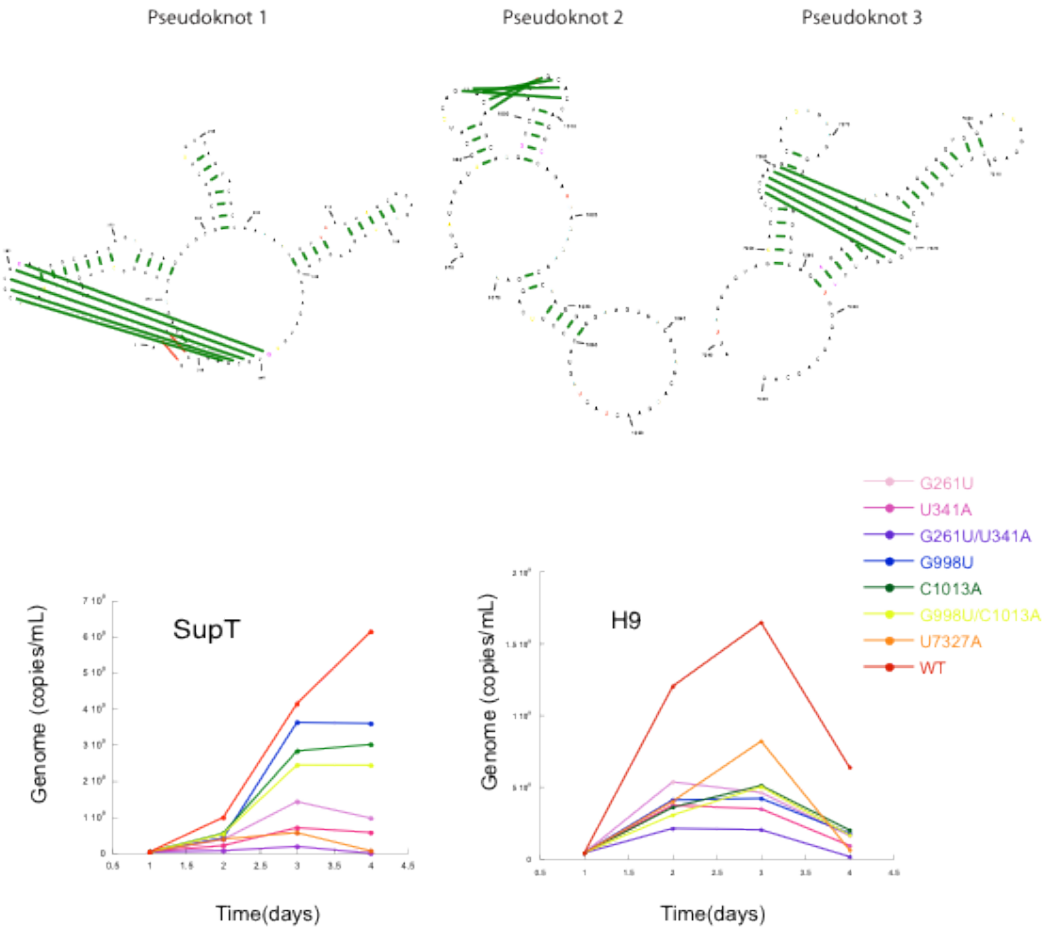

Figure 3.4: Activities of HIV-1 mutants confirm importance of nucleotides in predicted pseudoknots.

Mutations in the pseudoknotted helices are highlighted in pink on secondary structure diagrams. Other colors denote SHAPE reactivity in an analogous manner to Figure 3.3. Viral activities (measured as genome copies/mL) are shown for the SupT and H9 cell lines on the bottom for the following mutants: G261U (light pink), U341A (dark pink), G998U (blue), C1013A (green), U7327A (orange), G261U:U341A (purple), G998U:C1013A (yellow), and wild type (red).

Figure 3.4 shows the genome counts for point mutations made at G261U, U341A, G998U, C1013A, U7327A, G261U:U341A, and G998U:C1013A. The single mutants were designed to be synonymous mutations that disrupted each side of the potential pseudoknotted helix without disrupting the protein coding sequence. The double mutants were designed allow formation of the pseudoknotted helix. All single mutants resulted in a decrease in genome counts relative to the wild type. The double mutants also showed a decrease in replication, indicating that even when the base-pairing pattern was complementary, the base pairs did not reform or reformed but they were not functional. It is possible that all mutations caused a switch in structure away from the wild-type functional structure or that the pseudoknotted helix was a false prediction.

To detect changes in structure due to the mutations, we performed SHAPE on all of the mutants *in virio*. The SHAPE data suggested that both the single and double mutants disrupted the pseudoknotted helices. An example of this is shown in Figure 3.5 for pseudoknot 2. The two single mutants had differences in SHAPE reactivity relative to the wild-type genomic RNA. Additionally, the SHAPE reactivity for the double mutant indicated that the pseudoknotted structure did not reform. We calculated the fold adopted by the pseudoknotted region using the ShapeKnots algorithm with input SHAPE data. The predicted structures for pseudoknot 2 are shown in Figure 3.5. This analysis suggests that the pseudoknot is broken by the single mutations and does not reform with the double mutation. This would account for the low viral counts in the *in virio* study for both single and double mutants.

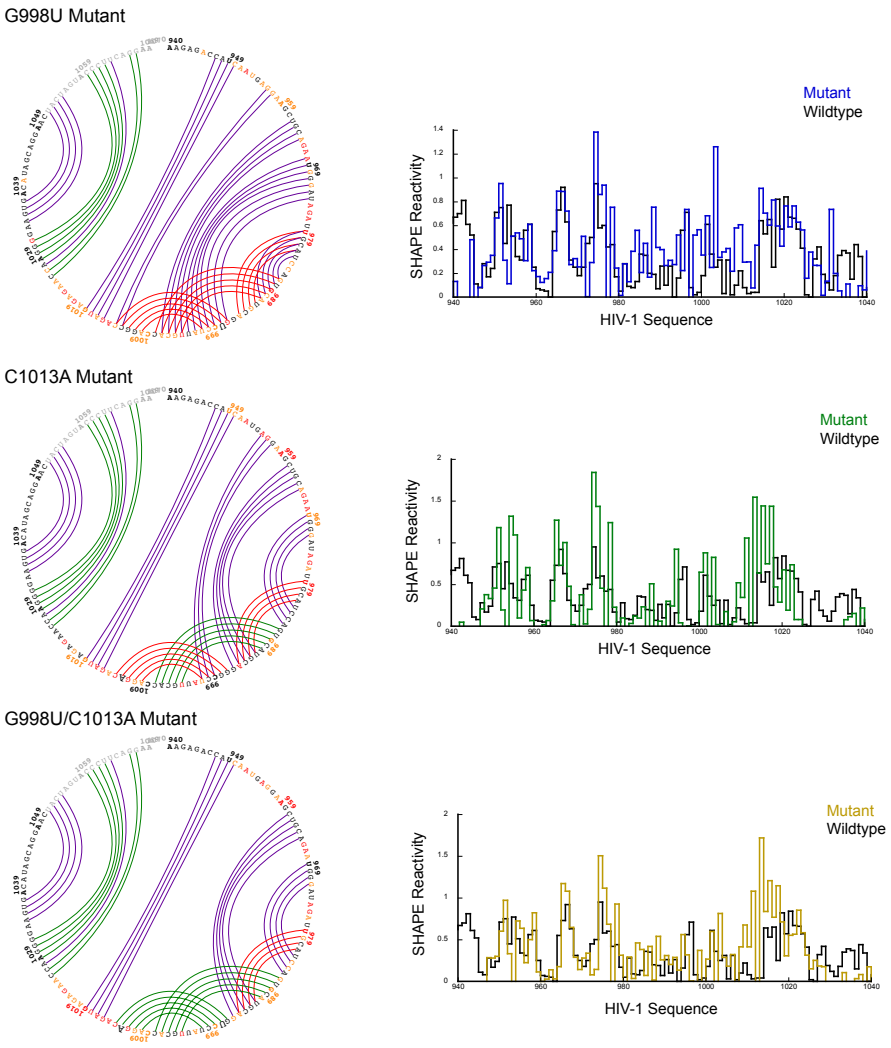

Figure 3.5: Effects of mutations on SHAPE reactivities.

The SHAPE reactivity traces for the G998U, C1013A, and G998U:C1013A mutants verses the wildtype are shown on the right. Nucleotide are colored by their SHAPE reactivity. Highly reactive (>.85) nucleotides are colored in red, medium (.4-.85) orange and low (<.4) black. Base paired predicted in both the mutant and wildtype are shown in green, missing (in mutant) in red, and different (in mutant) in purple.
The SHAPE-directed predictions of secondary structures of mutants are shown on the left.

3.3 Discussion

In this work, I utilized the ShapeKnots algorithm to identify three pseudoknots in the HIV-1 genomic RNA. I then evaluated whether these pseudoknots form using mutual information analysis, LNA binding, and mutagenesis. Upon testing, I found that three of the identified pseudoknots showed high levels of co-variation, indicating evolutionary support for their existence. I also saw changes in SHAPE reactivity compatible with the disruption of the pseudoknotted helix when an LNA was bound, suggesting the formation of the pseudoknots *in vitro*. Finally mutations that should disrupt pseudoknot formation resulted in decreases in viral replication, indicating not only that the pseudoknots likely form but also play a functional role in the replication cycle of the virus.

Two of the identified pseudoknots are predicted to form near the 5' end of the HIV-1 RNA. The 5' end of the HIV-1 RNA is very highly structured and contains many important functional elements such as the tat responsive element (TAR) binding element, the 5' poly A signal, and the primer binding site (PBS)[16]. One pseudoknot forms between nucleotides 242-253, 257-261, 263-276, and 339-343 (pseudoknot 1); this region encompasses the start of the protein coding sequencing (AUG start codon nucleotides 335-338). The structure of pseudoknots allows them to break and reform easily, making them ideal biological switches. We hypothesize that the pseudoknot at the beginning of the coding sequence acts as a method of translational control: turning on and off translation.

Another pseudoknot appears to form between nucleotides 977-981, 986-1000, 1003-1007,1009-1014 (pseudoknot 2) within the *gag-pol* gene. The Gag protein provides the physical infrastructure of the virus; this *gag* gene encodes the viral capsid protein

p24, nucleocapsid proteins p6 and p7, and the matrix protein p17[17]. The small compact nature of this pseudoknot suggests that it may slow translation. Therefore we hypothesize that this pseudoknot may help to slow or even pause translation and to allow processing or folding of the proteins encoded by the *gag* gene.

The third pseudoknot forms between nucleotides 7249-7253, 7256-7260, 7275-7279, 7318-7322, and 7324-7328 (pseudoknot 3). This region is within the portions of the *env* gene that code for gp120 and gp41. These two proteins are processed from the primary translation product of the *env* gene, gp160. We hypothesize that this pseudoknot serves to switch between these two protein-coding regions. This pseudoknot may also provide interesting insight into the structure of the Rev response element (RRE). Previously, the RRE was thought to fold into a rigid, long stem structure,[18, 19] but the formation of the pseudoknot would suggest that the RRE folds into a less rigid domain. The flexibility that would be provided by a pseudoknotted structure is consistent with the general idea that viruses do not fold into long helical structures but rather small base-paired domains[20-22]. This flexibility might be important if the pseudoknot played a role in switching the coding sequence between gp120 and gp41.

One reason that the structure of the RRE domain proposed here differs from previously proposed models may be that I considered the entire HIV-1 genome. In the previous studies, the RNA was characterized by cutting the RRE at the base of the predicted stem and then folding the RNA. This fragment of RNA does not contain bases necessary for pseudoknot formation. Studies suggest that this type of "end folding" effect can cause significant structural rearrangements[23]. Therefore, we hypothesize that the RRE does not form a long helical stem in the context of the full-length viral RNA.

3.3.1 Conclusion

This work demonstrated that the ShapeKnots algorithm accurately identified three pseudoknots in HIV-1. The presence of these pseudoknots was confirmed using a combination of computational and experimental studies. The ability to identify pseudoknots is critical to understanding the function of RNA. As known pseudoknots occur in motifs critical for biological function, this suggests that they can be used as effective drug targets. In the future, the ShapeKnots algorithm can be used to identify pseudoknots in other viral genomes such as HCV, Dengue, and severe acute respiratory syndrome virus (SARS).

3.4 Experimental

3.4.1 SHAPE on HIV-1 RNA

NL4-3 HIV-1 RNA was purified from virions as reported by Watts *et al.*[18]. The extracted RNA was refolded in 50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), and 3 mM $MgCl_2$ and treated with 1M7 (50 mM) in DMSO or with DMSO as a control. Locations of adducts were resolved using capillary electrophoresis as described[24]. Data was processed using custom software[25], and reactivities were scaled from ~0 to 1 using a boxplot normalization[26].

3.4.2 Identification of Pseudoknots

The base-pairing pattern of HIV-1 RNA was calculated using the ShapeKnots algorithm (see Chapter 2) in 600-nucleotide sliding windows that were moved in 100-nucleotide steps, resulting in overlapping 600-nucleotide windows. Input parameters were as follows: m=1.8, b=-0.6, p1=0.35, p2=0.65, window size=0, max structure=100.

To be scored as legitimate, predicted pseudoknots were required to appear in more than one folding window and have low median SHAPE reactivity consistent with structured elements.

3.4.3 Comparison of Mutual Information

HIV-1 sequences were obtained from the Los Alamos HIV database[27]. Mutual information for each possible nucleotide pairing in HIV-1 was calculated using MIFold[8]. Mutual information values were normalized to the degree of variation at each nucleotide. These normalized mutual information values were fit to a Gaussian curve and the 75% and 90% percentiles were computed. The mutual information for each base pair in each possible pseudoknotted helix was compared to the 75% confidence value. If the mutual information was below the 75% confidence value for all base pairs in a helix, that pseudoknotted helix was eliminated from consideration.

3.4.4 Binding of LNAs

To test if my proposed pseudoknots formed *in vitro,* a complementary LNA oligonucleotide was bound to one side of the pseudoknot helix, and SHAPE was performed. Differences between the LNA-bound and unbound SHAPE reactivity profiles were calculated by subtracting reactivities corresponding to the bound state from those corresponding to the unbound state. The statistical significance of these differences was determined by comparing the differences to a standard t-test.

3.4.5 *In Virio* Mutants

To test the existence and functionality of the pseudoknots *in virio,* point mutations were made: G261U, U341A, G998U, C1013A, and U7327A. Additionally, the double

mutants G261U:U341A and G998U:C1013A were constructed as described in Gorelick *et al.*[28]. These mutations were subsequently tested for growth using a virion associated reverse transcription assay in the culture media in H9 and Supt-1 cell lines. The amount of RNA genome copies for each cell line was determined by methods described in Gorelick *et al.*[28].

3.5 References

1. Leontis, N.B., Lescoute, A. & Westhof, E. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16**, 279-287 (2006).

2. Staple, D.W. & Butcher, S.E. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **3**, e213 (2005).

3. ten Dam, E.B., Pleij, C.W. & Bosch, L. RNA pseudoknots: translational frameshifting and readthrough on viral RNAs. *Virus genes* **4**, 121-136 (1990).

4. Stammler, S.N., Cao, S., Chen, S.J. & Giedroc, D.P. A conserved RNA pseudoknot in a putative molecular switch domain of the 3'-untranslated region of coronaviruses is only marginally stable. *Rna* **17**, 1747-1759 (2011).

5. Berry, K.E., Waghray, S. & Doudna, J.A. The HCV IRES pseudoknot positions the initiation codon on the 40S ribosomal subunit. *Rna* **16**, 1559-1569 (2010).

6. Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C. & Marquet, R. In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA. *J Biol Chem* **277**, 5995-6004 (2002).

7. Akmaev, V.R., Kelley, S.T. & Stormo, G.D. A phylogenetic approach to RNA structure prediction. *Proc Int Conf Intell Syst Mol Biol*, 10-17 (1999).

8. Freyhult, E., Moulton, V. & Gardner, P. Predicting RNA structure using mutual information. *Appl Bioinformatics* **4**, 53-59 (2005).

9. Lindgreen, S., Gardner, P.P. & Krogh, A. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics* **22**, 2988-2995 (2006).

10. Eddy, S.R. & Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res* **22**, 2079-2088 (1994).

11. Gutell, R.R., Lee, J.C. & Cannone, J.J. The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol* **12**, 301-310 (2002).

12. Barciszewski, J., Medgaard, M., Koch, T., Kurreck, J. & Erdmann, V.A. Locked nucleic acid aptamers. *Methods Mol Biol* **535**, 165-186 (2009).

13. Braasch, D.A. & Corey, D.R. Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. *Chem Biol* **8**, 1-7 (2001).

14. Rajwanshi, V.K. et al. The Eight Stereoisomers of LNA (Locked Nucleic Acid): A Remarkable Family of Strong RNA Binding Molecules We acknowledge the Danish Natural Science Research Council, the Danish Technical Research

Council, and Exiqon A/S for financial support. Ms Britta M. Dahl is thanked for oligonucleotide synthesis, Dr. Carl E. Olsen for MALDI-MS analysis, and Ms. Karen Jorgensen for recording CD spectra. *Angew Chem Int Ed Engl* **39**, 1656-1659 (2000).

15. Veedu, R.N. & Wengel, J. Locked nucleic acid as a novel class of therapeutic agents. *RNA Biol* **6**, 321-323 (2009).

16. Damgaard, C.K., Andersen, E.S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5' region of the HIV-1 genome. *J Mol Biol* **336**, 369-379 (2004).

17. Freed, E.O. HIV-1 gag proteins: diverse functions in the virus life cycle. *Virology* **251**, 1-15 (1998).

18. Watts, J.M. et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716 (2009).

19. Ippolito, J.A. & Steitz, T.A. The structure of the HIV-1 RRE high affinity rev binding site at 1.6 A resolution. *J Mol Biol* **295**, 711-717 (2000).

20. Zoll, J. et al. Unusual loop-sequence flexibility of the proximal RNA replication element in EMCV. *PLoS One* **6**, e24818 (2011).

21. Fulle, S. & Gohlke, H. Constraint counting on RNA structures: linking flexibility and function. *Methods* **49**, 181-188 (2009).

22. Silverman, A.P., Garforth, S.J., Prasad, V.R. & Kool, E.T. Probing the active site steric flexibility of HIV-1 reverse transcriptase: different constraints for DNA- versus RNA-templated synthesis. *Biochemistry* **47**, 4800-4807 (2008).

23. Wilkinson, K.A. et al. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol* **6**, e96 (2008).

24. Wilkinson, K.A., Merino, E.J. & Weeks, K.M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols* **1**, 1610-1616 (2006).

25. Karabiber, F., McGinnis, J.L., Favorov, O.V. & Weeks, K.M. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *Rna* **19**, 63-73 (2013).

26. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).

27. Kuiken C, F.B., Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, Rambaut A, Wolinsky S, and Korber B in Theoretical Biology and Biophysics Group, Vol. Los Alamos National Laboratory 10-036842010).

28. Gorelick, R.J. et al. Noninfectious human immunodeficiency virus type 1 mutants deficient in genomic RNA. *Journal of virology* **64**, 3207-3211 (1990).

Chapter 4: Using Shannon entropies to calculate the accuracy of secondary structure predictions

4.1 Introduction

4.1.1 Predicting accurate RNA structures is an important goal

RNA molecules are involved in many fundamental cellular processes such as catalysis, transcription, translation, RNA splicing, and RNA editing[1, 2]. These multiple functions are governed largely by the ability of an RNA to fold into complex secondary and tertiary structures[3, 4]. To fully understand the function of RNA in cells and how these macromolecules regulate biological processes it is necessary to understand their structures.

Computational folding algorithms provide an efficient method for determining RNA secondary structure by employing various methods including: statistical sampling, partition function folding, and free energy minimization (MFE)[5-8]. However, many of these traditional RNA folding algorithms suffer from two problems: incomplete and inaccurate energy rules and an inability to predict pseudoknots[9]. Pseudoknots are relatively rare RNA structure motifs that have been identified in or near functional regions in a number of RNAs [10-12].

To overcome these problems, we recently developed ShapeKnots, a dynamic programming algorithm that identifies potential structures by MFE. It is one of the most

accurate prediction algorithms available and is able to refine structures with an average sensitivity of 94% (see Chapter 2). This represents a 20% increase over traditional mFold class algorithms. The success of ShapeKnots can be attributed to its ability to (1) successfully allow and identify pseudoknotted base pairing and (2) incorporate experimental selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) data to refine incomplete energy models[13]. Chapter 2 provides details on the development of and theory behind ShapeKnots.

4.1.2 Identifying the mistakes in predicted structures

Despite the significant advances of the ShapeKnots algorithm, for a few RNAs the accuracy as low as 66%. The lack of accuracy results from one of three problems: (1) Mistakes result from lack of base pairing at the ends of helices or from slightly shifted helices. Most of the mistakes in the ShapeKnots test set and training set fall into this category (see Chapter 2). Such mis-predictions do not change the overall structure of the RNA, are generally viewed as minor mistakes[14] and only change the accuracy about 10%. (2) Most of the structure is correctly predicted, but one or two helices are incorrectly predicted. For example, the *E. coli* 5S rRNA is predicted at a sensitivity of 85%; however, one of the main helices is not predicted correctly, significantly altering the structure (see Chapter 2). (3) The RNA is severely mis-predicted. RNAs with structure induced by protein or ligand interactions or that likely sample multiple conformations create large problems for RNA folding algorithms because most prediction algorithms can only consider one structural conformation at a time and do consider protein binding in folding. This causes RNAs, such the RNA components of RNase P and the human

signal recognition particle (SRP) RNA, to be severely mis-predicted (see Chapter 2). The average accuracy of these two RNAs is only 66%.

Despite these types of mistakes, ShapeKnots predictions give key insights into the likely conformation of a particular RNA. If we could somehow tell *a priori* which regions of the RNA are correctly or incorrectly predicted, we would know which parts of the structure to trust. For this reason, I developed a heuristic way to evaluate, at nucleotide resolution, the regions of the RNA that are likely predicted correctly.

One way to assess the confidence in the predicted fold is to use partition function calculations. The partition function describes the statistical properties of a system in thermodynamic equilibrium and allows for the calculation of base-pair probabilities[5, 6, 15]. The partition function incorporates both the nearest neighbor energy rules and the energies associated with base pairing. Most of the aggregate thermodynamic variables of the system, such as the total energy, free energy, entropy, and SHAPE reactivities, can be expressed in terms of the partition function or its derivatives[5, 6, 15]. Partition function algorithms that are based on these calculations can provide a measure of confidence for MFE structure predictions; however, these algorithms often suffer from the same problems as their original mFold RNA prediction counterparts. They are still built upon incomplete energy rules and often do not allow for pseudoknotted pairings[5, 6, 15].

In this chapter, I describe how we utilized the algorithmic advances that made the ShapeKnots algorithm possible to improve the partition function calculation. To do this, we first expanded the experimental input of our energy function to include both traditional, 1M7, SHAPE[16] and differential, NMIA and 1M6, SHAPE data[17]. The differential data can be used to limit the number of possible structures[17]. Second, we

utilized the pseudoknot prediction capabilities of ShapeKnots. The ShapeKnots algorithm identifies pseudoknots using a topological model for pseudoknot formation. This model estimates the three-dimensional distance over which the pseudoknot forms and then relates that distance to an entropic penalty for pseudoknot formation. Finally, we calculated a Shannon entropy term, as introduced by Huynen *et al.*[18], for each nucleotide, *n*, by summing the probabilities of all potential base-pairing partners for *n*. Since the Shannon entropy is calculated over all possible base-pairing partners, it allows global representation of secondary structure conformations without limitation to a single predicted structure (as is necessary when using a basic partition function calculation to identify accuracy).

By incorporating these changes into the structure prediction algorithm, we were able to calculate the Shannon entropy across an RNA structure and identify regions of structure that are likely accurate and regions of low probably that are likely inaccurate. We also showed that the Shannon entropy and SHAPE data could be used to identify regions of an RNA that are likely to have multiple conformations.

4.2 Results

4.2.1 Identifying the accuracy of secondary structure prediction

As described in this chapter, we sought to develop a useful method of determining the accuracy of structure prediction at nucleotide resolution. To do this, we modified the energy function to incorporate both traditional 1M7 SHAPE[16] and differential SHAPE[17], pseudoknots, and offsets in helices. This modified energy function was incorporated into a partition function calculation and used to sum the Shannon entropy for each nucleotide in an RNA.

In order to test the new algorithm, we used a test set of RNAs chosen to represent those RNAs with complex and generally difficult to predict structures. These RNAs included (*i*) seven RNAs with pseudoknots, (*ii*) four RNAs with structures that are predicted especially poorly with accuracies <60% using nearest-neighbor thermodynamic parameters, and (*iv*) ten RNAs whose structures are modulated by protein and ligand binding.

4.2.2 Calculating the partition function as a Shannon entropy

The Shannon entropy for each nucleotide in an RNA sequence calculated by the equation:

$$\text{Shannon entropy} = \sum P * \log P \qquad (1)$$

where P is the base pairing probability of each *i, j* base pair combination in the RNA. The lower the Shannon entropy the more likely the nucleotide is to exist in a single, highly probable conformation[18]. Unlike traditional partition function calculations, the Shannon entropy can identify both highly probable base pairing and single-stranded regions. It identifies nucleotides when they have one highly probable base pairing partner and when all base pairing probabilities are low and the nucleotide is likely to be single stranded. Therefore, low Shannon entropy corresponds to high probability of a single conformation and provides a convenient way to visualize the data on a single scale.

By calculating the probability as a Shannon entropy, we also allow for a global representation of the structure. Traditional partition function calculations work by identifying the probability of a singular *i, j* base pair conformation. When superimposing these partition function values onto a secondary structure, the partition function changes for each possible suboptimal structure. Although this is helpful in some cases, this

technique requires a pre-identified secondary structure and cannot be used as a general measure to identify regions of RNA that are prone to structural inaccuracies. In one suboptimal structure particular base pairs may be improbable, whereas in another suboptimal structure these same pairs may be highly probable. The Shannon entropy sums over all possible base pair combinations, and thus it can identify regions of high and low probability that are not limited to a single RNA conformation. An example of a global Shannon entropy calculation is shown in Figure 4.1. Figure 4.1 shows the first 2000 nucleotides in HIV-1 genome. Low Shannon entropies (blue) are observed in regions with that have previously characterized structures: the trans-activation response element (TAR) and the dimer initiation site (DIS)[19, 20]. In contrast, higher Shannon entropies (black) correspond to more flexible regions of the RNA.

4.2.3 Subdividing the Shannon entropy

Although the Shannon entropy values are a useful metric for determining the probability of a structure, the raw Shannon entropy values are hard to interpret. They scale based upon the length of the RNA and with the inclusion of SHAPE and differential data. Therefore, we created a method for scaling the Shannon entropies, denoted here as scaled Shannon entropies.

The scaled Shannon entropies were determined by fitting the Shannon entropy distributions for predictions with and without SHAPE data to a beta probability distribution[21, 22] calculated using the following equation:

$$P = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad (2)$$

Here α and β are fit parameters determined to be .27 and 2.33, respectively.

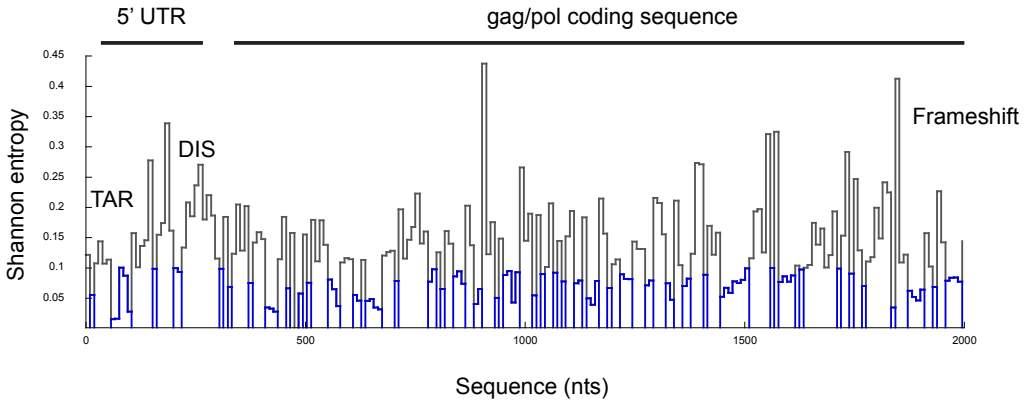

Figure 4.1: Shannon entropies values over the HIV-1 genome.

Shannon entropy values for the first 2000 nucleotides of the HIV-1 genome. Shannon entropy values less than 0.1 are colored in blue. Shannon entropies greater than 0.1 are colored in black. Key functional elements are labeled: TAR, DIS, Frameshift element. The coding region for the 5' UTR and gag/pol coding sequence is also highlighted.

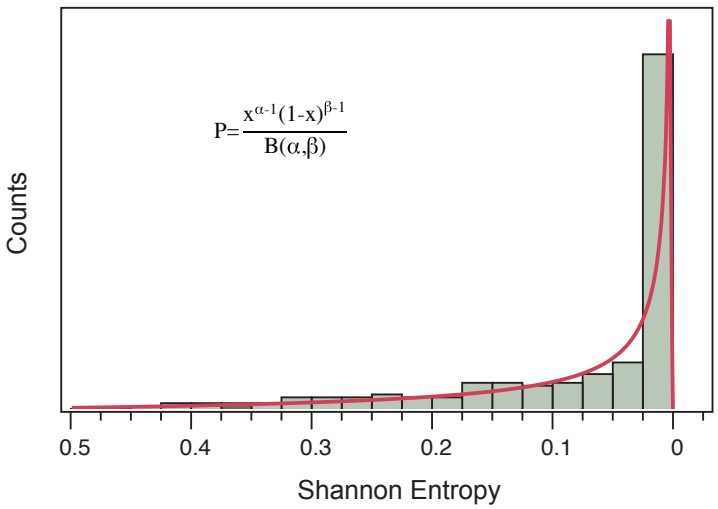

Figure 4.2: Distribution of the Shannon entropies from no SHAPE and SHAPE directed predictions.

The combined distribution for calculated Shannon entropies for the no SHAPE and SHAPE is shown in grey. The x-axis represents the Shannon entropy values and the y-axis represents the number of counts. The Beta curve fit is shown in red.

The distribution of the Shannon entropies is shown in Figure 4.2. The Shannon entropy values were rescaled so that they matched the percentiles of the distribution. For instance, a Shannon entropy value of 0.03, the mean, corresponds to 50%. Since the average sensitivity of an mFold class algorithm is ~73% and the average sensitivity of ShapeKnots is 94% (see Chapter 2), I choose cut off values at 60%, 75%, and 90%. This range assumes that a majority of the Shannon entropies we study come from correctly predicted base pairs.

In Figures 4.3-4.9 the scaled Shannon entropy values are superimposed around the outside of the circleplot (see Chapter 2) as colored stars. Scaled Shannon entropy below 60% are colored black, those between 60 and 75% are dark blue, those between 75 and 90% are light blue, and those between 90 and 100% are white. Stars in dark blue and black indicate high Shannon entropy and poorly defined structures, whereas light blue and white stars indicate low Shannon entropy and highly defined structures. By representing the Shannon entropy in this manner, it is easy to identify regions of the structure that have high and low probabilities.

4.2.4 Offset Helices

Nucleotide offsets occur frequently when predicting RNA structure[23]. An offset occurs when a nucleotide is incorrectly predicted to base pair with a nearest neighbor of the correct nucleotide. Nucleotide offsets can occur for single nucleotides, but usually occur for all nucleotides in a helix. When we examine these offset helices using alternative structure prediction techniques like X-ray crystallography, we generally see that both helices are compatible with the overall fold and topology of the RNA (see Chapter 5). This suggests that this shift in a helix up or down one nucleotide probably

does not affect the overall structure of an RNA and may just reflect a flexible region. Thus, both base pairing patterns are often deemed to be correct. However, since the partition function calculates the probability of each base pair partner individually, the accepted and the shifted helices are identified as low probability. In other words, even though two helices are effectively the same structurally, the partition function treats them like competing structures and the resulting Shannon entropy is artificially high. To account for this local base pairing flexibility we replace the sum of each offset base pair with the sum of the probabilities of the two offset base pairs.

An example of this type of calculation is shown in Figure 4.3. In Figure 4.3A, the dot plot of the most probable base pairs of the TPP riboswitch is shown. Helices are identified by groups of base pairs on diagonal lines. Although most of the probable base pairs occur in distinctive isolated helices, some helices appear in pairs: one member of the pair is correct, the other is offset. For example, this occurs for the helix between nucleotides 6-9 and 38-40 (Figure 4.3B).

The accepted base pairing pattern[24] and the structure predicted for the TPP riboswitch by ShapeKnots are shown in Figure 4.3C-D. The scaled Shannon entropies are superimposed around the outside of the circle. The predicted structure (Figure 4.3C) includes two helices that have low Shannon entropies, but are nearly correctly predicted. In the dot plot (Figure 4.3A-B), these helices correspond to regions where two highly probable helices are next to one another. We treat these two probable helices as a single structure by summing the probabilities for each nucleotide across both helices. When we do this, the Shannon entropy in this region decreases, which more accurately represents the pairing probabilities of these nucleotides (Figure 4.3D). When this factor was

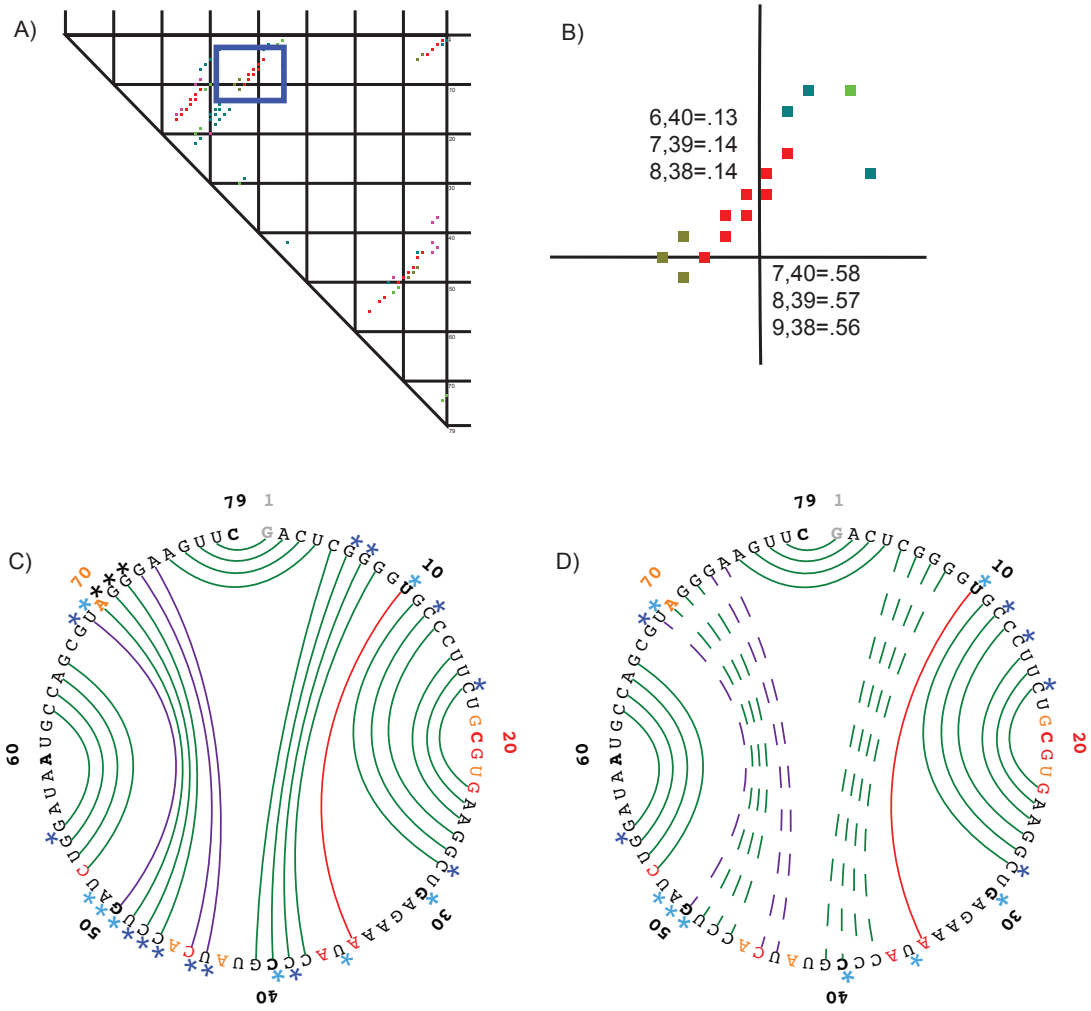

Figure 4.3: Identifying single nucleotide offsets in helices.

A) The dot plot for the TPP riboswitch. Both the x and y axis plot the sequence and base pairs are represented by dots on the plot. The plot is limited to show only the most probable base pairs. The base pairs are ranked by their probability from red (highly probable) to blue (lowly probable).
B) Zoomed –in version of part A corresponding to the highlighted box in A. The probabilities are shown for each nucleotide in decimal format.
C-D) The correct[24] and predicted structures of the TPP riboswitch for when the inline helix correction was not taken into account (left) and when it was taken into account (right). Correctly predicted base pairs are shown in green, mis-predicted in purple, and missing in red. The Shannon entropies are shown around the outside of the circle from black (low probability) to white (high probability). The dotted lines in part D indicate where the helix has two high probable conformations and the Shannon entropies have been altered accordingly.

incorporated into our algorithm, there was a significant increase in the correlation between true low and high probability helices.

4.2.5 Pseudoknot prediction

Pseudoknots are traditionally excluded from RNA structure prediction algorithms due to their tendency to increase calculation time and decrease structurally accuracy. However, pseudoknots are key functional elements and therefore, the correct prediction of them is critical to understanding RNA biology[10, 11, 25, 26]. Previously, we developed a method for identifying correct pseudoknots by RNA topology. To incorporate pseudoknots into the partition function, pseudoknots were first identified using the entropic penalty term found in the ShapeKnots algorithm (see Chapter 2).

$$\Delta G^{\circ}_{PK} = P1\ ln\ (e^2\ SS + f^2\ NE) + P2\ ln\ \Sigma_{IL(n)}(\lambda_n{}^2) \qquad (3)$$

where *P1* is 0.35 kcal/mol, *P2* is 0.65 kcal/mol, and λ_n is the penalty constant for in-line helices of length *n*. The first term penalizes formation of pseudoknots with long single-stranded regions and many nested helices, whereas the second term enforces an optimal geometry for in-line helices. Once identified, the pseudoknot is incorporated into the partition function by breaking the pseudoknot into two sets of helices (Figure 4.4). The first set considers the one pseudoknot helix that crosses the most base pairs. The second set considers all remaining pseudoknotted helices. For each pseudoknot, the partition function is run twice, while holding out each set of pseudoknotted helices. The final base probabilities are then calculated as the geometric average of each *i. j* base pair probability (see Experimental).

By incorporating pseudoknots into the partition function calculation, the probabilities more accurately identify regions of instability compared to traditional partition function that do not include pseudoknots. For example, Figure 4.4 shows the ShapeKnots predictions for the secondary structures, including pseudoknots, of the SAM I riboswitch[27] and the *Azoarcus* group I intron[28] compared to structures determined by crystallography. In both cases, the predicted structure matches well with the accepted structure. The left side of the figure shows the structures predicted and the Shannon entropies when pseudoknotting is not incorporated into the partition function. In these cases, the Shannon entropies are high around the pseudoknot. These high entropies suggest that the pseudoknot is incorrect. When the partition function includes the pseudoknot calculation (right side), the Shannon entropies are low for the entire RNA indicating that it is a high quality prediction.

4.2.6 Incorporating differential SHAPE data

Previous studies have demonstrated how incorporating 1M7 SHAPE data increases accuracy of RNA structure prediction algorithms[13]. In this work, we incorporate a second SHAPE energy term called differential SHAPE. Differential SHAPE is calculated by subtracting NMIA reactivity at each nucleotide from 1M6 reactivity[17]. NMIA and 1M6 have very different half-lives. The difference in reactivity between these two reaction times provides structural information. For instance, nucleotides that are more reactive to 1M6 than to NMIA tend to occur at the ends of helices and near tertiary interactions. Conversely, nucleotides that are more reactive to NMIA than to 1M6 tend to be extremely flexible[17]. By incorporating these data into our

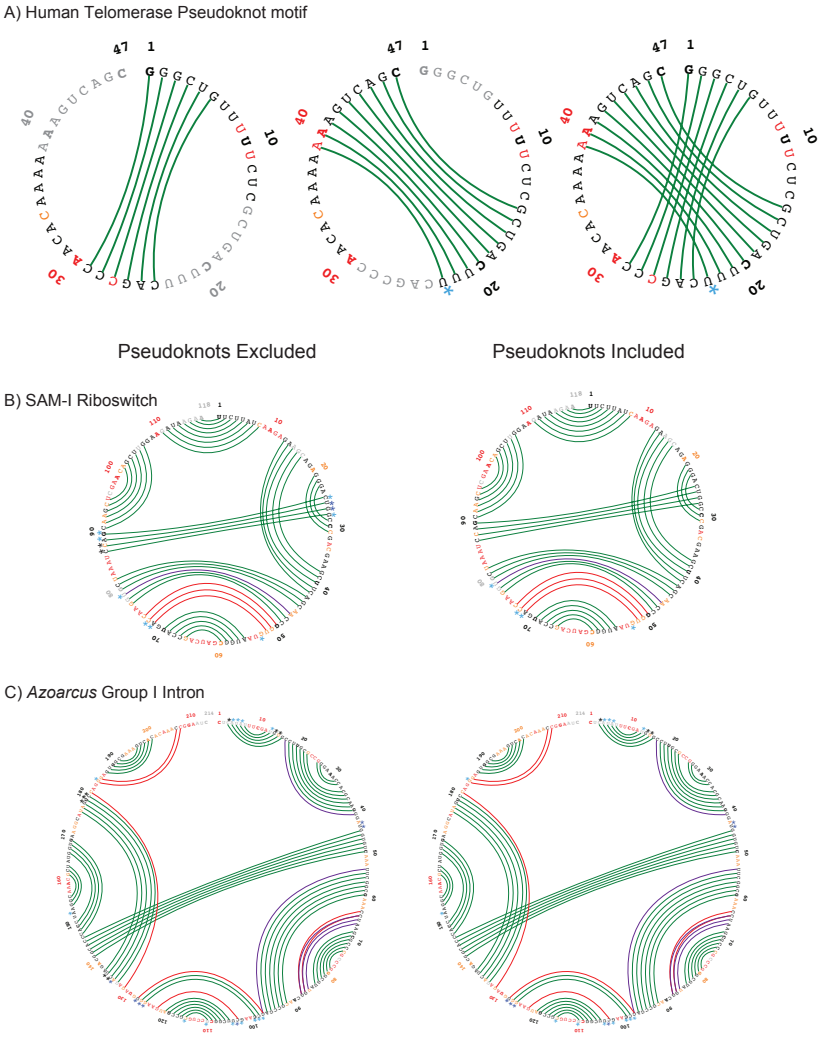

Figure 4.4: Predicting Shannon entropies with pseudoknots.

(A) Pseudoknot prediction method. The pseudoknotted helices are grouped into two sets (see Experimental). For each set, the corresponding nucleotides are held out (forced single stranded) and the Shannon entropies for the remaining nucleotides are calculated. On the left the pseudoknot in grey is held out, and the calculated Shannon entropies are super imposed. In the middle, the process is repeated for the other pseudoknotted helix. Finally the two sets of Shannon entropies are combined (see Experimental) to produce the final structure.

(B-C) The Shannon entropies of the SAMI riboswitch [27], and Azoarcus Group I[28] compared to the conventional crystal structures, both without(left) and with(right) the inclusion of pseudoknots. Structural annotations that same as described in Figure 4.2.

Shannon entropy calculation, we can further refine the probabilities and more correctly distinguish which parts of the RNA are correct. Figure 4.5 shows the predictions of the secondary structure of *E. coli* 5S rRNA[29] incorporating no SHAPE data, SHAPE data obtained with 1M7, and SHAPE data obtained with 1M7 plus differential SHAPE. Without SHAPE data, a majority of the structure is mis-predicted (see Chapter 2). Shannon entropy values are high indicating that the confidence in the structure predicted is low. The single helix that is correctly predicted has low Shannon entropies, shown in white. When SHAPE data was incorporated, the prediction accuracy increased to 85% sensitivity, and the Shannon entropies decreased. The single mis-predicted helix has high Shannon entropy. When the differential reactivity data was included, the prediction increased to nearly 100%, and most of the nucleotides have low Shannon entropies. This indicates that the structure is most likely correctly predicted. This example shows us that 1) use of SHAPE data increases accuracy of structure prediction and use of differential SHAPE data increases it even more and 2) regions that are predicted incorrectly compared to the accepted structure always have higher Shannon entropies than correctly predicted regions.

4.3 Discussion

Identifying the probability of accuracy of a structure allows for the differentiation between correctly predicted structures that can be used for future biological hypothesis and those that are not meaningful. In this work, we refined the partition function calculation to better differentiate between highly and lowly probable base pairs.

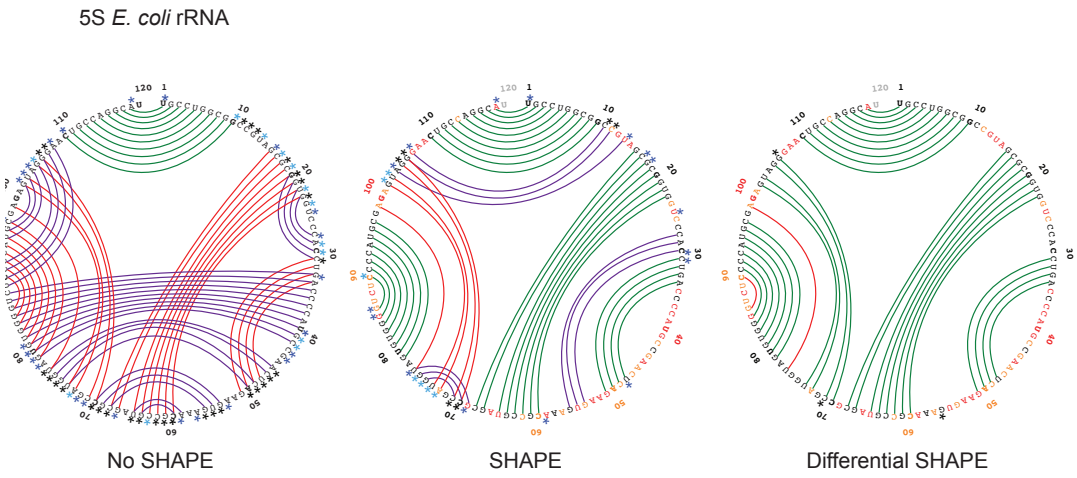

Figure 4.5: 5S *E. coli* no SHAPE, 1M7 and differential SHAPE

5S *E. coli* NoSHAPE (left), SHAPE (middle) and differential(right) predictions as compared to conventional crystal structure[29] . Structural annotation is the same as Figure 4.2.

4.3.2 *E. coli* 16S and 23S rRNAs

Figure 4.6 shows the comparison of the predicted structures versus secondary structures obtained from crystallographic data on the 16S and 23S rRNAs[30-32]. When the 16S and 23S rRNAs were folded without SHAPE data and the calculated Shannon entropies were superimposed on the structures, the predictions correlate well with the Shannon entropies. This means that mistakenly predicted nucleotides have high Shannon entropy and correctly predicted areas have low Shannon entropy.

When 1M7 SHAPE data was incorporated into the ShapeKnots algorithm for 16S and 23S rRNAs, the accuracy of structure prediction increased from ~65% to 90% and from 75% to 88%, respectively (Figure 4.6). In the 23S rRNA SHAPE-based prediction, the accuracy of the structure improved, but four helices observed in the accepted structure are mis-predicted. The Shannon entropies are high in these mis-predicted areas.

In the SHAPE-directed 16S rRNA structure prediction, most of the structure is correctly predicted and has low Shannon entropy. An exception is the region between nucleotides 117-193. The Shannon entropies are relatively low for this region, indicating that the predicted 16S structure is highly probable. This seems contradictory, but when we look closer at this region, the SHAPE reactivities match the predicted structure better than the accepted structure. For instance, nucleotides 117-122 are single stranded in the accepted structure but were relatively unreactive to SHAPE reagent. Previously, Deigan *et al.* used a similar SHAPE-directed folding approach to predict the 16S rRNA structure and saw the same alternative structure between nucleotides 117-122[13]. They tested the structure with follow up experiments and showed that their predicted structure was likely to occur *in vitro.* The calculated Shannon entropies, therefore, support the conclusions

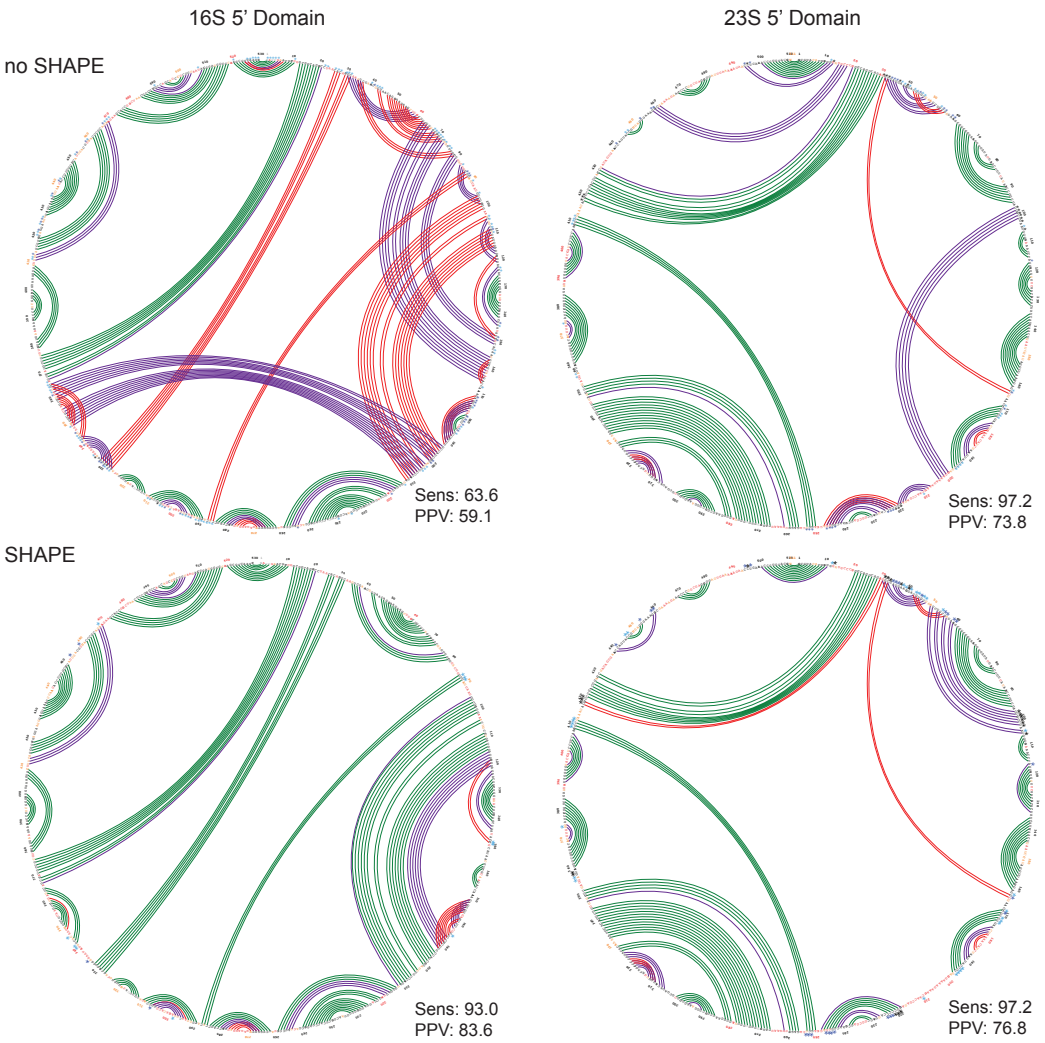

Figure 4.6: 16S and 23S rRNA no SHAPE and 1M7 SHAPE predictions and superimposed Shannon entropies

No SHAPE (top) and SHAPE (bottom) directed predictions of the 16S(left) and 23S (right). Structural annotations the same as Figure 4.2.

made by Deigan *et al.* and suggest that the SHAPE predicted structure is highly probable. This example shows that the Shannon entropies can identify regions of high and low probability and that we can use the Shannon entropies to identify structures in RNA that should be evaluated experimentally.

4.3.3 Signal Recognition Particle

The SHAPE-directed prediction of the SRP protein is shown in Figure 4.7 in comparison to the crystal structure[33, 34 35]. The sensitivity for the prediction is very low. When 1M7 data were included, the sensitivity was 66%. When both 1M7 and differential SHAPE were used to direct the prediction, the sensitivity was only 45%. In particular, the region between 130-270 is entirely mis-predicted. Despite the differences between the predicted and the accepted structure, the Shannon entropies are low, suggesting that the probability of the structure is high. As in the 16S rRNA case, the SHAPE reactivities match the SHAPE-directed structure, but contradict the structure determined from X-ray crystallography[33, 34]. Therefore, again, the SHAPE data suggest that the RNA is not forming the same structure under probing conditions as it does upon crystallization.

One hypothesis for this alternative structure is the difference in experimental conditions between crystallization and SHAPE. In the crystal, several SRP proteins are bound to the RNA. A schematic of the structure in the region between 140- 236 [33] is shown in Figure 7C. As three proteins bind in the region between 130-270[33, 34], we hypothesized that these proteins cause a conformational change relative to the free RNA.

To test this hypothesis I refolded the SRP RNA in the presence of proteins and repeated the SHAPE experiment.

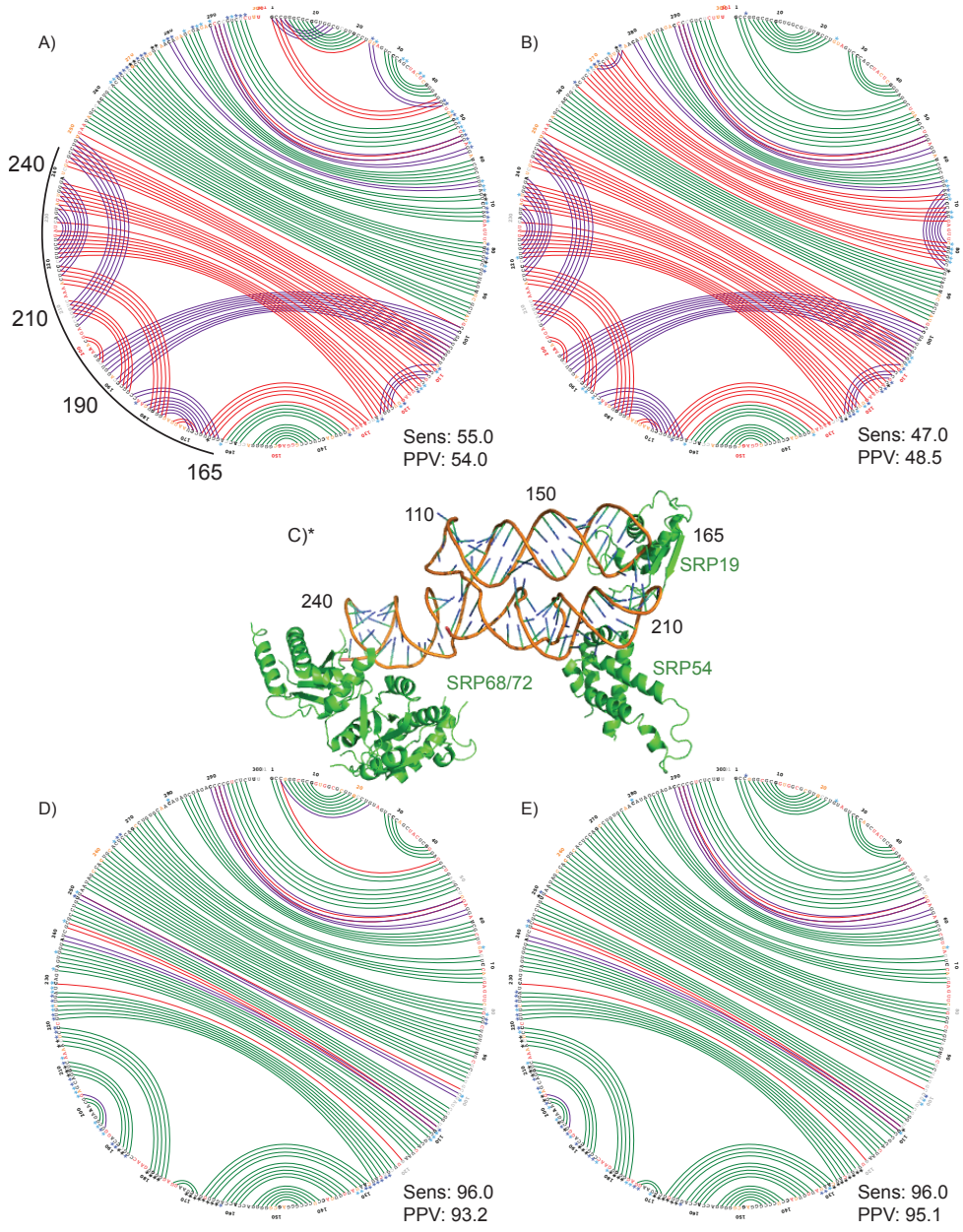

Figure 4.7: Signal Recognition Particle 1M7 and differential SHAPE predictions

A-B,D-E) SHAPE predicted structure for the SRP RNA versus the conventional structure[35 33]. Protein free structures are shown on the top, while protein bound structures are shown on the bottom. Traditional 1M7 SHAPE predictions are shown on the left and differential SHAPE is shown on the right.
C) Portion of the SRP crystal structure highlighting bound protein regions[35 33]

The SHAPE-directed protein-bound structures are shown in Figure 4.7D and 4.7E. The sensitivity of the prediction increased to 93% for the 1M7 and 95% for the differential SHAPE. The Shannon entropies of the majority of the nucleotides are low, indicating that the structure is highly probable. However, the Shannon entropies are still relatively high around the region 160-220. This is the region where the SRP-19 protein is bound and where SRP-54 and SRP-68/72 should bind[35 33] (Figure 4.7C). Since only SRP-19 was used in this experiment, we expect that there may be some nucleotide flexibility and lower probability in this region due to these missing protein interactions. This analysis demonstrates that the Shannon entropies can be used to determine highly probable structures and can even help to identify structure prone to conformational switches as proteins bind.

4.3.4 Other small RNAs

The Shannon entropy calculation was repeated for the no SHAPE, SHAPE, and differential SHAPE cases for all small (<100 nts) RNAs in our test set. Examples of two small RNAs (the cyclic di-GMP and the adenine riboswitch) are shown in Figure 4.8. These RNAs tend to have well-defined structures that are accurately predicted even without the inclusion of SHAPE data with sensitivities for the cyclic di-GMP and adenine riboswitch of 85% and 100%, respectively, without SHAPE data. The Shannon entropies for the cyclic di-GMP nucleotides correctly identify the mistakes within the structure relative to the accepted structure[36]. The Shannon entropies are relatively high for the adenine riboswitch. When 1M7 and differential SHAPE data were incorporated the accuracy was high and the Shannon entropies were lower. This suggests

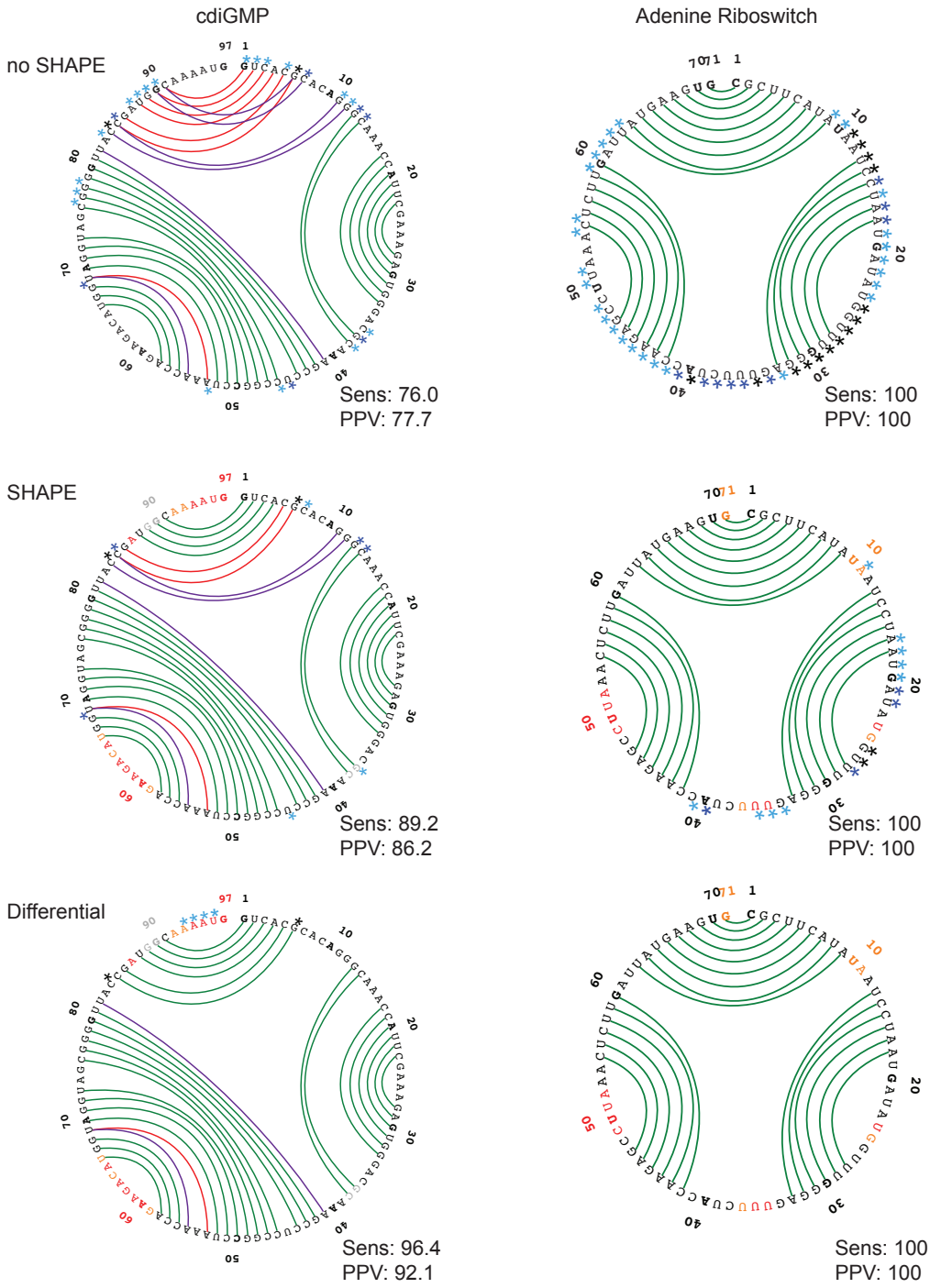

Figure 4.8: Shannon entropy calculations for small RNA predictions.

The no SHAPE(top), 1M7(middle) and differential(bottom) SHAPE predictions for the Adenine riboswitch(right) and cyclic diGMP(left). Structural annotations are the same as Figure 4.2.

that the Shannon entropy calculation is robust, but without the inclusion of SHAPE data probabilities of structural predictions may be low.

4.3.5 Other large RNAs

Large RNAs often pose potential problems for RNA structure determination. Two sample predictions for large RNAs (the lysine riboswitch and a group I intron) are shown in Figure 4.9. The incorporation of SHAPE data increased the structural prediction accuracies and decreased the Shannon entropies for both. However, in the case of the group I intron, incorporation of differential data caused mis-incorporation of a pseudoknot. Due to the nature of differential reactivity, slow nucleotides tend to be more reactive around pseudoknots. This is a potential flaw with the method. Further studies will need to be done to identify whether or not differential reactivity is consistent with the prediction of pseudoknots.

4.3.6 Conclusion

Identifying the probability of accuracy of a structure allows researchers to distinguish between correctly predicted structures that can be used to make hypotheses regarding function and those that are not meaningful. In this work, we refined the partition function calculation to better differentiate between high and low probability base pairs.

The ShapeKnots algorithm has drastically increased the accuracy of RNA secondary structure prediction, but the algorithm still has a range of accuracies between

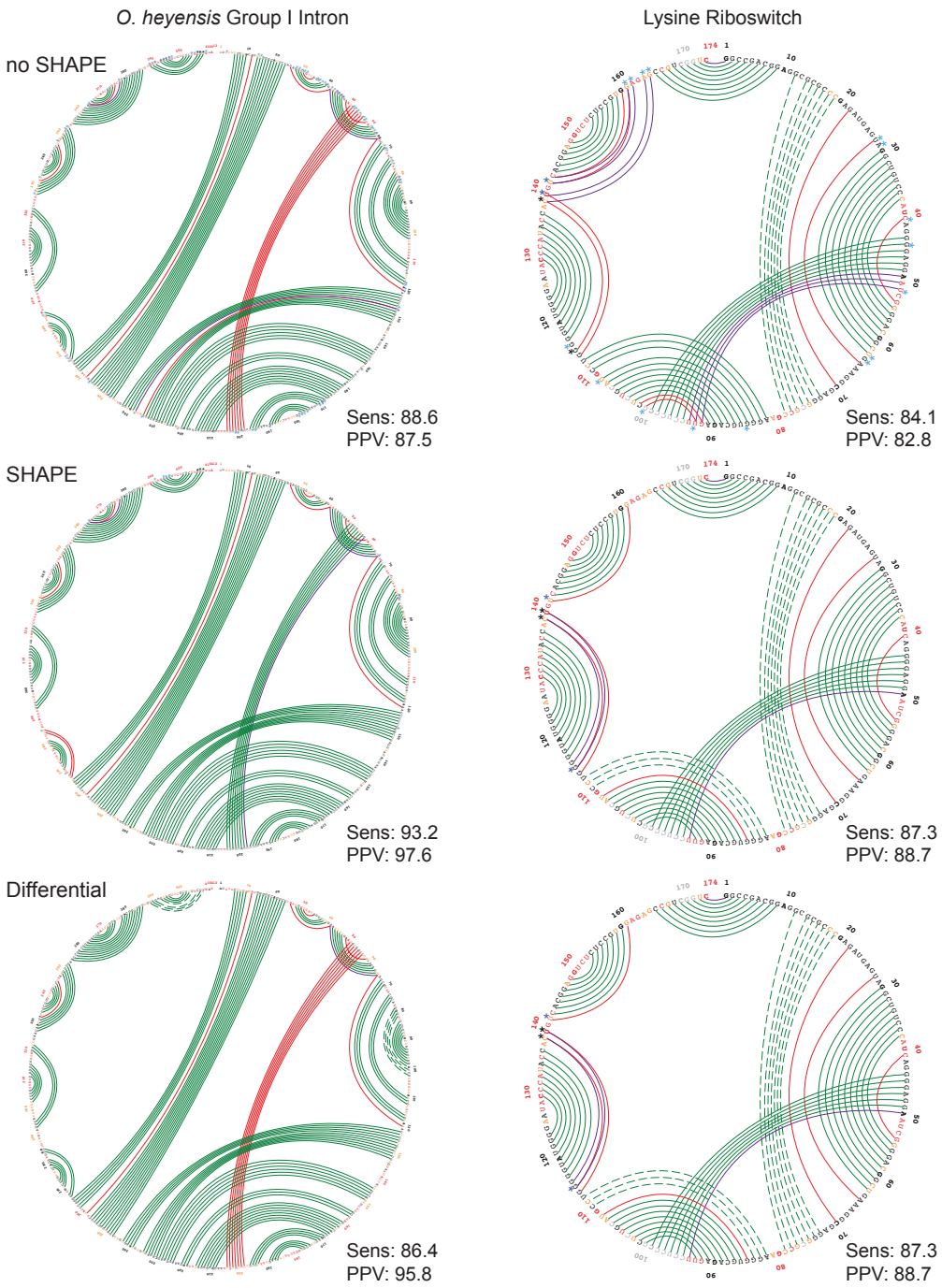

Figure 4.9: Shannon entropy calculations for large RNA predictions.

The no SHAPE(top), 1M7(middle) and differential(bottom) SHAPE predictions for the Lysine riboswitch(left) and Group I Intron(right). Structural annotations are the same as Figure 4.2.

66-100 %. In this work, I showed how we can utilize the Shannon entropy calculations to differentiate between regions in an RNA where there is high and low confidence in the structure predicted. Although this does not directly improve the accuracy of RNA structure prediction, it tells us which parts of the predicted structure to trust. I show that we can use this to gain a greater understanding of the folding patterns of an RNA without a structure prediction and that we can use these values to identify the probability that a predicted structure is accurate. I also show how we can utilize Shannon entropies and SHAPE data to identify regions in RNAs that can form different structures depending on solution conditions or the presence or absence of co-factors. In the future, this technique can be further used to identify the probability of structures or identify structures that have the potential to undergo a conformational switch.

4.4 Experimental

4.4.1 RNA preparation and SHAPE modification.

The PreQ1 riboswitch, human telomerase RNA, TPP riboswitch, adenine riboswitch, cyclic diGMP, SAMI riboswitch, mBox riboswitch, P546, *E. coli* 5S rRNA, *Azoarcus* group I Intron, lysine riboswitch, RNase P RNA, *Tetrahymena* group I Intron, *Oceanobacillis inheyensis* group II Intron, 5' domain of the 23S rRNA, the 5' domain of the 16S rRNA, and the HIV-1 genome were purified as described in Chapter 2. The purified RNAs were then folded in a standard buffer with physiologically relevant ion concentrations (as described in Chapter 2) and treated with 1M7. The TPP riboswitch, cyclic diGMP, adenine riboswitch, 5S rRNA, *Azoarcus* group I Intron, lysine riboswitch, RNase P RNA, *Tetrahymena* group I intron, and *O. inheyensis* Group II intron RNAs

were also treated 1M6 and NMIA for 2 minutes and 30 minutes, respectively, in separate experiments. Sites of 2'-*O*-adduct formation were detected by primer extension using a previously described high-throughput SHAPE approach[16] and processed using custom software. SHAPE reactivities for 1M7, 1M6, and NMIA were normalized to place them on a scale from zero (unreactive) to ~1.5 (highly reactive) as described in Chapter 2. In Figures 3-9 the SHAPE data and SHAPE-predicted secondary structure are plotted on circle plots against the conventional accepted structure. The nucleotide sequence is arrayed on the outer circle: unreactive nucleotides (SHAPE reactivities < 0.4) are colored black, moderately reactive nucleotides (0.4 – 0.85) are yellow, and highly reactive nucleotides (reactivities > 0.85) are red.

4.4.2 Signal Recognition Particle Protein and RNA preparation and modification

SRP-19 protein was purified using the previously described methods[37] and placed in protein dilution buffer (300 mM KOAc, 20 mM Hepes, pH 8, 5 mM $MgCl_2$). The RNA was made from linear transcripts as described in Chapter 2. The protein and RNA were mixed in a 1:1 ratio in 1:5 volume then heated to 80 °C and snapped cooled. The RNA-protein complex was then allowed to fold at 37 °C for 2-3 minutes. After folding, the complex modified using 1M7, 1M6, and NMIA [38] in separate reactions and the RNA was purified using a RNA cleanup kit (Biogen). Sites of 2'-*O*-adduct formation were detected by primer extension as previously described[16, 39] and were processed using custom software. SHAPE reactivities were normalized as described above.

4.4.3 Shannon entropy calculation

This is a heuristic method that allows for the differentiation of highly confident predicted structural features from low confidence features. It is based upon the partition function calculation that is found in the RNA structure platform[5, 6]. The processing steps of the modified algorithm are as follows:

1) Pseudoknotted helices are generated using a method analogous to that found in the ShapeKnots algorithm (see Chapter 2). In this method, a candidate pseudoknot helix list, H, along with the corresponding helix energies is generated from the energy dot plot. Helix H_i is accepted into H if it spans at least three base-pairs and occurs in a structure with a $\Delta G°$ within 25% of the free energy of minimum free energy structure (S_{mfe}), H is trimmed to include a maximum of 100 of the most thermodynamically stable helices.

   For each H_i, a new set of structures, composed of the lowest free energy structure and up to 100 suboptimal structures, is generated by the dynamic programming algorithm, where all nucleotides in H_i are prohibited from pairing (forced single-stranded)[40]. After these structures have been generated, base pairs from H_i are restored to the structures. The $\Delta G°$ of each structure is incremented by the free energy of the corresponding helix H_i. For each structure that contains a pseudoknot, the entropic cost of pseudoknot formation is penalized by $\Delta G°_{PK}$ (Eqn. 3). For pseudoknots in structures within the top 10% of the S_{mfe}, limited to no more than 100 structures, the pseudoknotted helices are added to a list of pseudoknots, P.

2) For each P_i in P, the pseudoknotted helices are grouped into two categories. The first represents the singular helix that once removed will abolish the pseudoknot (P_k). When only two helices make up the pseudoknot, the longer of the two is identified. If they are the same length the 5' most pseudoknotted helix is identified. The second category represents all other helices that are part of the pseudoknot. The partition function and Shannon entropy calculation is run twice for each pseudoknot. First, the nucleotides within P_k is forced to be single stranded, and then all other helices involved in the pseudoknot are prohibited from pairing. For each separate pseudoknot structure, the Shannon entropy is determined when its P_k is not prohibited from pairing. For all base pairs not included in the pseudoknot, the entropy is defined as the geometric average of the two Shannon entropies. This process is shown in Figure 4.3 for the human telomerase RNA.

3) If two helices with greater than 10% probability[41] are found within one nucleotide either direction, the probabilities of each base pair in the helix is replaced by the sum of the two probabilities. This accounts for the tendency of base pairs to slip by one nucleotide.

4) SHAPE data is read into the program using the equation:

$$\Delta G_{SHAPE} = m * \ln(\text{SHAPE} + 1) + b \quad (4)$$

Where m=1.8 and b=-0.6 (See optimization of m and b in Chapter 2)

5) Differential SHAPE is read into the program using the equation[42]:

$$\Delta G_{\text{Dif}} = m * \text{DifferentialSHAPE} \quad (5)$$

Where m=2.1

6) The Shannon entropy is calculated as in (Eqn. 1).

4.4.4 Color Distribution.

To identify which Shannon entropies are low (indicating well-defined structure) and which entropies are high (indicating a less defined structure), we fit the Shannon entropies of the 16 RNAs in our test set to a beta distribution[21, 22] (Eqn. 2). The 60%, 75%, and 90% intervals of the distribution were calculated.

4.5 References

1. Brierley, I., Pennell, S. & Gilbert, R.J. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol* **5**, 598-610 (2007).

2. Coffin, J.M., Hughes, S.H. & Varmus, H.E. in Retroviruses. (eds. J.M. Coffin, S.H. Hughes & H.E. Varmus)Cold Spring Harbor (NY); 1997).

3. Montange, R.K. & Batey, R.T. Riboswitches: emerging themes in RNA structure and function. *Annu Rev Biophys* **37**, 117-133 (2008).

4. Berkhout, B. Structure and function of the human immunodeficiency virus leader RNA. *Prog Nucleic Acid Res Mol Biol* **54**, 1-34 (1996).

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

6. Mathews, D.H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna* **10**, 1178-1190 (2004).

7. Mathews, D.H., Turner, D.H. & Zuker, M. RNA secondary structure prediction. *Curr Protoc Nucleic Acid Chem* **Chapter 11**, Unit 11 12 (2007).

8. Ren, J., Rastegari, B., Condon, A. & Hoos, H.H. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *Rna* **11**, 1494-1504 (2005).

9. Liu, B., Mathews, D.H. & Turner, D.H. RNA pseudoknots: folding and finding. *F100 Biol. Rep.* **2**, 8 (2010).

10. Staple, D.W. & Butcher, S.E. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* **3**, e213 (2005).

11. Brierley, I., Pennell, S. & Gilbert, R.J. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.* **5**, 598-610 (2007).

12. Pleij, C.W. Pseudoknots: a new motif in the RNA game. *Trends Biochem. Sci.* **15**, 143-147 (1990).

13. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).

14. Lu, Z.J., Gloor, J.W. & Mathews, D.H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *Rna* **15**, 1805-1813 (2009).

15. Waldispuhl, J. & Clote, P. 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 (2007).

16. Wilkinson, K.A. et al. Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *Rna* **15**, 1314-1321 (2009).

17. Steen, K.A., Rice, G.M. & Weeks, K.M. Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *J Am Chem Soc* **134**, 13160-13163 (2012).

18. Huynen, M., Gutell, R. & Konings, D. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol* **267**, 1104-1112 (1997).

19. Freisz, S. et al. Sequence and structure requirements for specific recognition of HIV-1 TAR and DIS RNA by the HIV-1 Vif protein. *RNA Biol* **9**, 966-977 (2012).

20. Frankel, A.D. & Young, J.A. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem* **67**, 1-25 (1998).

21. Pearson, K. Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of LondonARRAY* **186**, 343–414 (1895).

22. Pearson, K. Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **216**, 538–548 (1916).

23. Freyhult, E., Moulton, V. & Gardner, P. Predicting RNA structure using mutual information. *Appl Bioinformatics* **4**, 53-59 (2005).

24. Thore, S., Leibundgut, M. & Ban, N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312**, 1208-1211 (2006).

25. Roth, A. & Breaker, R.R. The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem.* **78**, 305-334 (2009).

26. Reiter, N.J., Chan, C.W. & Mondragon, A. Emerging structural themes in large RNA molecules. *Curr. Opin. Struct. Biol.* **21**, 319-326 (2011).

27. Deng, T., Sharps, J.L. & Brownlee, G.G. Role of the influenza virus heterotrimeric RNA polymerase complex in the initiation of replication. *J Gen Virol* **87**, 3373-3377 (2006).

28. Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J.M. & Strobel, S.A. Crystal structure of a self-splicing group I intron with both exons. *Nature* **430**, 45-50 (2004).

29. Leontis, N.B. & Westhof, E. The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. *Rna* **4**, 1134-1153 (1998).

30. Gutell, R.R., Larsen, N. & Woese, C.R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological reviews* **58**, 10-26 (1994).

31. Dunkle, J.A. et al. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332**, 981-984 (2011).

32. Dunkle, J.A., Xiong, L., Mankin, A.S. & Cate, J.H. Structures of the Escherichia coli ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. *Proc Natl Acad Sci U S A* **107**, 17152-17157 (2010).

33. Hainzl, T., Huang, S. & Sauer-Eriksson, A.E. Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. *Nature* **417**, 767-771 (2002).

34. Keenan, R.J., Freymann, D.M., Stroud, R.M. & Walter, P. The signal recognition particle. *Annu Rev Biochem* **70**, 755-775 (2001).

35. Larsen, N., Samuelsson, T. & Zwieb, C. The Signal Recognition Particle Database (SRPDB). *Nucleic Acids Res* **26**, 177-178 (1998).

36. Smith, K.D., Lipchock, S.V., Livinston, A.L., Shanahan, C.A. & Strobel, S.A. Structural and Biochemical Determinants of Ligand Binding by the c-di-GMP Riboswitch. *Biochemistry* **49**, 7351-7359 (2010).

37. Ling, J.Q., Hou, A. & Hoffman, A.R. Long-range DNA interactions are specifically altered by locked nucleic acid-targeting of a CTCF binding site. *Biochim Biophys Acta* **1809**, 24-33 (2011).

38. Gastaminza, P. et al. Antiviral stilbene 1,2-diamines prevent initiation of hepatitis C virus RNA replication at the outset of infection. *J Virol* **85**, 5513-5523 (2011).

39. Reblova, K. et al. An RNA molecular switch: Intrinsic flexibility of 23S rRNA Helices 40 and 68 5'-UAA/5'-GAN internal loops studied by molecular dynamics methods. *J Chem Theory Comput* **2010**, 910-929 (2010).

40. Sharp, P.A. The discovery of split genes and RNA splicing. *Trends in biochemical sciences* **30**, 279-281 (2005).

41. Dong, H., Ding, L., Yan, F., Ji, H. & Ju, H. The use of polyethylenimine-grafted graphene nanoribbon for cellular delivery of locked nucleic acid modified

molecular beacon for recognition of microRNA. *Biomaterials* **32**, 3875-3882 (2011).

42. Rice, G.M., Steen, K.A. & Siegfried, N.A. Unpublished. (2013).

Chapter 5: Testing Alternative 16S rRNA state

5.1 Introduction

5.1.1 RNA structure can be divided into three different levels

RNA structure can be discussed on three different structural levels. The primary structure is the nucleotide sequence of the RNA. The secondary structure is defined by the base pairing patterns and provides a blue print for the RNA structure. The tertiary structure is the most complex: Tertiary interactions define the three-dimensional structure of an RNA. Although tertiary interactions can involve base pairing, these interactions are not usually seen in the secondary structure.

5.1.2 Using X-ray crystallography to determine RNA structure

One of the most common techniques for determining three-dimensional structures is X-ray crystallography. X-ray crystallography identifies structural features by measuring the diffraction pattern of electrons. This diffraction pattern can then be mathematically transformed into an electron density map. Finally, atoms are modeled within the electron density. X-ray crystallography can be a powerful technique, but it is not compatible with most RNAs. For instance, molecules must be crystallized for analysis by X-ray crystallography. Only rigid, well-ordered RNAs form crystals of sufficient quality for analysis. Since the accuracy of X-ray crystallography depends

heavily on the quality of the crystals, the wide range of crystal types formed by RNAs poses potential problems[1, 2]. To stabilize the structure of flexible RNAs, high concentrations of metals, ligands, and proteins can be added to the RNA. This limits the number of conformations the RNA can adopt but also represent conditions that vary greatly from biology. Additionally, most crystallographic studies of large RNAs rely on assumptions regarding the RNA secondary structure to trace the general topology of the RNA[3]. This makes it easier to de-convolute electron density and identify a structure. If the assumed secondary structure determined based on co-variation analysis or another method is incorrect, it can significantly bias the structural prediction.

5.1.3 Using SHAPE directed prediction to determine the structure of the 16S rRNA

Previously, Deigan *et al.* attempted to recapitulate the secondary structure from the previously crystallized 16S rRNA[4] (conventional structure) using SHAPE technology (for selective 2' hydroxyl acylation analyzed by primer extension)[5]. In this method, the RNA is treated with a small molecule SHAPE reagent that preferentially reacts with single-stranded nucleotides, and the reactivities of each nucleotide are determined. By inputting SHAPE data into a folding algorithm, the number of possible structures is decreased and the accuracy of predictions is increased (see Chapter 2). The SHAPE-directed structure (alternative structure) for the 16S rRNA is shown in Figure 5.1. Most of the structure corresponds well with the conventional structure; however, there are some regions, including positions 140-220, 1064-1210, 946-1235, and 920-1410 that differ significantly. In these regions, neither the predicted secondary structure nor the SHAPE data agree with the conventional model. Conversely, the alternative structure

model matches well with the experimental data. These data suggest that these regions are in a different conformation than the conventional model.

To test the biological significance of these refolded regions and the validity of the SHAPE data, Deigan *et al.* performed two additional experiments[5]. The first tested the structure of the region of nucleotides 140-220. By binding DNA oligonucleotides to potential helices and looking for changes in SHAPE reactivity, Deigan *et al.* demonstrated that the SHAPE-directed structure occurred *in vitro*[5]. The second experiment tested the region between 920-1410. This region is near the tRNA binding site that is critical to the translation mechanics of the ribosome. Without the binding of tRNA, the amino acids cannot be integrated into the protein sequence. Since the alternative structure was determined without proteins or ligands bound, it is believed that differences between the alternative and the conventional structures may be due to the lack of critical ligands. To test this, the 16S rRNA was folded in the presence of tRNA, and the SHAPE experiment was repeated. When the 16S rRNA was bound to tRNA, the SHAPE reactivities were more similar to the conventional structure than reactivities in the absence of tRNA. This suggested that the alternative structure in this region might correspond to a lowest energy state of the RNA that forms before proteins and ligands bind.

The formation of different structures with and without tRNA bound has significant implications on the possible mechanism of the 16S rRNA. If the alternative structure identified by SHAPE-directed modeling is biologically relevant, it must fit within the general topology of the conventional model. Since the proposed structure is quite different from the conventional structure, it may not be compatible with the overall

structure and mechanism of the ribosome. To determine whether or not the proposed secondary structure of the 16S rRNA could be topologically compatible, I utilized Discrete Molecular Dynamics (DMD)[6] to create a tertiary structure model of the 16S rRNA. I then compared this model to the published electron density to determine whether the SHAPE-based model was compatible with the X-ray crystallography data[4].

5.2 Results

5.2.1 Alternative SHAPE directed prediction is different than conventional structure

Previous work by Deigan *et al.* suggested that the SHAPE-directed structure model of the 16S rRNA had significant structural changes compared to the conventional model[5]. To test whether or not these structural differences were compatible with the general topology of the 16S rRNA, I utilized DMD[6] to model the structures predicted based on SHAPE data into the electron density of the 16S rRNA[4].

5.2.2 Using modeling techniques to identify topology of alternative SHAPE directed structure

I first analyzed two regions with limited differences: nucleotides 1064-1210 and 946-1235. The secondary structures and conventional and alternative tertiary structure are shown in Figures 5.1 and 5.2. Only limited refolding and remodeling was necessary to fit the alternative models for these regions into the electron density.

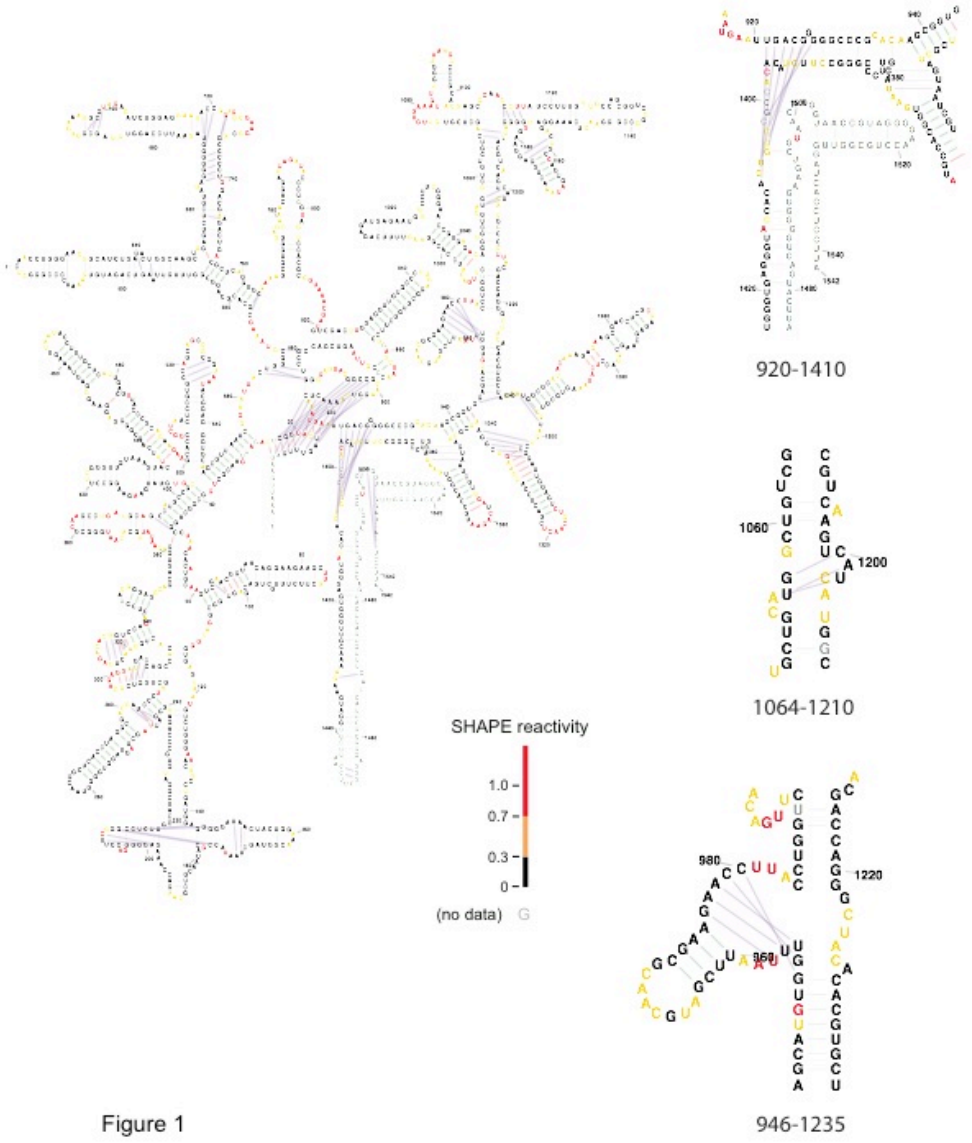

Figure 5.1: SHAPE-directed secondary structure model of the 16S rRNA compared to the conventional model.

Base pairs that are present in both models are shown as green lines connecting residues, those base pairs proposed based on the SHAPE-direct model are shown in purple, and base pairs present in the conventional model but not in the SHAPE-directed model are indicated with red. The regions highlighted in this chapter are shown on the right. The first is between nucleotides 920-1410, the second between 1064-1210, and the third between 946-1235.

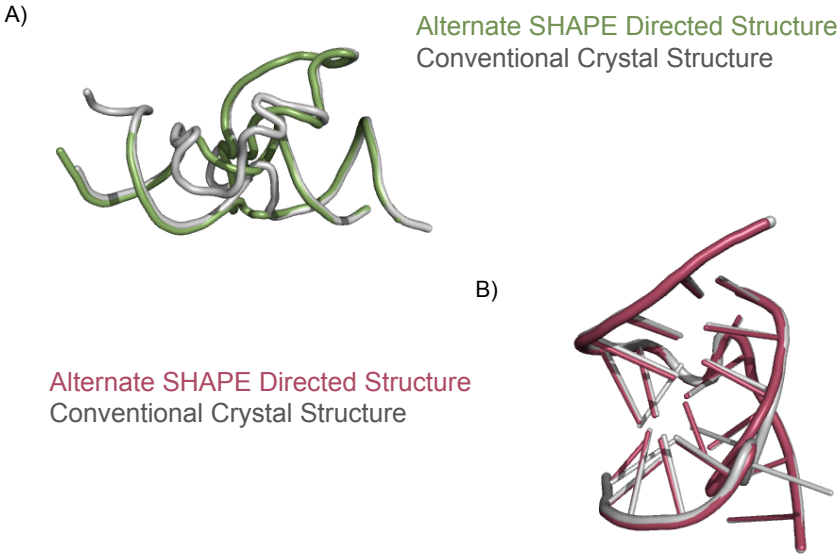

Figure 5.2: SHAPE-directed structural models of small refolded regions

Regions between nucleotides 1064-1210 (red) and 946-1235 (green). The conventional structure is shown in grey.

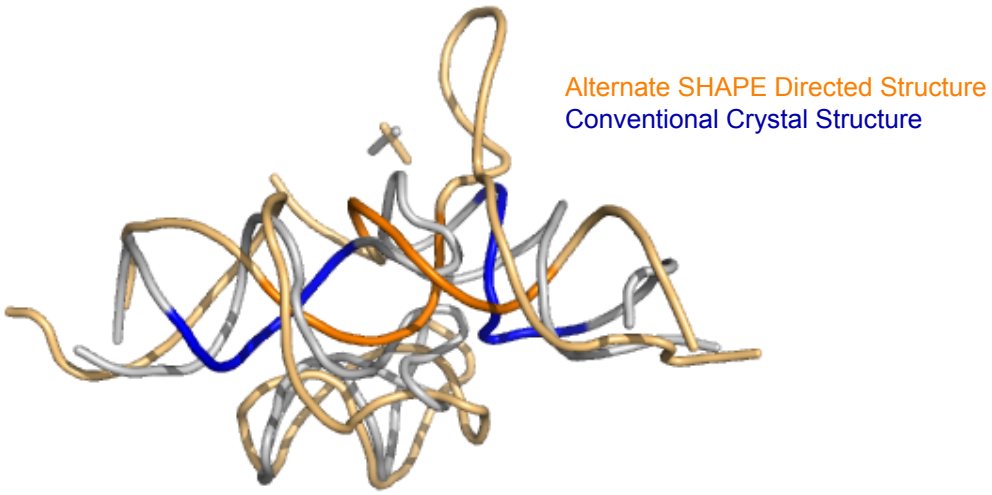

Figure 5.3: The structure for the region between nucleotides 920-1410 predicted by DMD based on SHAPE data.

The conventional structure is shown in grey. The helices that differ are highlighted in blue (conventional) and orange (SHAPE-based model).

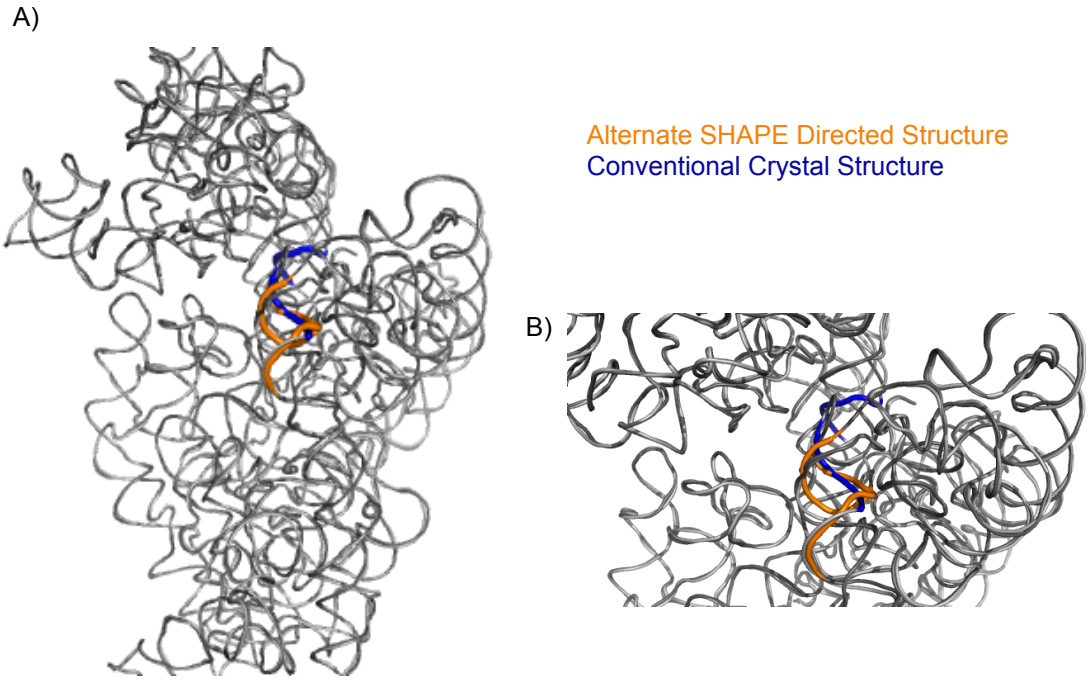

Figure 5.4: SHAPE-directed structure refined using PHENIX for the region between nucleotides 920-1410.

The helices that differ are highlighted in blue (conventional) and orange (SHAPE-based model). The remainder of the structure has small difference between the conventional(light grey) and alternative (dark grey).

The region between 920-1410 differs significant between the two models. I first modeled the secondary structure predicted based on SHAPE data using DMD. This structure and conventional structure are shown in Figure 5.3. The general topologies of the two structures are nearly identical despite differences in base-pairing patterns. The DMD model based on SHAPE data was inserted into the 16S rRNA electron density, further modified in Coot[7], and finally refined using PHENIX[8, 9]. The final alternative model of the 16S rRNA based on SHAPE data is compared to the crystallographic model shown in Figure 5.4. This alternative model fits the same topology of the conventional structure and fits well with the electron density data determined from X-ray crystallography.

5.3 Discussion

SHAPE-directed studies suggested that the 16S rRNA secondary structure may differ from that proposed based on conventional co-variation analysis that was used to fit the electron density in the reported X-ray crystallography structure[4, 10, 11] (Figure 5.1). Although most of the differences between the models are small and localized, the region from nucleotide 920 through nucleotide 1410 represents a significant change in folding. Despite the changes in base pairing interactions proposed based on SHAPE data, this work showed that the SHAPE-directed structure could form a compact tertiary structure that is consistent with topology of the conventional structure and the electron density data (Figure 5.4). There are two possible explanations for the base pairing differences in the SHAPE-based model and the model proposed by Gutell *et. al* and crystallized by Dunkle *et al*[4, 10, 11].

First, the fitting algorithm used to create the model based on X-ray crystallographic data is biased by the input secondary structure. The conventional structure previously determined from co-variation was used to model the electron density[10, 11]. It is not surprising that the crystal structure closely resembles the co-variation model. When the secondary structure based on SHAPE-directed modeling used as input to PHENIX[8, 9] and the R-value was recalculated, the value was nearly analogous to that obtained with the conventional structure, suggesting that the SHAPE-directed secondary structure fits the X-ray diffraction data as well as the conventional secondary structure does.

Second, the SHAPE-directed structure model was based on data collected in the absence of ligands or proteins. In contrast, the 16S rRNA was crystallized in the presence of tRNA, proteins S2-S21, and mRNA[4]. We hypothesize that these multiple bound proteins and ligands caused the 16S rRNA to move away from its lowest energy state; whereas, the alternative SHAPE directed structure was probed alone in solution and represents the lowest energy state of the RNA. The tRNA binding study by Deigan *et al*, which study showed that when tRNA was bound, the SHAPE reactivities were more consistent with the conventional structure, further supports this hypothesis[5].

5.3.1 Conclusion

In this work, we utilized the information determined from SHAPE probing of the 16S rRNA to predict the tertiary structure of the molecule. This worked showed that the SHAPE-directed model is compatible with the 16S rRNA topology and electron density[4] and that the ribosome might be in a different conformation when proteins are not bound. It also provided us with a unique perspective as to the accuracy of SHAPE-directed

predictions and their ability to provide alternative models for use in interpretation of X-ray crystallographic data. Lastly, this method demonstrated a unique way to utilize SHAPE data to probe both secondary and tertiary structures. In the future these methods can be used to identify other tertiary structures from SHAPE data.

5.4 Experimental

5.4.1 Performing SHAPE on 16S rRNA

The 5' domains of the *E. coli* 16S rRNA was equilibrated in buffer [50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 5 mM $MgCl_2$] at 37 °C for 30 minutes and treated with 1M7[12]. Frequencies of 2'-hydroxyl modification were identified by primer extension, resolved using capillary electrophoresis, and quantified using custom software[13]. After determining the inter-quartile range of the data, nucleotides with reactivities greater than 1.5 times interquartile range were taken to be outliers[5]. SHAPE reactivities were then divided by the mean of the 10% most reactive non-outlier data, which ultimately placed reactivities on a scale from 0 (no reactivity) to ~1.5. All SHAPE datasets obtained in this work are available at the SNRNASM community structure probing database[14].

5.4.2 Folding using RNAstructure Fold:

Since pseudoknots could be ignored in this study, the 16S rRNA was folded using RNAstructure Fold[5]. The parameters m=2.6 and b=-0.8 were used for folding[5].

5.4.3 Discrete Molecular Dynamics calculations:

The DMD algorithm models each nucleotide as three separate pseudo atoms: a sugar, a base, and a phosphate. Pair-wise interactions, including base pairing, base stacking, packing interactions, and electrostatic repulsion, were approximated using square-well potentials[6]. For this model, base-pairing information determined from the RNAstructure prediction was loaded into the program.

The simulations began at a high temperature with the RNA strand in an extended linear conformation. In the first step, the RNA was subjected to a folding phase designed to allow base pairs and local helical structure to form. Then, the RNA was cooled through automated steps as described in Lavender *et al.*[15]. To select a representative structure, potential structures from the final step were subjected to hierarchical clustering as described in Gherghe *et al.*[16]. Structures were binned by RMSD value into five clusters. The centroid of the cluster with the highest population was taken to be the representative structure. Refinements were performed on a Linux server (2.67 GHz with 48GB memory).

5.4.4 Modeling:

The initial model determined from DMD was corrected and improved in several rounds using automated restrained refinement with the program PHENIX[8, 9] and interactive modeling with Coot[7]. Source electron density files provided by the Cate Lab at the University of Berkeley[4]. The DMD model was initially read into Coot and the phosphate backbone was fit to the electron density map. Helices were built from the electron density. The final model was analyzed using the program MolProbity[17-19].

5.5 References

1. Lu, J., Li, N.S., Sengupta, R.N. & Piccirilli, J.A. Synthesis and biochemical application of 2'-O-methyl-3'-thioguanosine as a probe to explore group I intron catalysis. *Bioorg Med Chem* **16**, 5754-5760 (2008).

2. Mooers, B.H. Crystallographic studies of DNA and RNA. *Methods* **47**, 168-176 (2009).

3. Robertson, M.P., Chi, Y.I. & Scott, W.G. Solving novel RNA structures using only secondary structural fragments. *Methods* **52**, 168-172 (2010).

4. Dunkle, J.A. et al. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332**, 981-984 (2011).

5. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).

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

7. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).

8. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).

9. Adams, P.D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58**, 1948-1954 (2002).

10. Gutell, R.R., Larsen, N. & Woese, C.R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological reviews* **58**, 10-26 (1994).

11. Wu, J.C., Gardner, D.P., Ozer, S., Gutell, R.R. & Ren, P. Correlation of RNA secondary structure statistics with thermodynamic stability and applications to folding. *J Mol Biol* **391**, 769-783 (2009).

12. Mortimer, S.A. & Weeks, K.M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc* **129**, 4144-4145 (2007).

13. Karabiber, F., McGinnis, J.L., Favorov, O.V. & Weeks, K.M. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *Rna* **19**, 63-73 (2013).

14. Rocca-Serra, P. et al. Sharing and archiving nucleic acid structure mapping data. *Rna* **17**, 1204-1212 (2011).

15. Lavender, C.A., Ding, F., Dokholyan, N.V. & Weeks, K.M. Robust and generic RNA modeling using inferred constraints: a structure for the hepatitis C virus IRES pseudoknot domain. *Biochemistry* **49**, 4931-4933 (2010).

16. Gherghe, C.M., Leonard, C.W., Ding, F., Dokholyan, N.V. & Weeks, K.M. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J Am Chem Soc* **131**, 2541-2546 (2009).

17. Chen, V.B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).

18. Davis, I.W., Murray, L.W., Richardson, J.S. & Richardson, D.C. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res* **32**, W615-619 (2004).

19. Davis, I.W. et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, W375-383 (2007).

Chapter 6: Principles for understanding the accuracy of SHAPE-directed RNA structure modeling.

6.1 Introduction

6.1.1 RNA modeling may provide a useful alternative method for experimental techniques

The universe of biologically important RNAs with true three-dimensional tertiary folds, mediated by long-range and higher-order interactions, is likely to be very large. However, only a small fraction of these structures have been characterized at high-resolution. Moreover, there exist many functionally important RNA states, including folding intermediates and elements containing flexible motifs, whose structures cannot be established by direct high-resolution structure determination approaches. Structure-function relationships for these RNAs can, in principle, be addressed by accurate three-dimensional RNA structure modeling.

The field of RNA modeling is developing rapidly and many new ideas have been introduced for obtaining useful structures. Strategies for three-dimensional RNA structure prediction and modeling differ in whether they use all-atom or simplified representations of RNA structure, allow or require expert user intervention, facilitate incorporation of experimental information, or are designed for small versus large RNA motifs (reviewed in [1, 2]). Ultimately, the goal of all modeling approaches is the same: to generate an accurate structural model that is useful for designing, testing, confirming, or rejecting chemical and biological hypotheses.

RNA molecules are built up from just four nucleotide building blocks and form a single predominant secondary structure, the A-form RNA duplex. Thus, RNA structure prediction

might be easier than for proteins[3]. Even with these simplifying features, a given RNA can fold into a very large number of potential structures. An RNA of *N* nucleotides can form roughly 1.8^N base paired secondary structures[4] and a large number of tertiary folds.

6.1.2 Identifying methods of determining the accuracy of RNA tertiary models

The best way of summarizing the quality of an RNA structure model will vary depending on the prediction goals and methods. The quality of a tertiary structure model at the level of its overall fold can be summarized in a simple way as the root mean square difference (RMSD) between predicted and accepted RNA structures over a representative sets of atoms, typically a ribose atom or the phosphate position. A strength of using the RMSD to characterize structure prediction is that this metric can be applied to both simplified and all-atom models. Other metrics are necessary to characterize the accuracy of local interactions. For example, local base pairing and stacking interactions are sensitive to the all-atom RMSD, the global distance test (GDT, widely used to assess template-based models of protein structure) [5, 6], or the recently introduced interaction network fidelity (INF) which applies specifically to RNA [7]. The decision to focus on the global fold versus local interactions depends on the specific modeling objective. For longer RNAs with long-range tertiary interactions, it currently remains a major challenge to predict the overall architecture correctly; whereas, predictions for small helical RNAs or of individual motifs within large RNAs can sometimes correctly identify many individual hydrogen bonding and base stacking interactions.

In this work, we sought to develop an approach for characterizing algorithms designed to predict the overall architecture of relatively large RNA (50-200 nts), characterized by extensive long-range interactions that involve more than individual helices (for example, Figure 6.1A). We focus on metrics for assessing the global fold of an RNA at roughly "nucleotide resolution".

This is also the level of RNA structural information that is obtained from most biochemical experiments when applied to large RNAs. This class of experiments includes chemical probing, through-space cleavage and crosslinking, and solution hydrodynamic measurements. To this end, we address the magnitude of RMSD that constitutes a successful prediction, as opposed to models that are not significantly different from those expected by chance. Throughout this work, we will emphasize RMSDs calculated over all phosphate positions, although our conclusions are likely to apply to correlations calculated at any backbone position.

6.1.3 Identifying variables that will effect the accuracy of prediction

Success and failure for tertiary structure prediction are obvious at the extremes. For example, for an RNA of moderate size like the SAM-I riboswitch (94 nts) [8], a model with 4.5 Å RMSD relative to the crystallographically determined structure [9] clearly corresponds to a good prediction; whereas, a prediction at 18 Å RMSD is unlikely to be helpful in generating strong, testable biological hypotheses (Figs. 1A,C). At 13.2 Å RMSD, a model for this RNA clearly resembles the experimentally determined structure (Figure 6.1B). However, given the intrinsic rigidity of RNA helices and the limited number of nucleotide building blocks, it is not clear whether a model that differs from the accepted structure by 13.2 Å RMSD constitutes a successful prediction, especially if the secondary structure is used as a constraint during modeling.

RNA chain length is an important variable in establishing the RMSD value that describes a non-random prediction. The range of RMSD values that correspond to similar RNA structures increases with chain length. For example, two RNAs with a 4.5 Å RMSD are similar if their lengths are 94 nts (Figure 6.1A),

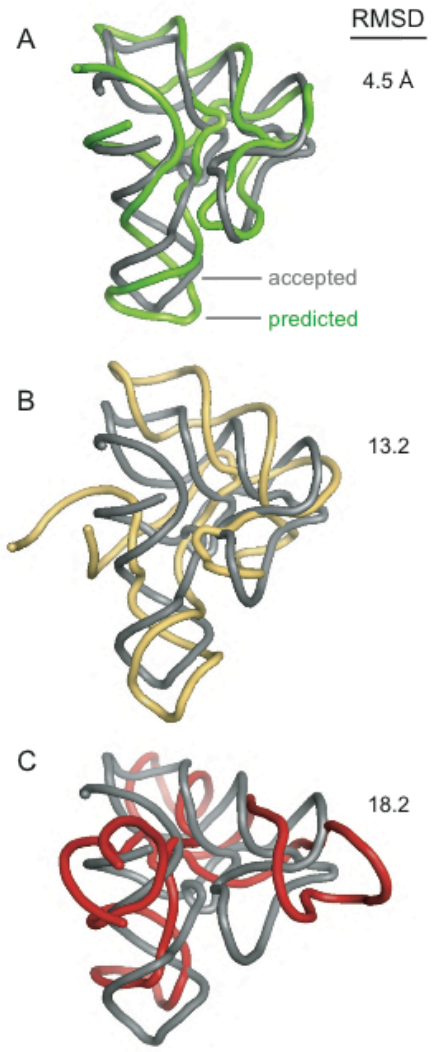

Figure 6.1: Comparison of an accepted RNA structure with modeled tertiary structures as a function of RMSD similarity.

The experimentally determined [9] and simulated structures of the SAM riboswitch (94 nts, 2gis) are shown as gray and colored backbones, respectively.

but are dissimilar if they comprise short base paired duplexes. This feature is common to both protein [10, 11] and RNA structure prediction, but may be more pronounced with RNA for two reasons. First, structured RNAs tend to be more elongated and less globular compared to proteins of similar mass. Second, stacked helices comprise the major structural building block for RNA, are relatively rigid, and can span large linear dimensions. If a helix is modeled to be in roughly the right place but is angled relative to the correct orientation, this error can propagate to produce large RMSD values with modest degrees of angular deflection.

A second criterion distinctive to RNA structure prediction is that the pattern of base pairing that comprises an RNA secondary structure is often known with perfect or near-perfect accuracy prior to three-dimensional modeling. Accurate RNA secondary structures can be obtained from comparative sequence analysis[12-14] and experimentally-constrained prediction [15]. Most RNA helices, including those that incorporate mismatched and non-canonical base pairs, will show good (< 2 Å RMSD) alignments if the structure is simply assumed to be A-form. For large RNAs, enforcement of native-like base pairing dramatically reduces the allowed conformational space. RMSD values for predicted structures should therefore be significantly smaller if information regarding base pair constraints is included in the modeling algorithm.

In this work, I develop a framework for assessing the confidence that a predicted RNA tertiary structure is significantly different from a chance prediction. I generate a large number of decoy structures using replica exchange DMD and then calculate the magnitude of RMSD that indicates any two structures are more similar than two randomly generated, but still RNA-like, chains. I also establish an empirical power law relationship for mean RMSD as a function of chain length that makes it possible to define analytical expression for the confidence, and non-randomness, of RNA structure prediction.

RNA	PDB ID	N (nts)	imposed base pairing: – <RMSD> (Å)	– σ	– RMSD $p = 0.01$	+ <RMSD> (Å)	+ σ	+ RMSD $p = 0.01$
Sarcin/ricin domain	1q9a	27	8.3	1.7	7.8	4.2	1.7	0.1
Viral RNA pseudoknot	1l2x	28	12.4	1.7	8.2	2.7	0.8	0.1
Vitamin B12 aptamer	1ddy	35	16.0	1.9	10.6	7.9	1.9	1.9
4.5S RNA fragment	1duh	45	19.8	1.7	13.6	8.5	1.4	4.3
SARS virus pseudoknot	1xjr	47	20.5	1.7	14.1	7.4	1.8	4.7
Guanine riboswitch	1u8d	68	24.0	1.9	19.2	14.1	1.6	8.8
$tRNA^{Asp}$	2tra	75	24.7	1.7	20.7	18.7	1.7	10.0
Thi-box riboswitch	3d2g	83	27.0	1.9	22.3	11.7	1.9	11.2
SAM riboswitch	2gis	94	29.4	2.0	24.3	17.7	2.0	12.9
SRP RNA	1z43	101	27.9	1.8	25.6	16.5	1.7	13.8
glmS ribozyme	2gcs	125	35.4	2.0	29.4	24.0	2.0	16.9
RNase P specificity domain	1nbs	155	38.6	2.1	33.6	24.5	1.8	20.3
Tetrahymena P546 domain	1gid	158	36.5	1.8	34.1	25.3	1.8	20.7
Lysine riboswitch	3dou	161	39.5	1.9	34.5	23.9	1.8	21.0

Table 6.1: RNA targets with decoy structures generated by DMD

6.2 Results

6.2.1 Selection of Target Structures.

RNA structures, ranging in size from 27 to 161 nts, were obtained from the RCSB structure database (Table 6.1). RNA structures were required to (*i*) be solved at a resolution of 3.3 Å or better, (*ii*) have non-trivial higher-order tertiary interactions, defined as having close helix packing, long-range intrastrand interactions, or a pseudoknot, (*iii*) contain a single complete or nearly complete chain, and (*iv*) form a stable tertiary structure in the absence of protein binding. We excluded RNAs that form simple A-form helices or stem-loops or that form Y-shaped structures without significant long-range tertiary interactions. For RNAs with multiple structures, the example with the best resolution or that was most complete was selected. The RNA structures were also chosen to be distributed evenly over the 27-161 nt length range, given the examples available in the current RCSB database [16].

6.2.2 Generation of Decoy Structures by DMD.

Ideally, the quality of an RNA tertiary structure prediction would be determined by comparing the agreement between a predicted versus an experimentally determined model. This similarity would then be compared to the differences observed between members of a diverse group of experimentally determined decoy structures of similar size. Unfortunately, even with the recent increase in high-resolution structures, there are still too few known RNA structures to serve as a statistically valid set of decoys in any given size range.

I therefore used replica exchange DMD simulations [17] to generate decoy structures for representative RNAs. RNA decoys were generated by DMD using a coarse-grained model in which each nucleotide is represented as three pseudo-atoms corresponding to the phosphate,

sugar, and base moieties [17]. Interactions between pseudo-atoms include bonded, non-bonded, and loop entropy terms. This coarse-grained RNA model yields topologically reasonable RNA-like folds for a large number of small RNAs [17] and for tRNA when constrained by pair-wise experimental information [18]. Replica exchange DMD makes it possible to efficiently overcome energy barriers in a rugged energy landscape and to explore conformational space broadly while simultaneously maintaining conformational sampling in a regime that corresponds to a physically relevant free energy surface [19, 20].

A *priori* knowledge of the secondary structure dramatically increases the correlation (and therefore reduces the RMSD) between simulated and experimentally determined structures. I therefore also generated decoy structures for each target RNA in which the DMD pseudo-atoms corresponding to the bases were constrained to pair. In all cases, I selected for compact decoy structures by requiring that the radius of gyration be within 1.2-fold of the native structure.

6.2.3 Analysis of RNA Decoy Structures.

To generate an ensemble of statistically significant and structurally reasonable decoy structures, the replica exchange DMD simulations must reach equilibrium in conformational sampling. I therefore evaluated whether the DMD ensembles generated from different starting states converged. I initiated simulations starting from two very different starting states, the experimentally determined native structure and a linear, extended, structure generated *in silico* for seven of the target RNAs (1q9a, 1l2x, 1xjr, 1u8d, 2gis, 1nbs, 1gid; Table 1). Both the pair-wise RMSD distributions (Figure 6.2) and DMD energies (not shown) were nearly identical for simulations initiated from either the native or fully extended states. This similarity in the final distribution of structures holds independent of whether the native pattern of base pairing is imposed during the simulation (Figure 6.2).

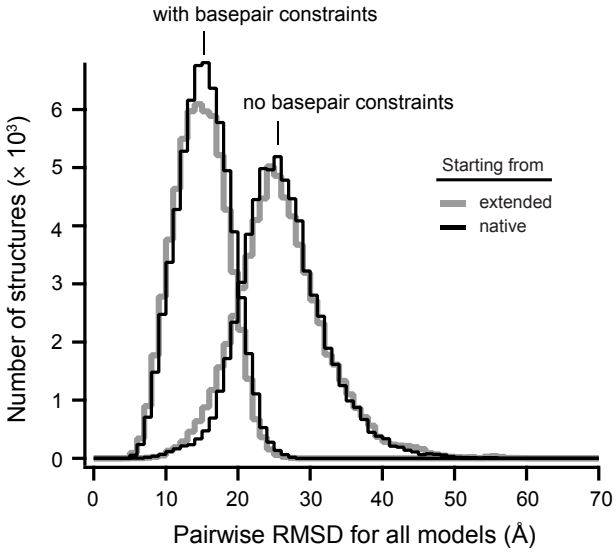

Figure 6.2: Replica exchange DMD simulations as a function of starting state and of enforcing native base pairing.

Simulations were initiated either from the crystallographic structure or from a linear, extended state for the purine riboswitch (68 nts, 1u8d) [21].

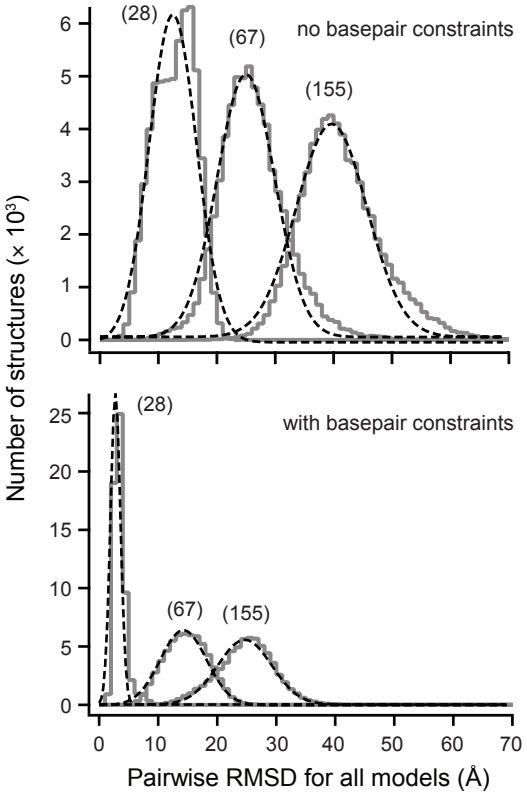

Figure 6.3: Distributions of decoy structures.

RNA decoy structures were stimulated using replica exchange DMD starting from fully extended linear structures either without or with constraints that enforce the native pattern of base pairing (solid gray lines). Distributions show good Gaussian-like behavior (dashed lines). RNAs shown are a viral RNA pseudoknot (28 nts), the purine riboswitch (68 nts), and the specificity domain of RNase P (155 nts) [21-24].

Thus, replica exchange DMD yields fully equilibrated sets of RNA decoy structures for RNAs as large as 160 nts. We then used replica exchange DMD to generate decoy structures for our complete set of RNAs (Table 6.1) and calculated RMSD values for all pair-wise combinations of decoy structures. Representative RMSD distributions for a viral RNA pseudoknot (28 nts), the purine riboswitch (68 nts), and the specificity domain of RNase P (155 nts) are shown in Figure 6.3. These profiles have three critical features. First, the pair-wise RMSD distributions are Gaussian-like (compare solid and dashed lines, Figure 6.3). A Gaussian-like distribution in pair-wise RMSD distribution is consistent with the Central Limit Theorem that holds that the sum of a large number of random variables (structures) should be normally distributed. Gaussian-like behavior also means that each distribution can be characterized by its mean RMSD value and a standard deviation.

Second, mean RMSD values increase as a function of chain length (Figure 6.3, Table 6.1). Hence, no single RMSD value represents a non-random prediction. An RNA modeling algorithm must therefore produce structures with comparatively smaller RMSD values for short RNAs, if these structures are to be better than those expected by chance.

Third, imposing the native pattern of base pairing has a large effect on the RMSD distributions. Constraining structures to have native base pairing biases the distribution to smaller RMSD values by 4-15 Å, depending on RNA length (Figure 6.3, Table 6.1).

6.2.4 A Power Law Relationship for the Radius of Gyration and Chain Length in RNA.

Given the mean and standard distribution for each RMSD profile, I sought to determine a proper mathematical relationship between the mean, the chain length (N) and the RMSD (derived below). The mean RMSD for protein structure prediction is approximately proportional to the radius of gyration.

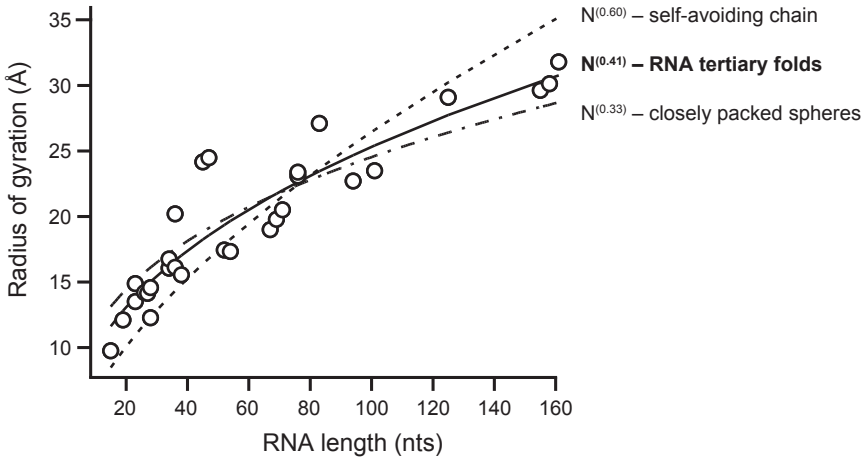

Figure 6.4: Dependence of radius of gyration on chain length for compact RNAs with higher-order tertiary structure interactions.

Fits to the 0.33 and 0.60 exponents (but not to the 0.41 exponent) show systematic deviations from the points.

This relationship reflects that the distances between corresponding atoms in two structures scale with the overall dimensions of the macromolecule [11]. I also expect that the mean RMSDs will scale in a similar way with chain length and the radius of gyration for RNA. I calculated the radius of gyration, R_g, for all of the RNAs in our target set (Table 6.1) plus a set of additional RNAs to more fully populate the R_g versus N curve (Figure 6.4). The best fit gives:

$$R_g \sim 3.8\, N^{0.41} \qquad (1)$$

The key result is the exponent, 0.41, which lies between the values expected for a molecule composed of closely packed spheres (1/3) and for a self-avoiding chain (3/5) [25]. This exponent is different from a prior analysis that suggested R_g for RNA scales with an exponent of 0.33 [26]. The earlier work did not filter simple helices of 25 nts or less and included the 16*S* and 23*S* ribosomal RNAs, which achieve their structures only as ribonucleoprotein complexes. Excluding these two sets of RNAs yields an exponent consistent with this work.

Both Pearson's correlation coefficient and the non-parametric Wald-Wolfowitz test indicate that the 0.41 exponent better fits the R_g data than either of the other two limits (Figure 6.4). This result is intrinsically satisfying because it suggests that folded RNAs are more structured than random self-avoiding chains but do not fully maximize their packing density. This exponent is also slightly larger than the 0.33 value found for proteins [11], consistent with the less-globular structures of most RNAs relative to proteins of the same mass [27].

6.3 Discussion

We have used DMD to calculate statistically significant sets of decoy structures for a representative set of RNAs. These decoy structures correspond to compact, RNA-like, but largely incorrect, structures for each target RNA. Mean RMSD values increase with chain length, both when base pairing was allowed to vary or was constrained to correspond to that in the

accepted structure (Figure 6.5A). In both cases, these distributions are well fit by a power law relationship, $a\ N^{0.41} - b$, where the exponent 0.41 is derived from R_g and N (Figure 6.4 & 6.7). Since the mean RMSD value defined by the empirical relationship with respect to RNA length should be positive, the RNA length should be $N > N_c = (b/a)^{1/0.41}$. The critical length, N_c, is approximately 5.3 when no base-pair information is imposed during modeling and 16 Å when the base-pair constraints are enforced (a & b for a chance prediction are given in Figure 6.7). These values are sensible and correspond to the minimal lengths of RNA with significant secondary and tertiary structures. Mean RMSD values increase by roughly 5-fold as chain length increases from 27 to 160 nt.

In contrast, the standard deviation in RMSD for each distribution is approximately constant at 1.8 Å (Figure 6.5B). It is not clear what physical property of RNA explains the relative constantness of the standard deviation in RMSD; interestingly, a similar behavior appears to hold for protein structure [11].

These distributions (Figure 6.5) represent a measure of the agreement between any two structure predictions for an RNA of a given size as expected by chance. Although we generated these distributions based on a specific DMD model for the RNA decoy structures, we believe these relationships will be general because our DMD model captures the driving forces of RNA folding and is able to predict the native structures of many small RNAs from a large set of competitive decoys [17]. Moreover, the replica exchange simulation efficiently samples RNA conformational space, which is populated by many thermodynamically viable decoy structures with competitive base pairing and higher-order packing interactions.

Using the empirical relationships for RMSD distribution as the function of RNA length (Figure 6.5), it is possible to create a scoring function for the significance of an RNA tertiary

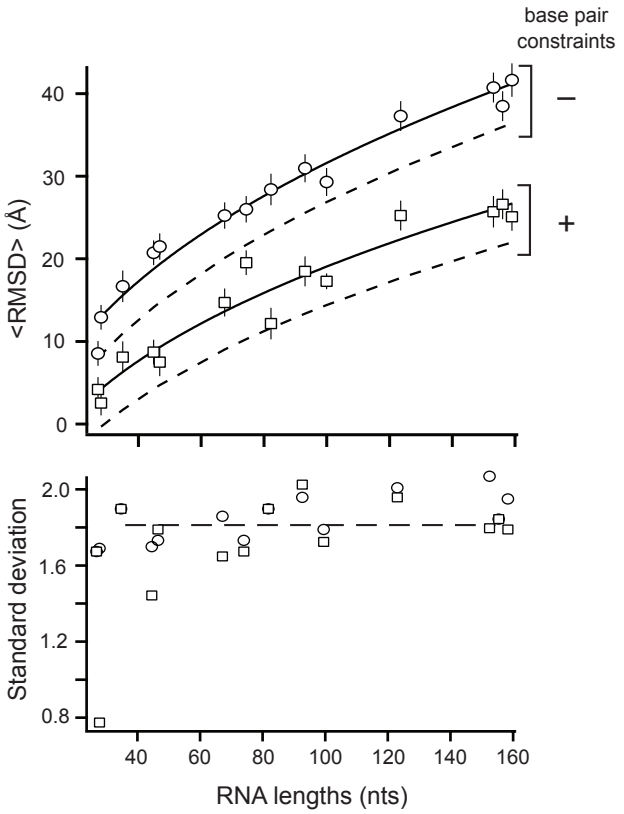

Figure 6.5: Mean pair-wise RMSD as a function of RNA chain length.

Decoy structures either constrained to form base pairs found in the experimentally determined native structure or allowed to form any energetically favorable set of base pairs are shown. Solid lines correspond to distributions expected for RNA-like, but chance, folds. Dashed lines indicate the RMSD cutoff corresponding to a prediction better than that expected by chance at the $p < 0.01$ level. Lines indicate fits to the power law relationship $\langle RMSD \rangle \approx a\, N^{0.41} - b$; a and b values are given in Figure 7. The mean and standard deviation for each distribution are shown with symbols and error bars.

structure prediction given the chain length (N) and the RMSD relative to the accepted structure (Figure 6.7). This assessment of RNA tertiary structure prediction can be summarized as a *p*-value. Smaller *p*-values correspond to more statistically significant predicted structures. The *p*-value calculation provides a broad measure of prediction quality for RNAs between 35 and 160 nts and can be used to evaluate predictions for both small and large RNAs and for algorithms that make use of prior information about base pairing versus those that predict all interactions *de novo*. The mean and standard deviation obtained for each distribution can also be used to calculate the RMSD between a known and predicted three-dimensional structure that corresponds to a predicted structure that differs from a random prediction at a chosen confidence level. We suggest that $p < 0.01$ represents a successful prediction (dashed lines, Figure 6.5A). Analytical expressions for the distributions corresponding to chance predictions and to successful predictions at the $p < 0.01$ level are given in Figure 6.7.

Our laboratories are developing accurate and efficient methods for modeling complex RNA structures [15, 17, 18, 28, 29]. Many other laboratories are also making innovative contributions to the RNA modeling field [2, 30-34]. We undertook the present study in order to create a framework for benchmarking any RNA modeling algorithm. We illustrate the usefulness of the *p*-value approach outlined here by considering two studies that have focused on refining the tertiary fold of tRNA.

For an RNA the size of yeast $tRNA^{Asp}$ (75 nts), a model should have an RMSD over all phosphate atoms of 10.0 Å or better to reach $p \leq 0.01$. For comparison, RMSD values between $tRNA^{Asp}$ and two unrelated RNAs of similar size, the HDV and Thi-box RNAs, are 23 and 27 Å which correspond to the near-maximal *p*-value of 0.99; whereas, the free $tRNA^{Asp}$ and its protein-bound form superimpose with an RMSD of 6.5 Å (*p*-value = .00001) (Figure 6.6).

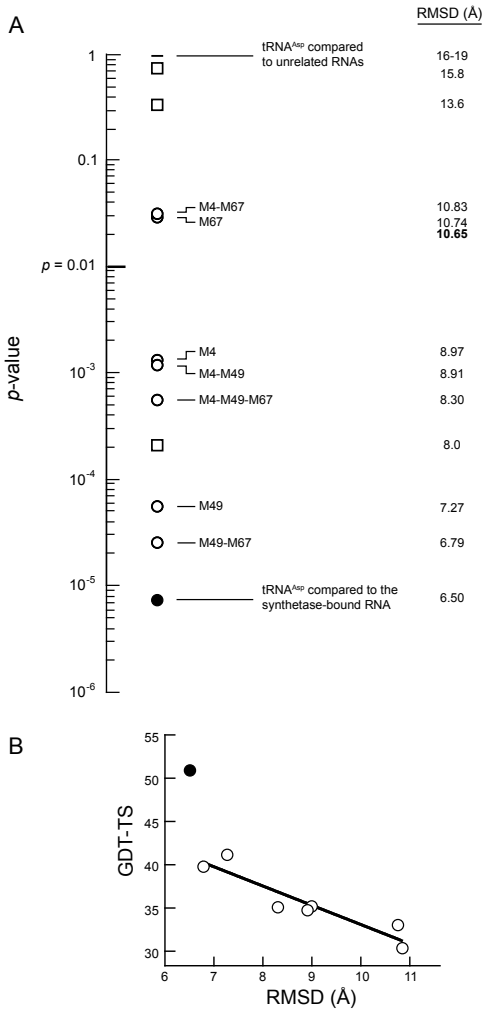

Figure 6.6: Use of *p*-values to benchmark RNA tertiary structure models.

(A) Spheres represent p-values for seven models (indicated with Mx) of tRNAAsp (2tra, 37) based on experimentally-derived tertiary structure information, refined by DMD 18. Squares indicate p-values for three refinements of tRNA using a one-bead model for RNA and filtering by hydroxyl radical and SAXS data 2. For comparison, p-values for two unrelated RNAs of similar size, the HDV ribozyme (1vby, 76 nts) [38] and the Thi-box riboswitch (3d2g, 77 nts) [39] plus tRNAAsp as it exists when bound by its synthetase (1asy) [40], are shown as horizontal bars. (B) Comparison of RMSD and GDT-TS values for the seven Mx tRNA models (open circles), plus the comparison between the 2tra and 1asy structures (filled circle).

Figure 6.7: Significance (*p*-value) analysis for RNA tertiary structure prediction.

Relationship between <RMSD> and N (from Figure 5):

$$\text{<RMSD>} = a\,N^{(0.41)} - b$$

imposed base pairing:		–		+	
		chance	$p < 0.01$	chance	$p < 0.01$
where	a =	6.4	6.4	5.1	5.1
	b =	12.7	16.9	15.8	19.8

Given N and the RMSD between predicted and accepted structures, m, the prediction significance (p-value) is:

$$p\text{-value} = \frac{1 + erf(Z/\sqrt{2})}{2}$$

$$\text{where} \quad Z = \frac{m - \text{<RMSD>}}{\sigma_m}$$

$$\text{and} \quad \sigma_m \approx 1.8 \text{ Å}$$

In one approach, native-like tertiary structures for yeast $tRNA^{Asp}$ were obtained given only the sequence and using a combination of SHAPE chemistry [35, 36] and pair-wise constraints generated using a sequence-directed cleavage agent. This biochemical information was then refined using DMD [18]. The cleavage agent was placed at nucleotide positions 4, 49 and 67 in $tRNA^{Asp}$ and structures were refined using the tertiary constraints provided by any one, two, or all three experiments for seven possible total refinements (summarized as spheres, Figure 6.6A). Of the seven refinements, five yielded models with *p*-values significantly lower than 0.01 (Figure 6.6A). These refinements correspond to *p*-values of 2.0×10^{-5} to 2.0×10^{-3} (calculated given the correct pattern of base pairing as established by SHAPE). Two structures refined to RMSDs of ~10.8 Å, corresponding to a *p*-value of 0.03, which represent fair predictions, but not equivalent to the $p < 0.01$ level. RNA

In a second approach, tRNA was modeled by representing each nucleotide as a single bead centered at the C3' atom, enforcing base pairing, and filtering structures based on hydroxyl radical cleavage and SAXS data yielded models for *E. coli* $tRNA^{Phe}$ (76 nts) with RMSDs of 8.0, 13.6 and 15.8 $Å^2$. Although these RMSD values were calculated at the C3' position, comparison with the framework developed here is appropriate because RNA backbone atom positions are highly correlated (see Methods). These RMSD values correspond to *p*-values of .00023, 0.36 and 0.80 (squares, Figure 6.6A). Overall, this analysis of two recent, and different, approaches for refining RNA structure models makes clear that experimentally-constrained modeling of complex RNA structures has substantial promise for refining structures to *p*-values ≤ 0.01, but that additional effort is required to reach this level consistently.

An alternative to the RMSD, the global distance test (GDT) is a good indicator of similarity between two structures. The GDT-TS (total score), as implemented in the LGA

program [5], has been widely used to rank protein models [6, 41] and, recently, to evaluate RNA structures [2, 7]. LGA uses multiple alignments and calculates the largest set of atoms that deviate by less than a user-defined cutoff. GDT scores span a uniform scale with zero equal to no similarity and 100 indicating near perfect agreement. It is not clear what GDT-TS score corresponds to a significant tertiary fold prediction for RNA. We find that RMSD and GDT-TS are highly correlated (r^2 = 0.86) for RNA models at medium resolution (open circles, Figure 6.6B). A GDT-TS value ≥35 indicates a strong prediction, with a *p*-value > 0.01 (as defined in Figure 6.7). However, the GDT-TS increases rapidly as structures become highly similar. This is exemplified in the comparison of free $tRNA^{Asp}$ with it synthetase-bound form. Of the 75 nucleotides that comprise these two structures, 70 positions have RMSDs less than 5. The remaining nucleotides have large variations, with RMSDs >10. This gives a GDT-TS of 51, whereas the overall RMSD is 6.5 (filled circle, Figure 6B). Thus, for very detailed analyses involving threading, homology modeling, or evaluating single site mutations, the GDT-TS is more discriminating. However, for evaluating RNA modeling at the level of the global fold, especially for RNAs with long-range tertiary interactions, the RMSD and GDT-TS are both good metrics for determining similarity.

Returning to our original example, a 4.5 Å RMSD for an RNA of 94 nts using an algorithm that enforces native base pairing (Figure 1A) corresponds to a highly significant prediction (*p*-value ≤ 10^{-6}). In contrast, a 18.2 Å RMSD (Figure 6.1C) is readily identified as a poor prediction by its *p*-value = 0.74. For an RNA of 94 nts, the 13.2 Å prediction falls at the *p* = 0.016 level. Inspection of the agreement between this structure and the accepted structure (Figure 6.1B) supports the view that this prediction lies near the lower limit at which the model might be useful for designing instructive biological hypotheses. We believe that *p*-value analysis

will prove broadly useful in ongoing efforts to benchmark and improve RNA tertiary structure prediction and modeling algorithms.

6.4 Experimental

6.4.1 Target RNAs and analysis of Power Law relationships for RNA.

RNA structures were obtained from the RCSB structure database [16]. For RNAs with multiple structures, the example with the best resolution or that was most complete was selected. If the U1A protein was present to facilitate crystallization [42], this protein component was removed. To establish a power law relationship between the radius of gyration and RNA length, we calculated the radius of gyration (R_g) for the structures in Table 1, plus the following (listed by PDB code): 1ato, 1nem, 2tob, 1q9a, 1l2x, 437d, 1eht, 1rnk, 1fmn, 1q8n, 1mme, 1xjr, 2qwy, 3e5c, 1kh6, 2goz, 1u8d**,** 1y26, 1eov, 1tra, 1vby, 3d2g, 2hoj, 2gis, 1z43, 2gcs, 1nbs, 1gid, 2qbz, 1u9s, 3djz, 1u6b, 1x8w, 3bwp, 2a64. The radii of gyration were fit to Eqn. 1. We used both Pearson's correlation coefficient, *r*, and the non-parametric Wald-Wolfowitz test to evaluate whether the best fit exponent of 0.41 is better than the limits for closely packed spheres (0.33) or a self-avoiding chain (0.60). *p*-values for the latter two values were 0.0096 and 0.0003 which indicate statistically significant deviations; in contrast, the *p*-value for the 0.41 exponent was 0.24, indicating no significant deviation from the proposed power law model. We also calculated the exponent for a complete dataset of all RNA structures in the RCSB database (as described by [26]). The exponent over all deposited structures is 0.33, exactly as reported previously; however, if short (< 25 nt) and ribosomal RNAs are excluded and only single chain RNAs are considered, the exponent is 0.46.

6.4.2 Generation of RNA decoys by Replica Exchange DMD.

We used replica exchange DMD [17, 43] to explore RNA conformational space and generate statistically valid ensembles of decoy structures. Each RNA nucleotide is represented as three pseudo-atoms representing the phosphate, sugar, and base moieties [17]. Bonded terms included bond angles and dihedrals; non-bonded terms included base pairing, stacking, hydrophobic, and phosphate-phosphate repulsion interactions; an explicit term was included for loop entropy. Replica DMD simulations were performed in parallel over temperatures ranging from low (T = 0.20) to high (T = 0.24); this temperature range covers the folding temperatures of the coarse-grained RNA model [17]. Replicas with neighboring temperature values were periodically [every 2000 time units (tu)] exchanged in a Metropolis manner. Temperatures were exchanged between two replicas, i and j, at temperatures T_i and T_j, and with energies E_i and E_j according to the exchange probability ρ, where $\rho = 1$ if $\Delta = (1/k_B T_i - 1/k_B T_j)(E_j - E_i) \leq 0$, and $\rho = exp(-\Delta)$, if $\Delta > 0$. Simulations were carried out for 800,000 tu, yielding 12,000 structures. Decoy generation for a 150 nt RNA requires approximately 20 hrs on a single core equivalent Xenon CPU (2.3 GHz). Individual structures were accepted for pair-wise analysis subject to the following: (*i*) simulations were allowed to equilibrate for 2000 frames to exclude structures that reflected residual memory of the starting state, (*ii*) frames were required to be different by 200 steps to exclude correlated consecutive structures, and (*iii*) structures were required to be compact and have a radius of gyration that was within 1.2-fold of the accepted structure.

6.4.3 Pair-wise RMSD and Gaussian Distribution calculations.

The RMSD was calculated as:

$$RMSD = \min\left\{\sqrt{\sum_{i=1}^{N}(\vec{r}_i^{\,1} - \mathbf{A}\vec{r}_i^{\,2})^2 / N}\right\}, \qquad (2)$$

where **A** is an arbitrary rotation matrix. The calculation was performed using the Kabsch algorithm [44] over all phosphate positions in each RNA. RMSD distributions were fit to a Gaussian curve,

$$y = A\ e^{\frac{-(x-x_0)^2}{2\sigma^2}} \tag{3}$$

where A is the amplitude, x_0 is the mean, and σ is the standard deviation.

6.4.4 Effect of calculating RMSD values over other RNA atoms.

I calculated RMSDs for free $tRNA^{Asp}$ (2tra) [37] relative to this tRNA as bound by the tRNA synthetase [40] (RNA molecule in 1asy). RMSD values as a function of atom are: phosphate, 6.80 Å; C3', 6.37 Å; C4', 6.66 Å; N1, 6.59 Å; N3, 6.68 Å; and over all atoms, 7.11 Å. The single atom RMSD values are essentially identical; the all-atom value is larger by 0.3-0.6 Å.

6.4.5 Calculation of Confidence Intervals.

The $p < 0.01$ line in Figure 6.5 was calculated from a standard Z-score relationship. For $p < 0.01$, the RMSD value is obtained as:

$$\mathrm{RMSD}_{p<0.01} = x_0 - 1.8\sigma \tag{4}$$

The RNA prediction significance, or p-value, is also calculated from the Z-score, given a predicted structure that differs from an accepted structure by an RMSD of m:

$$Z = \frac{m - <\mathrm{RMSD}>}{\sigma_m} \tag{5}$$

where <RMSD> is the expected RMSD obtained from the best-fit relationship in Figure 6.7 and is a function of chain length, N; σ_m is the standard deviation for decoy structures of length N (Figure 6.5). For predictions of RNAs with lengths ≥35 nts, this value is approximately constant at 1.8 Å. The statistical probability of obtaining a given RMSD value is estimated as the p-value:

$$p(Z) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{z} e^{\frac{-x^2}{2}} dx$$

$$= \frac{1}{\sqrt{\pi}} \left(\int_{-\infty}^{0} e^{\frac{-x^2}{2}} dx + \int_{0}^{z} e^{\frac{-x^2}{2}} dx \right)$$

$$= (1 + erf(Z/\sqrt{2}))/2 \qquad (6)$$

where *erf*(*x*) is the standard Gauss error function and *Z* is given by Eqn. 5. A simplified summary of this calculation is provided in Figure 6.7.We provide a spreadsheet for calculating the RNA tertiary structure prediction significance *p*-value, given *N* and the RMSD between the predicted and accepted structure. This calculation and source code are also available at the iFoldRNA server (http://iFoldRNA.dokhlab.org) [29].

6.5 References

1. Shapiro, B.A., Yingling, Y.G., Kasprzak, W. & Bindewald, E. Bridging the gap in RNA structure prediction. *Curr. Opin. Chem. Biol.* **17**, 157-165 (2007).

2. Jonikas, M.A. et al. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *Rna* **15**, 189-199 (2009).

3. Tinoco, I. & Bustamante, C. How RNA folds. *J. Mol. Biol.* **293**, 271-281 (1999).

4. Zuker, M. & Sankoff, D. RNA secondary structures and their prediction. *Bull. Math. Bio.* **46**, 591-621 (1984).

5. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucl. Acids Res.* **31**, 3370-3374 (2003).

6. Keedy, D.A. et al. The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins* **77 (Suppl 9)**, 29-49 (2009).

7. Parisien, M., Cruz, J.A., Westhof, E. & Major, F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *Rna* **15**, 1875-1885 (2009).

8. Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E. & Breaker, R.R. An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nature Struct. Biol.* **10**, 701–707 (2003).

9. Montange, R.K. & Batey, R.T. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441**, 1172-1175 (2006).

10. Cohen, F. & Sternberg, M.J.E. On the prediction of protein structure: the significance of the root-mean-square deviation. *J. Mol. Biol.* **138**, 321-333 (1980).

11. Reva, B.A., Finkelstein, A.V. & Skolnick, J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Folding Des.* **3**, 141-147 (1998).

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

13. Gutell, R.R., Lee, J.C. & Cannone, J.J. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* **12**, 301-310 (2002).

14. Roth, A. & Breaker, R.R. The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem.* **78**, 305-334 (2009).

15. Deigan, K.E., Li, T.W., Mathews, D.H. & Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA* **106**, 97-102 (2009).

16. Berman, H.M. et al. The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242 (2000).

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

18. Gherghe, C.M., Leonard, C.W., Ding, F., Dokholyan, N.V. & Weeks, K.M. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J. Am. Chem. Soc.* **131**, 2541-2546 (2009).

19. Zhou, R., Berne, B.J. & Germain, R. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA* **98**, 14931-14936 (2001).

20. Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graph. Model.* **22**, 425-439 (2004).

21. Batey, R.T., Gilbert, S.D. & Montange, R.K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432**, 411-415 (2004).

22. Egli, M., Minasov, G., Su, L. & Rich, A. Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. *Proc. Natl. Acad. Sci. USA* **99**, 4302-4307 (2002).

23. Krasilnikov, A.S., Yang, X., Pan, T. & Mondragón, A. Crystal structure of the specificity domain of ribonuclease P. *Nature* **421**, 760-764 (2003).

24. Gherghe, C.M., Mortimer, S.A., Krahn, J.M., Thompson, N.L. & Weeks, K.M. Slow conformational dynamics at C2'-endo nucleotides in RNA. *J. Am. Chem. Soc.* **130**, 8884-8885 (2008).

25. Doi, M. in Oxford Science Publications 10-12 (Clarendon Press, Oxford; 1996).

26. Hyeon, C., Dima, R.I. & Thirumalai, D. Size, shape, and flexibility of RNA structures. *J. Chem. Phys.* **125**, 194905 (2006).

27. Holbrook, S.R. Structural principles from large RNAs. *Annu. Rev. Biophys.* **37**, 445-464 (2008).

28. Badorrek, C.S., Gherghe, C.M. & Weeks, K.M. Structure of an RNA switch that enforces stringent retroviral genomic RNA dimerization. *Proc. Natl. Acad. Sci. USA* **103**, 13640-13645 (2006).

29. Sharma, S., Ding, F. & Dokholyan, N.V. iFoldRNA: Three-dimensional RNA structure prediction and folding. *Bioinformatics* **24**, 1951-1952 (2008).

30. Massire, C. & Westhof, E. MANIP: an interactive tool for modelling RNA. *J. Mol. Graph. Model* **16**, 197-205 (1998).

31. Tan, R.K.Z., Petrov, A.S. & Harvey, S.C. YUP: A molecular simulation program for coarse-grained and multiscaled models. *J. Chem. Theory Comput.* **2**, 529–540 (2006).

32. Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. USA* **104**, 14664-14669 (2007).

33. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51-55 (2008).

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

35. Merino, E.J., Wilkinson, K.A., Coughlan, J.L. & Weeks, K.M. RNA structure analysis at single nucleotide resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *J. Am. Chem. Soc.* **127**, 4223-4231 (2005).

36. Wilkinson, K.A., Merino, E.J. & Weeks, K.M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1**, 1610-1616 (2006).

37. Westhof, E., Dumas, P. & Moras, D. Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallogr.* **A44**, 112-123 (1988).

38. Ke, A., Zhou, K., Ding, F., Cate, J.H. & Doudna, J.A. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature* **429**, 201-205 (2004).

39. Thore, S., Frick, C. & Ban, N. Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch. *J. Am. Chem. Soc.* **130**, 8116-8117 (2008).

40. Ruff, M. et al. Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). *Science* **252**, 1682-1689 (1991).

41. Zhang, Y. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.* **19**, 145-155 (2009).

42. Ferré-D'Amaré, A.R. & Doudna, J.A. Crystallization and structure determination of a hepatitis delta virus ribozyme: use of the RNA-binding protein U1A as a crystallization module. *J. Mol. Biol.* **295**, 541-556 (2000).

43. Ding, F., Tsao, D., Nie, H. & Dokholyan, N.V. Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure* **16**, 1010-1018 (2008).

44. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **A32**, 922-923 (1976).