# RNA STRUCTURE PREDICTION USING HIGH-THROUGHPUT CHEMICAL MODIFICATION TECHNIQUES

Christopher Andrew Lavender

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

Chapel Hill
2013

Approved by:

Nikolay Dokholyan

Howard Fried

Gary Pielak

Linda Spremulli

Kevin Weeks

**ABSTRACT**

Christopher Andrew Lavender: RNA Structure Prediction using High-Throughput Chemical Modification Techniques
(Under the direction of Kevin M. Weeks)

Functional RNA molecules require the formation of defined structures in order to perform their critical tasks in biology. Complete understanding of this structure-function relationship in RNA requires the elucidation of accurate RNA structural models. RNA chemical modification has proven to be an invaluable tool in the characterization of RNA structure. Recently, the throughput of RNA chemical modification approaches has increased significantly through the adaptation of chemical modification techniques to next-generation sequencing platforms. In this work, I create several new methodologies for the generation of accurate RNA structural models based on high-throughput RNA chemical modification analysis. First, I create a general methodology for predicting three-dimensional RNA structures based on RNA interactions implicated by biochemical and bioinformatic approaches. In this work, I develop a three-dimensional model for the hepatitis C virus internal ribosome entry site (HCV IRES) pseudoknot domain. This methodology is then applied to a new high-throughput chemical modification approach called RING-MaP (RNA interaction groups identified by mutational profiling). Implicated interactions from RING-MaP analysis allow for accurate prediction of RNA tertiary folds. Second, I create an algorithm for the comparison of high-throughput chemical modification data from related RNA sequences. Using SHAPE chemical modification alone, this approach allows recapitulation of ribosomal RNA alignments made using sequence identity. Chemical modification data for three HIV-related viral RNA genomes are then compared. Following creation of chemical modification-dependent alignments, statistically related RNA structures are found across the three viral genomes. Consensus secondary structures considering both chemical modification data and covariation are then made, recapitulating all known RNA structures in the HIV genome and suggesting previously undescribed functional elements.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS AND SYMBOLS

| | |
|---|---|
| 1M7 | 1-methyl-7-nitroisatoic anhydride |
| A | adenosine |
| Asp | aspartic acid |
| B | base pseudoatom |
| BLAST | Basic Local Alignment Search Tool |
| BHIS | brain heart infusion-supplemented medium |
| C | cytidine |
| cDNA | complementary DNA |
| $CO_2$ | carbon dioxide |
| CrPV | cricket paralysis virus |
| CRW | Comparative RNA Website |
| cryo-EM | cyro-electron microscopy |
| dA | deoxyadenosine |
| DMD | discrete molecular dynamics |
| DMS | dimethyl sulfate |
| DMSO | dimethyl sulfoxide |
| DNA | deoxyribonucleic acid |
| EDTA | ethylenediaminetetraacetic acid |
| FRET | fluorescence resonance energy transfer |
| G | guanosine |
| g | gram |
| HCV | hepatitis C virus |
| HEPES | N-2-hydroxyethylpiperazine-N'-ethanesulfonic acid |
| HHR | hammerhead ribozyme |
| HIV | human immunodeficiency virus |
| $H_2$ | molecular hydrogen |
| IRES | internal ribosome entry site |

| | |
|---|---|
| $k_B$ | Boltzmann constant |
| KCl | potassium chloride |
| kcal | kilocalorie |
| MaP | mutational profiling |
| $MgCl_2$ | magnesium chloride |
| min | minute |
| mM | millimolar |
| mol | mole |
| MSA | multiple sequence alignment |
| $N_2$ | molecular nitrogen |
| NAST | Nucleic Acid Simulation Toolkit |
| ng | nanogram |
| NGS | next-generation sequencing |
| nM | nanomolar |
| NMR | nuclear magnetic resonance spectroscopy |
| nt | nucleotide |
| $OD_{600}$ | optical density at 600 nanometers |
| P | phosphate pseudoatom |
| PARS | parallel analysis of RNA structure |
| PCR | polymerase chain reaction |
| Phe | phenylalanine |
| PK | pseudoknot |
| PPT | polypurine tract |
| ppv | positive predictive value |
| res | residue number |
| RING | RNA interaction group |
| RING-MaP | RNA interaction groups identified by mutation profiling |
| RMSD | root-mean-square deviation |

| | |
|---|---|
| RNA | ribonucleic acid |
| RNase H | ribonuclease H |
| RNase P | ribonuclease P |
| RPM | revolutions per minute |
| RRE | Rev response element |
| rRNA | ribosomal RNA |
| RT | reverse transcriptase |
| S | sugar pseudoatom |
| sens | sensitivity |
| SDS | sodium dodecyl sulfate |
| SIVcpz | chimpanzee simian immunodeficiency virus |
| SIVmac | rhesus macaque simian immunodeficiency virus, |
| SHAPE | selective 2'-hydroxyl acylation analyzed by primer extension |
| SHAPE-MaP | selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling |
| SHAPE-seq | selective 2'-hydroxyl acylation analyzed by primer extension sequencing |
| T | temperature |
| TAR | trans-activation response element |
| TE | 10 mM Tris (pH 8.0), 1 mM EDTA |
| TPP | thiamine pyrophosphate |
| Tris | tris(hydroxymethyl)aminomethane |
| tRNA | transfer RNA |
| tu | time unit |
| U | uridine |
| UTR | untranslated region |
| Å | Ångstrom |
| °C | degree Celsius |
| µL | microliter |

# CHAPTER 1: INTRODUCTION

## 1.1 Structure-function relationships in ribonucleic acid

As evidenced by its place in the central dogma of biology, ribonucleic acid (RNA) was originally thought to be a passive intermediate in the translation of genetic information from deoxyribonucleic acid (DNA) to protein. However, this viewpoint was challenged in the 1980s by the discovery of "ribozymes," discrete, structured RNA elements that catalyzed chemical reactions.[1] This, along with discovery of RNA interference pathways in the 1990s, challenged the traditional role of RNA in the central dogma. Today, RNA is considered an essential player in gene regulation.[2, 3]

Functional RNAs require the formation of defined three-dimensional structures in order to function properly.[3] As such, RNA follows the same structure-function paradigm established for proteins. Complete understanding of the function of a given RNA requires characterization of that RNA's structure. RNA structure is conceptualized as being hierarchical. Folding of an RNA structure begins with formation of base pairs, with the base-pairing arrangement of RNA defined as its secondary structure. Following formation of its secondary structure, an RNA molecule then folds into a final three-dimensional structure, known as its tertiary structure. Complete structural characterization of an RNA molecule requires knowledge of both its secondary and tertiary structures.

## 1.2 Established means of RNA structure characterization

Following the discovery of functional RNA molecules, approaches were developed to characterize RNA structure. RNA secondary structures could be predicted with high accuracy using covariance approaches.[4-6] These covariance models were developed considering a number of different RNA sequences that were related evolutionarily. An RNA covariance model describes the Watson-Crick base pairs that may be formed by all or most members of an RNA family. RNA tertiary structures were determined by adapting X-ray crystallography techniques, and later NMR techniques, to RNA.[7-10] X-ray crystallography techniques allowed atom-by-atom reconstructions of RNA structures based on X-ray diffraction patterns.

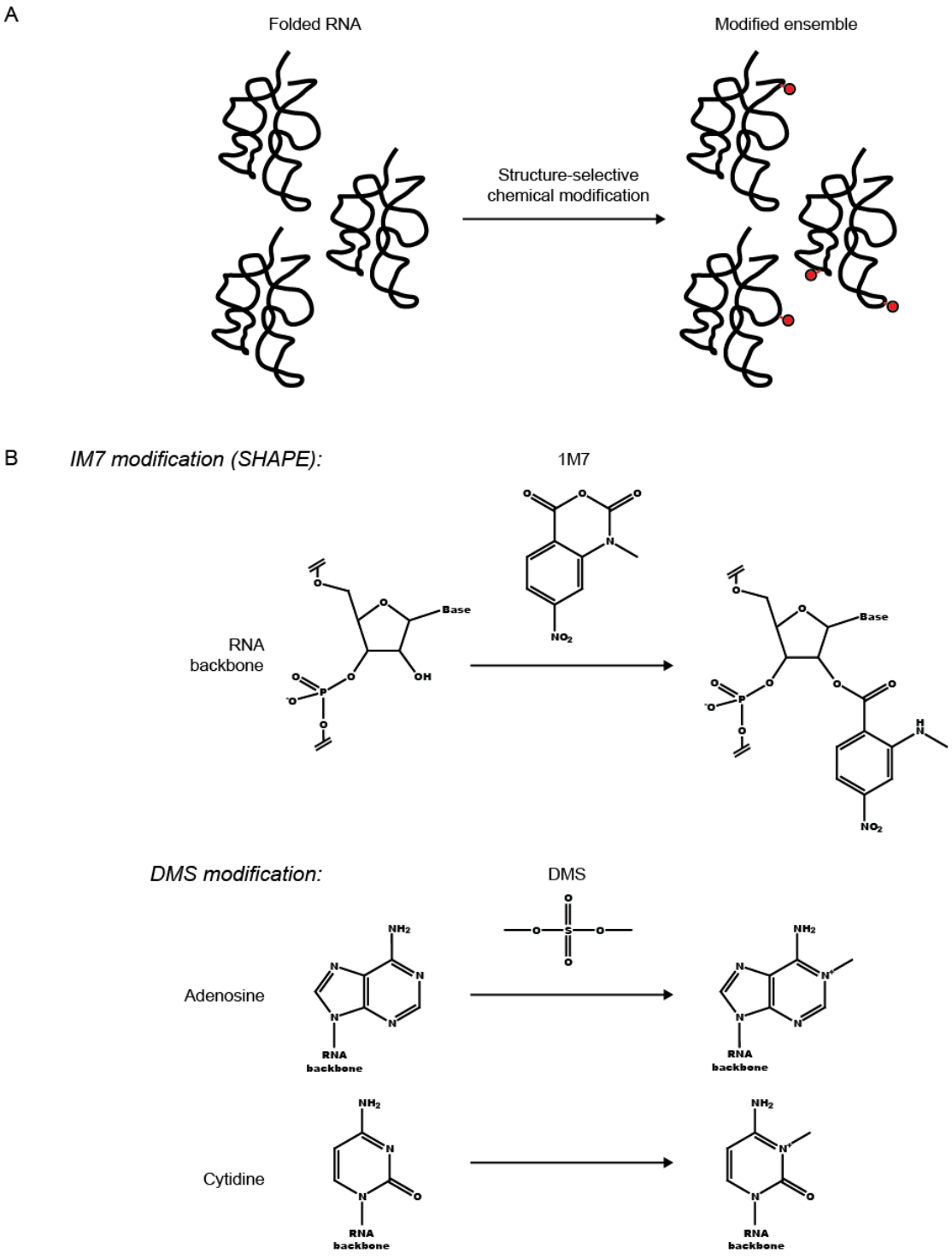



**Figure 1.1** Structure-selective chemical modification of RNA. (**A**) Generation of an ensemble of chemically modified RNA molecules. (**B**) Chemical modifications used in RNA structure characterization techniques. In SHAPE, 1M7 selectively modified the 2'-hydroxyl position of unstructured RNA residues. DMS selectively methylates the adenosines and cytidines of unstructured RNA positions.

However, covariation modeling and X-ray crystallography carry significant drawbacks. For an accurate covariation model to be developed, hundreds of different RNA sequences may be required, limiting the application of covariation modeling to only well-studied RNAs. X-ray crystallography determination has not been the wide-ranging success for RNA as it has been for proteins, owing largely to the difficulty in crystallizing RNA molecules. As such, only a small number of functional RNA structures have been able to be determined by X-ray crystallography. Both covariance modeling and X-ray crystallography are limited in the number of RNA molecules that are suitable targets.

**1.3 RNA structure prediction directed by biochemical experimentation**

To address the limitations of traditional techniques, new approaches combining computational structure prediction and biochemical characterization technique have been developed. Directing secondary structure prediction by SHAPE reactivity measurements has proven to be a robust and accurate means of predicting RNA secondary structure.[11, 12] In this approach, an electrophilic chemical moiety, termed a SHAPE reagent, reacts with the 2'-hydroxyl group of ribose sugars in a folded RNA molecule (Figure 1.1). The SHAPE reagent preferentially modifies nucleotides that are not constrained by base pairing or other molecular interactions. Positions of modification are then resolved using primer extension, with the reverse transcription reaction stopping at modified residues. Relative lengths of the cDNA library are used to assign positions of modification, with the extent of modification determined by comparison with a background control. The end result is a measure of nucleotide flexibility at each position in an RNA molecule. This information is then used to generate a pseudo-energy term that directs structure prediction in an RNA folding algorithm.[11]

Biochemical information has also been used to direct tertiary structure prediction of RNA molecules. In these approaches, biochemical measurements that describe or imply short contact distances between positions in an RNA molecule are used to constrain molecular dynamics simulations. Biochemical approaches used in these methods include chemical modification, crosslinking, and fluorescence resonance energy transfer (FRET) techniques.[13-18] These techniques have proven to be successful, but they are often not generalizable. Most have been applied to only a handful of RNAs, reflecting a difficulty in application or a limited number of suitable RNA targets. Moreover, modeling

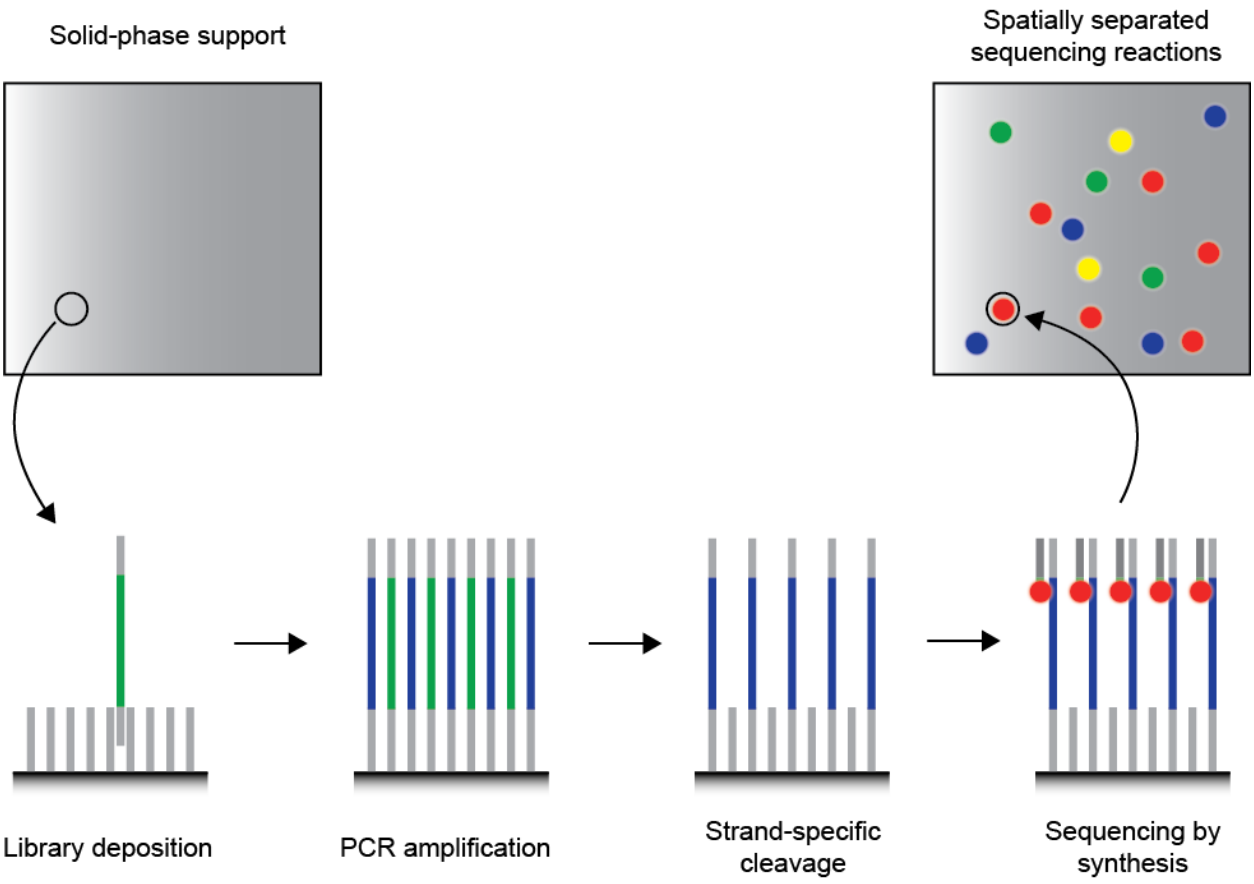



**Figure 1.2** Next-generation sequencing. Though there are many distinct next-generation sequencing platforms, in each approach many different sequencing reactions are performed in parallel while being spatially separated. In the Illumina platform, sequencing reactions are spatially separated on a glass chip. The glass surface is coated with sequencing library-specific primers. Individual DNA molecules anneal to these primers, and individual DNA colonies are amplified using PCR. Individual colonies are characterized using a sequencing-by-synthesis approach using fluorescently labeled DNA nucleotides.

approaches are often designed to incorporate information from only a single biochemical technique and may not be immediately adapted to another.

**1.4 The emergence of high-throughput RNA structure assays**

The development of biochemical RNA assays has been tightly coupled to the development of nucleic acid sequencing approaches. For instance, in many RNA structure assays such as SHAPE, chemical modifications or strand cleavages are resolved using size-based comparisons made by electrophoresis. This same methodology is used in Sanger sequencing, where base identities are determined by comparing the relative size of DNA fragments generated in the presence of strand-ending dideoxynucleotides.[19] Just as the throughput of Sanger sequencing was increased through the transition from gel electrophoresis to capillary electrophoresis, similar gains were made by applying that same transition to SHAPE.[20]

In recent years, the throughput of sequencing has increased exponentially as next-generation sequencing (NGS) techniques have emerged.[21] There are a number of NGS platforms available for use today that utilize different sequencing approaches, but they each share a common feature: A multitude of sequencing reactions are performed in parallel and resolved simultaneously (Figure 1.2). The number of sequences generated in a single NGS experiment range from tens of thousands to hundreds of millions.

Researchers have been quick to adapt RNA biochemical assays to NGS approaches. In the SHAPE-seq method, cDNA generated in a SHAPE experiment is used to generate an NGS library following ligation of DNA adapters and PCR amplification.[22] In PARS, RNA cleaved by structure-specific enzymes is converted to DNA following RNA-RNA ligation of primer binding sites.[23] In both techniques, the positions of chemical modification or enzymatic cleavage are found at sites of DNA-RNA or RNA-RNA ligation. Enhanced throughput is possible with both methods; however, these approaches do a poor job of recapitulating experimental results from low-throughput approaches.[24] This is believed to be due in part to strong biases introduced in ligating single-stranded DNA or RNA molecules.[25]

A new approach developed by the Weeks lab allows resolution of chemical modification positions by NGS while avoiding the ligation biases associated with other techniques. In this approach, termed mutational profiling (MaP), positions of chemical modification are resolved by analyzing mutation rates in cDNA following reverse transcription of modified RNA transcripts (Siegfried *et al*., in preparation) (Figure

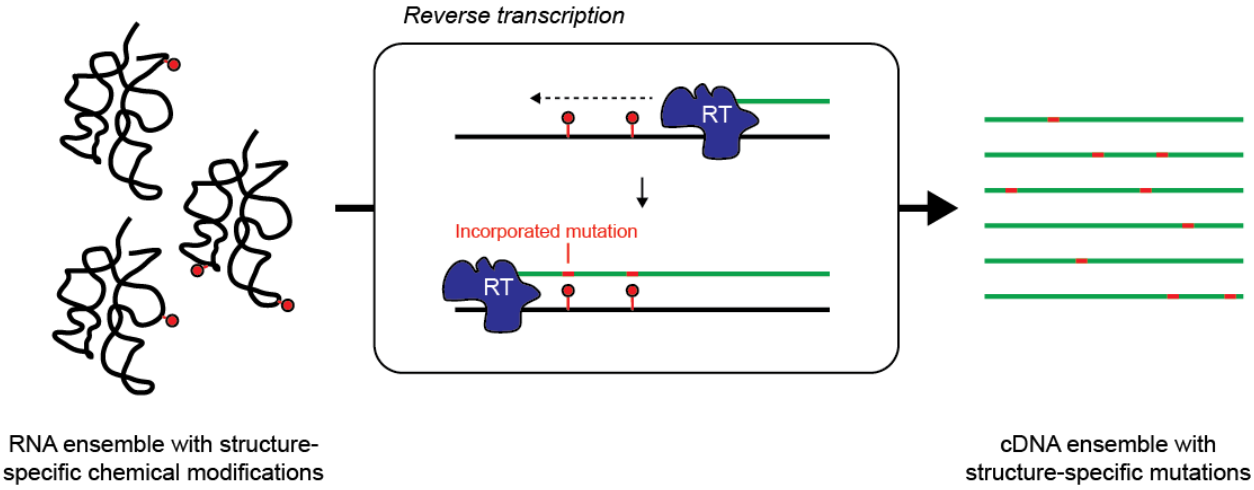



**Figure 1.3** Mutational profiling. In the mutational profiling approach, chemically modified RNA is reverse transcribed. Under specific reaction conditions, reverse transcriptase incorporates mutations in cDNA at positions that correspond to chemically modified RNA nucleotides. Reverse transcription generates an ensemble of cDNA molecules with mutations based on RNA chemical modifications.

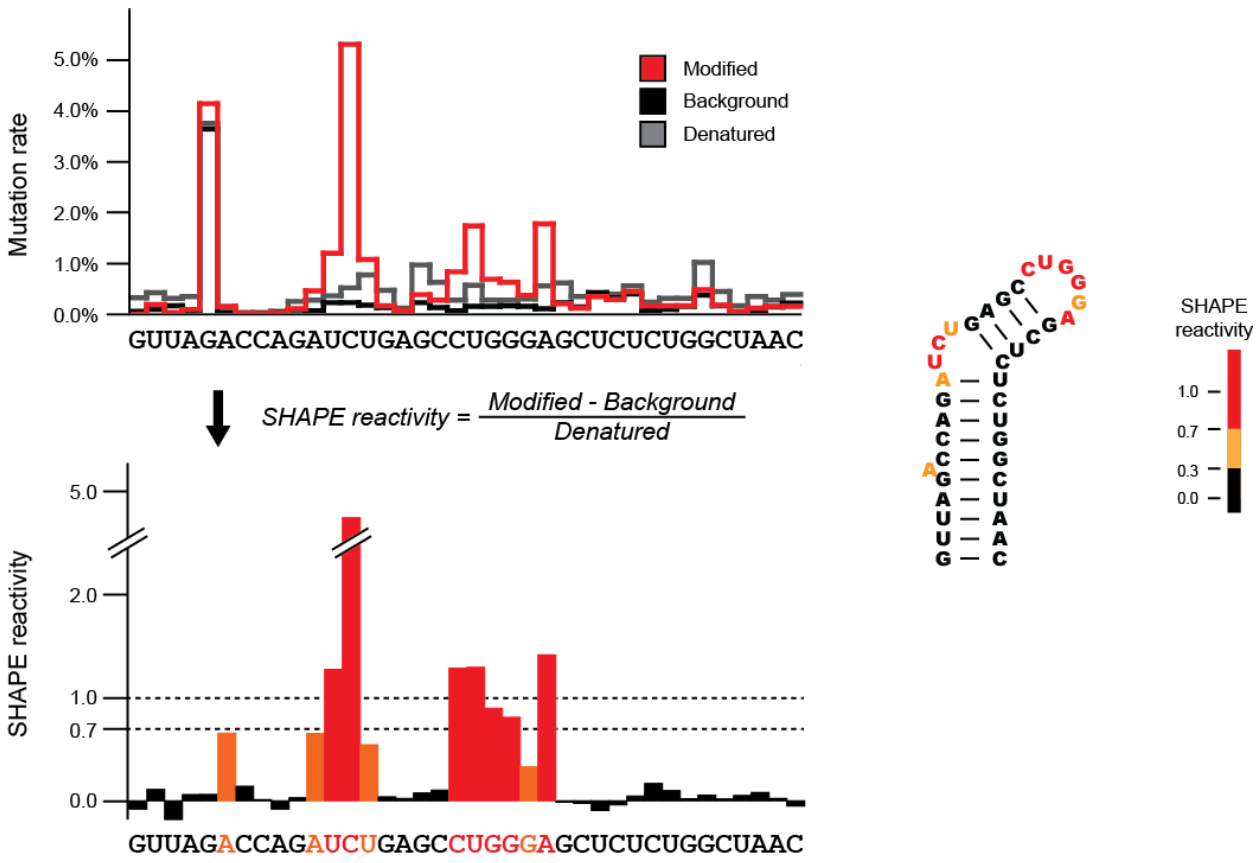



**Figure 1.4** Representative SHAPE data for the HIV-1 TAR stem loop (Siegfried *et al.*, in preparation). Mutation rates from chemically modified RNA are compared to background and denatured controls to determine a SHAPE reactivity value for each position in an RNA molecule. Positions with high SHAPE reactivity (orange and red) are found in looped regions, while positions with low SHAPE reactivity (black) are found in base-paired regions.

1.3). This approach exploits a recently described phenomenon in which reverse transcriptase will read through a chemical modification under specific conditions but with an increased chance of incorporating a mutation at the modified position. By analyzing mutation rates over a cDNA-based sequencing library, the extent of chemical modification at each position in the RNA may be found (Figure 1.4).

To date, MaP has been successfully applied to SHAPE and dimethyl sulfate (DMS) modification experiments. By coupling these experiments to NGS techniques, a number of small RNA molecules under different experimental conditions may be analyzed simultaneously. Additionally, genome- and transcriptome-scale interrogation of RNA structures is possible. However, the development of new analytical and bioinformatic techniques is required in order to take advantage of the high throughput of these methods.

## 1.5 Research overview

The goal of my research has been the development of new techniques to generate RNA structure models. Central to this goal has been the development of new techniques for the analysis of data from high-throughput RNA structure assays. The techniques developed fit broadly into two different research directions. The first is the directed modeling of three-dimensional RNA folds based on data from biochemical approaches. The second is the prediction of consensus secondary structures based on chemical modification data from related RNA molecules. Under both directions, new methods and new biologically relevant hypotheses have been made.

### 1.5.1 Tertiary structure prediction using RNA contacts implicated by high-throughput chemical modification techniques

A modeling methodology was created for incorporation of biochemical data into molecular dynamics simulations in order to predict RNA tertiary folds. Based on RNA-RNA contacts implicated by a variety of biochemical and bioinformatic techniques, energy potentials were used to bias molecular dynamics simulations of RNA molecules. In test predictions, statistically significant native-like folds ($p < 0.01$) were predicted for four RNA molecules ranging in length from 45 to 158 nucleotides.

Using this approach, a tertiary fold was predicted for the previously undescribed hepatitis C virus internal ribosome entry site (HCV IRES) pseudoknot domain. The model generated agreed with independent solvent accessibility measurements and fit well into published cryo-electron microscopy density maps. Based on the model, we hypothesize that the pseudoknot domain is involved in positioning

of the AUG start codon of the IRES into the ribosomal initiation site. The similarity of the model's topology to that of tRNA suggests that the IRES employs molecular mimicry as a functional strategy.

This general methodology was adapted to predict RNA folds based on RING-MaP, a new technique developed to find RNA interactions based on NGS-resolved DMS modification experiments. Despite the variability in implicated RNA interactions, native-like folds were predicted for three test case RNAs. By coupling this modeling methodology with a generalizable biochemical approach, structure determination may be applied to many new RNA targets.

**1.5.2 Sequence comparisons of high-throughput chemical modification data**

A method was developed to compare high-throughput chemical modification data across related RNAs. The method compares SHAPE reactivities between any two nucleotides by a pairwise scoring approach. Using this scoring system, sequence alignments for 16S and 23S rRNA were generated considering SHAPE chemical modification data alone. Alignment quality was similar to that of sequence-identity based approaches.

SHAPE-dependent sequence alignments were then performed across three HIV-related RNA genomes. Using linear regression analysis, areas across these genomes that were statistically interdependent were found. Based upon sequence alignment and SHAPE reactivities, consensus secondary structure models were found for the HIV-related strains. All known functional RNA elements of the HIV genome were recapitulated. A number of additional consensus elements were found, indicating possible functional elements newly discovered in this work.

## 1.6 References


1. Kruger, K.*, et al.* (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena, *Cell 31*, 147-157.

2. Fire, A.*, et al.* (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, *Nature 391*, 806-811.

3. Gesteland, R. F., Cech, T., and Atkins, J. F. (2006) *The RNA World*, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

4. Gutell, R. R.*, et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods, *Nucleic Acids Res 20*, 5785-5795.

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

6. Woese, C. R.*, et al.* (1980) Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence, *Nucleic Acids Res 8*, 2275-2293.

7. Kim, S. H.*, et al.* (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA, *Science 185*, 435-440.

8. Klug, A.*, et al.* (1974) Conservation of the molecular structure of yeast phenylalanine transfer RNA in two crystal forms, *Proc Natl Acad Sci U S A 71*, 3711-3715.

9. Cheong, C., Varani, G., and Tinoco, I., Jr. (1990) Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC, *Nature 346*, 680-682.

10. Heus, H. A., and Pardi, A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops, *Science 253*, 191-194.

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

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

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

14. Gherghe, C. M.*, et al.* (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics, *J Am Chem Soc 131*, 2541-2546.

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

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

17. Yu, E. T.*, et al.* (2008) MS3D structural elucidation of the HIV-1 packaging signal, *Proc Natl Acad Sci U S A 105*, 12248-12253.

18. Stephenson, J. D.*, et al.* (2013) Three-dimensional RNA structure of the major HIV-1 packaging signal region, *Structure 21*, 951-962.

19. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors, *Proc Natl Acad Sci U S A 74*, 5463-5467.

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

21. Ansorge, W. J. (2009) Next-generation DNA sequencing techniques, *N Biotechnol 25*, 195-203.

22. Lucks, J. B.*, et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), *Proc Natl Acad Sci U S A 108*, 11063-11068.

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

24. Weeks, K. M. (2011) RNA structure probing dash seq, *Proc Natl Acad Sci U S A 108*, 10933-10934.

25. Hafner, M.*, et al.* (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries, *RNA 17*, 1697-1712.

## CHAPTER 2: ROBUST AND GENERIC RNA MODELING USING INFERRED CONSTRAINTS

### 2.1 Introduction

Critical RNA structures directly regulate gene expression, splicing, and translation.[1] but the structures of most biologically important RNA folds are currently unknown. Recent studies highlight significant successes in *ab initio* structure prediction of local helical structure and of small RNA motifs.[2-4] However, the ability of current approaches to predict RNA structure accurately decreases rapidly with increasing RNA size. *De novo* prediction of large RNA structures with complex, nontrivial, three-dimensional folds from sequence alone remains beyond the realm of current automated algorithms. A compelling alternative is to develop modeling methods for facile incorporation of readily obtained experimental information.

Long-range constraints for RNA modeling can be inferred from a variety of biochemical and bioinformatic techniques, ranging from chemical footprinting and cross-linking to sequence covariation.[5-7] Algorithms devised thus far are making significant progress toward the goal of incorporating specific classes of tertiary structure information into RNA structure refinement.[8-11] However, current refinement approaches still make large assumptions about the nature of the constraint information used and are closely tied to the specific techniques employed to infer long-range interactions.

To address these challenges, we develop a generic and efficient approach for accurately predicting RNA folds using tertiary structure information as inferred from diverse biochemical or bioinformatic techniques. Distance constraints are incorporated into a discrete molecular dynamics (DMD) engine[3] that uses a single refinement approach for all classes of tertiary structure constraint information. RNA nucleotides are represented as three pseudoatoms corresponding to the phosphate (P), sugar (S), and base (B) moieties (Figure 2.1A). Three pseudoatoms are sufficient for the development of nucleotide-resolution RNA models with rigid base-paired helices and physically meaningful base stacking interactions, while still allowing large RNAs to be refined efficiently.

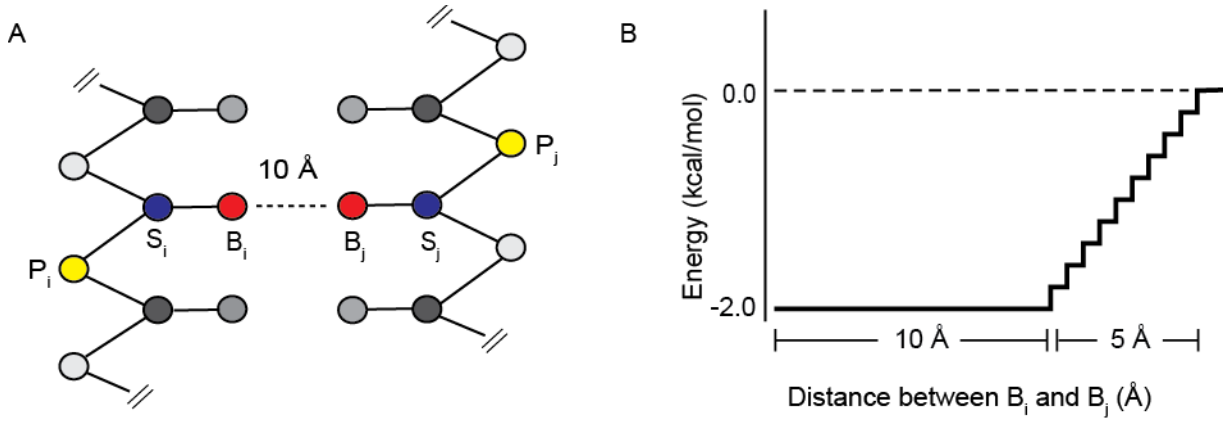



**Figure 2.1** Long-range constraints used to incorporate inferred RNA contacts. (**A**) Constraint shown on the DMD reduced representation of RNA. Constraint potentials were based on the distance between base pseudo-atoms. (**B**) Square-well representation of long-range constraints. A maximum energy bonus of -2.0 kcal/mol was applied when base pseudo-atoms were between 10 Å of one another.

**2.2 Generic tertiary constraint system**

Inferred pairwise tertiary constraints are incorporated via a generic constraint system that uses a potential well with an effective length of 15.0 Å and a depth of 2.0 kcal/mol between base pseudoatoms (Figure 2.1B). This constraint system is compatible with techniques that do not directly provide distance information but instead merely imply pairwise interactions, as with mutational studies.

**2.3 Selection of test-case RNAs**

Four RNAs were selected to benchmark constrained structure refinement: domain III of the cricket paralysis virus internal ribosome entry site (CrPV) (49 nucleotides), a full-length hammerhead ribozyme from *Schistosoma mansoni* (HHR) (67 nucleotides), *Saccharomyces cerevisiae* $tRNA^{Asp}$ (75 nucleotides), and the P546 domain of the *Tetrahymena thermophilia* group I intron (P546) (158 nucleotides). Each of these RNAs has a complex three-dimensional fold dependent both on local helical structure and on long-range tertiary interactions. Prior to publication of the high-resolution structures,[12-15] significant biochemical or bioinformatic data describing tertiary interactions were available for each RNA. The secondary structure was also known with good accuracy in each case. Only this prior information (Table 2.1) was used during refinement.

**2.4 Automated refinement protocol**

A single generic and completely automated refinement protocol was applied to each RNA. Simulations begin with the RNA strand in an extended conformation at a high temperature. Constraints based on the secondary structure are included, and the molecular system is annealed to allow helices to form. Constraints for inferred tertiary interactions are incorporated, and the RNA is cooled to a final target temperature. RNA structures from this step (100000) are subjected to automated clustering. The centroid of the most populated cluster is selected as the final predicted structure. Given our refinement model, this structure is representative of the lowest-free energy state.

**2.5 Final predicted models**

Refined models for all four test RNAs are accurate (Figure 2.2). The root-mean-square deviations (RMSDs) of the phosphate backbone relative to the accepted structures for the CrPV, HHR, $tRNA^{Asp}$, and P546 RNAs were 3.6, 5.4, 6.4, and 11.3 Å, respectively. Analysis of the RNA structure prediction

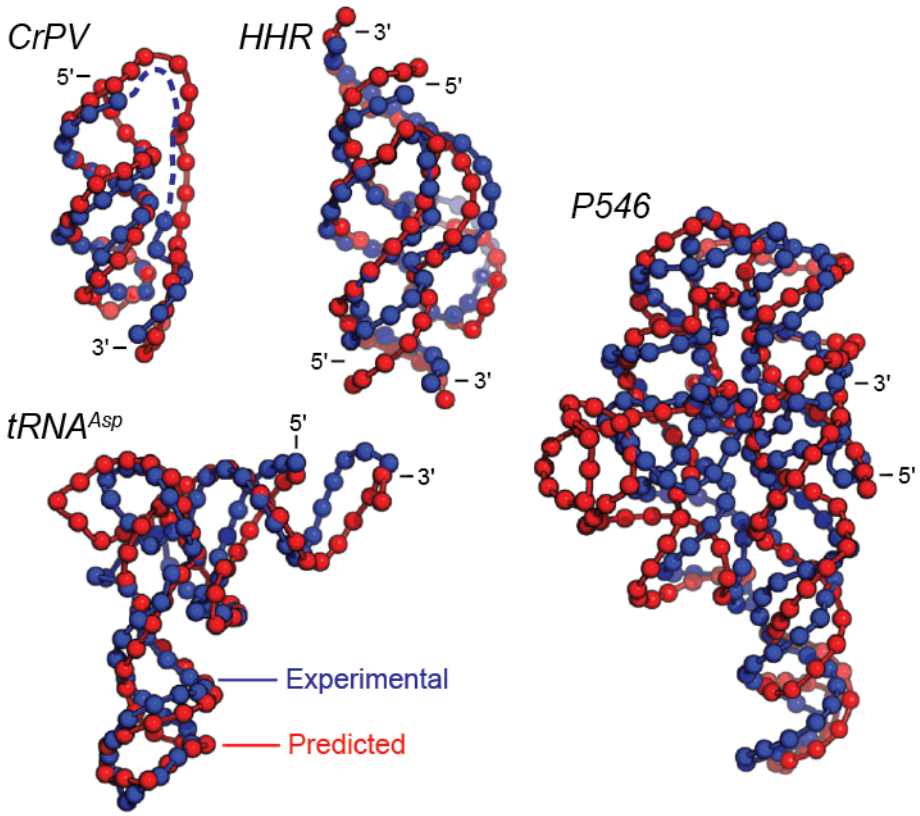



**Figure 2.2** Comparison of predicted models (red) and accepted high-resolution structures (blue). Spheres indicate phosphorus atoms (high-resolution structures) or phosphate pseudo-atoms (predicted models). The root-mean-square deviations for CrPV, HHR, tRNA, and P546 RNAs are 3.6, 5.4, 6.4, and 11.3 Å, respectively.

**Table 2.1** Secondary and tertiary interactions used in modeling. All secondary and tertiary structure information was available prior to crystallographic structure determination. Note that the numbering of the residues is that of the sequences used in modeling (each modeled RNA beginning with position 1) and is not necessarily the same as the source references.

| RNA | Secondary structure | Tertiary interactions | Inferred pair-wise tertiary contacts | Crystal structure |
|---|---|---|---|---|
| CrPV | ref. [16] | DMS and kethoxal probing,[16] mutational analysis[17] | U16-G47<br>A17-U46<br>G18-C45<br>G19-C44<br>U20-A43<br>A21-U42<br>G22-U41 | 3b31 [15] |
| HHR | [18] | Mutational analysis [19, 20] | A7-A27<br>A7-U30<br>C61-A27<br>C61-U30 | 2goz [14] |
| $tRNA^{asp}$ | [21] | Chemical protection,[21] sequence covariation [22] | U8-A14<br>A15-U47<br>G18-U54<br>U19-C55 | 2tra [12] |
| P546 | [23] | Mutational analysis, hydroxyl radical protection [24] | A51-C121<br>A51-G148 | 1gid [13] |
| | | Mutational analysis, DMS modification [25] | A81-C7<br>A81-110 | |
| HCV-PK | [26] | Mutational analysis, chemical and enzymatic probing, thermodynamic calculations [27] | U72-A96<br>G73-U95<br>C74-G94<br>G75-C93<br>A76-U92<br>G77-C91 | *none* |

significance (*p*-value)[28] shows that the probabilities that these models result from chance are small ($2 \times 10^{-3}$, $2 \times 10^{-5}$, $3 \times 10^{-6}$, and $\leq 10^{-6}$, respectively).

There are two critical results from this analysis of RNAs with known structures. First, native-like RNA folds were obtained in every case despite the diversity of structural information used to constrain refinement (Table 2.1). Second, prediction quality was maintained as RNA size increased from a 49-nucleotide pseudoknot to a 158-nucleotide RNA domain with a complex tertiary structure (Figure 2.2).

**2.6. Prediction of the HCV IRES pseudoknot domain**

Having shown that this fully automated approach recapitulates native-like folds for diverse, well-characterized RNAs, we sought to apply this algorithm to an RNA for which extensive biochemical information exists but whose structure is unknown. We focused on the pseudoknot domain in the hepatitis C virus (HCV) internal ribosome entry site (IRES).

IRES elements bypass canonical cap-dependent eukaryotic translation initiation by directly recruiting ribosomes to internal sequences in an mRNA.[29, 30] Structural studies have significantly improved our understanding of functional mechanisms of IRES elements.[15, 31] High-resolution structures are available for many elements of the HCV IRES;[32, 33] however, the three-dimensional fold for the pseudoknot domain (HCV-PK) has not been determined. The pseudoknot domain consists of a pseudoknot at the base of domain III (dIII) and its flanking structures (Figure 3A). Mutation of the pseudoknot inhibits translation initiation in HCV replication.[27, 34] Compensatory mutations that restore the pseudoknot do not always restore HCV translation activity, suggesting that sequence conservation is required for functions beyond base pairing. The pseudoknot domain contains the AUG start codon for translation of the HCV polypeptide (yellow in Figure 2.3A). Solvent accessibility experiments show the pseudoknot domain is the most highly structured element in the IRES.[35] Extensive available biochemical information and intense biomedical interest make the HCV-PK RNA an ideal candidate for deriving biological insights based on structural modeling.

A three-dimensional model for the HCV-PK domain RNA was refined using the same fully automated folding algorithm as for the four test RNAs. Base pairs in the pseudoknot were modeled as generic tertiary constraints.

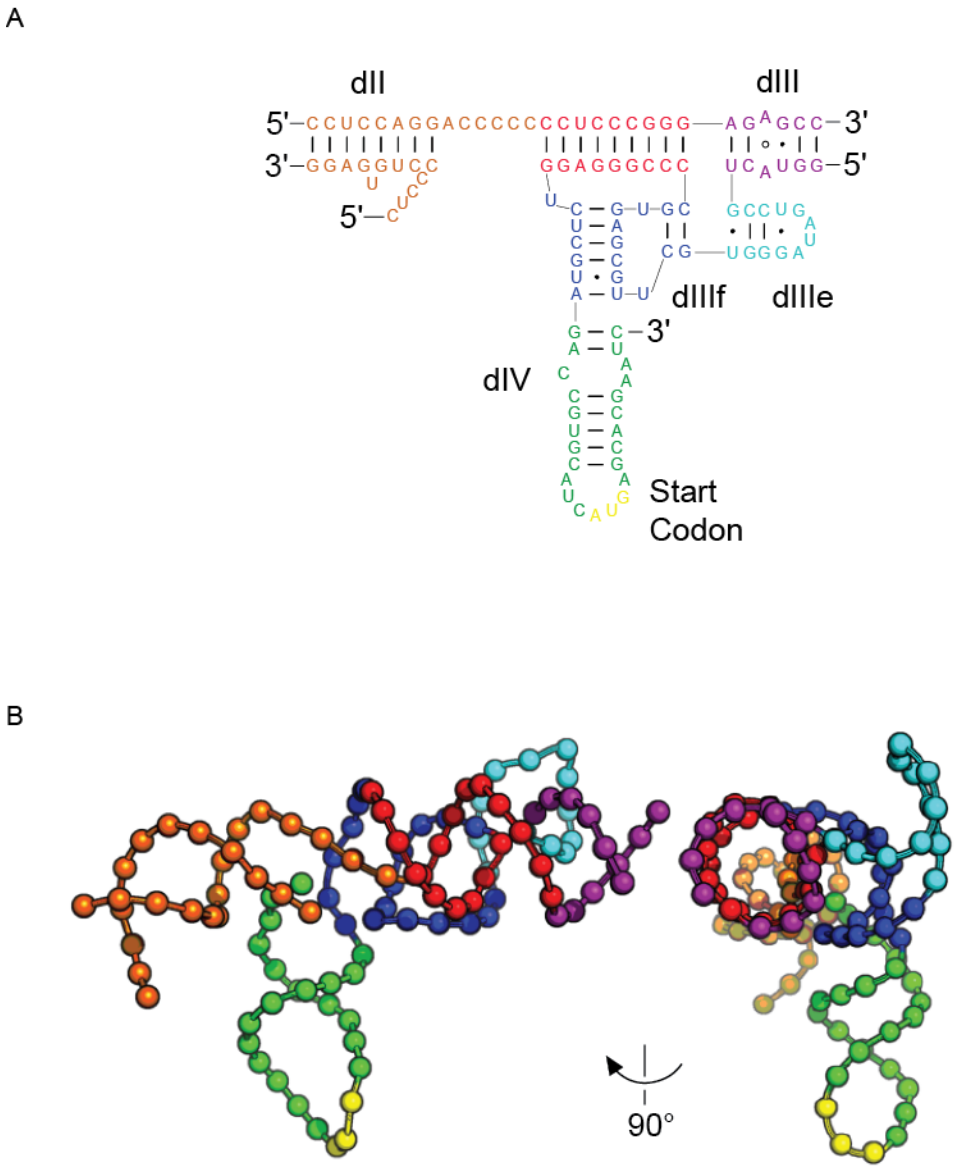



**Figure 2.3** Predicted model for the HCV-PK domain. (**A**) Secondary structure of the HCV-PK domain. Stem loop structures are annotated by color. (**B**) Predicted tertiary fold. Spheres indicate the positions of phosphate pseudo-atoms. The coloring of stem loops is consistent across (**A**) and (**B**).

The predicted HCV-PK structure is dominated by two structural features (Figure 2.3B). The first is the four-way junction comprised of stems at the base of dIII (red and purple), dIIIe (cyan), and dIIIf (blue). The second consists of base stacking interactions between the pseudoknot (blue) and dIV (green). The nucleotide linkages between these two motifs are short and lock the dIV helix in a conformation perpendicular to the plane described by the helices of the four-way junction.

**2.7 Independent support of the HCV-PK model**

Two classes of independent experiments support the proposed structure for the HCV-PK RNA. First, the predicted tight RNA folding in the four-way junction and pseudoknot is supported by protection from hydroxyl radical cleavage, indicative of solvent inaccessible regions of the RNA backbone.[35] Solvent inaccessible regions fall precisely in the interior of the four-way junction and at the interface of this element with the pseudoknot (red and yellow spheres in Figure 2.4B).

Second, the HCV-PK model is consistent with cryo-EM electron density maps of the IRES−ribosome complex. Our model of the HCV-PK is that of the uncomplexed IRES, and conformational changes occur in both the ribosome and IRES when the IRES interacts with ribosomal subunits and translation initiation factors.[36, 37] For example, domain IV likely unfolds to allow positioning of the start codon in the P-site.[38-40] Nevertheless, the core of our model fits well in the density assigned to the pseudoknot domain in the cryo-EM electron density maps of the IRES−ribosome complex.[37] The critical correlations are that dII and dIII (orange and purple, respectively, in Figure 2.4A) are positioned to connect sensibly with the rest of the IRES, and the perpendicular orientation of the pseudoknot (blue) allows the AUG start codon in dIV (yellow in Figures 2.4A and 2.5) to be positioned in or near the mRNA channel.

**2.8 Discussion**

**2.8.1 Comparison with similar prediction approaches**

Our approach compares favorably to other coarse-grained RNA modeling approaches. Folds for $tRNA^{Phe}$ and the P546 domain have been predicted with the program NAST in which each RNA nucleotide is represented by a single pseudoatom.[2] NAST modeling was constrained using structure information similar to that used in our refinements. Of the resulting models, the most accurate had RMSDs relative to the accepted structure of 8.0 and 16.3 Å for tRNA and P546, respectively, whereas our

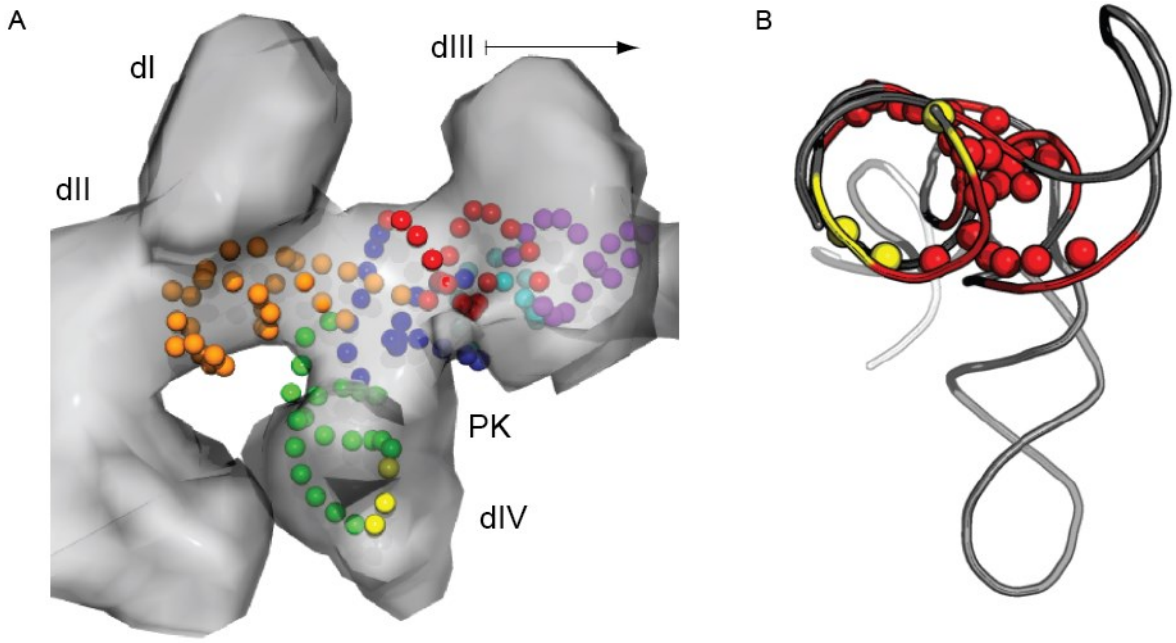



**Figure 2.4** Validation of the HCV-PK model by independent studies. (**A**) Fitting of the HCV-PK model into a cryo-EM density map.[37] (**B**) Positions in the HCV-PK auto-protected from hydroxyl radical cleavages.[35] Modest (yellow) and strong (red) auto-protection indicates solvent inaccessibility at a given RNA residue. The phosphate backbone for the model is shown with a cartoon tube, while regions of protection are additionally represented with a sphere at sugar pseudo-atom positions.

approach yields smaller RMSDs of 6.4 and 11.3 Å, respectively. NAST simulations used 300 h per RNA, as compared to 18−40 real-time computing hours for the DMD-based refinements (Figure 2). These comparisons highlight both the accuracy and efficiency of our constrained DMD approach.

### 2.8.2 Functional hypotheses for the HCV-PK domain

Several functional hypotheses are consistent with the predicted model. First, the HCV-PK RNA is L-shaped, similar to tRNA, and can be aligned with yeast $tRNA^{Asp}$ (Figure 2.6). Formation of a tRNA-like structure is consistent with biochemical studies showing that the HCV IRES is cleaved by the tRNA-recognizing ribonuclease RNase P.[41, 42] tRNA mimicry also rationalizes the presence of a seven-nucleotide loop at the end of domain IV, a structural feature that is generally uncommon in RNA but present in the anticodon loops of most tRNAs.

A recent structural study also yielded evidence of tRNA mimicry in domain III of the CrPV IRES.[15] Though the HCV-PK model and CrPV experimental structure have distinct folds, both support tRNA mimicry as a common strategy employed by IRES structures and are consistent with extensive examples of tRNA mimicry in biologically diverse RNAs.[43]

Second, the perpendicular orientation of the pseudoknot relative to the four-way junction may function to position the AUG start codon for translation initiation. In cryo-EM maps of both the 40S− and 80S−IRES complexes, density corresponding to the pseudoknot domain is adjacent to the channel occupied by the mRNA template during translation.[36, 37] Thus, dIV and, specifically, the AUG start codon will be positioned near the ribosome mRNA exit site.

These observations support a model in which the IRES pseudoknot domain docks initially near the ribosome exit channel, facilitated by its tRNA-like structure (Figure 2.5, left). Our model suggests additional conformational changes are required in the IRES and ribosome for the AUG start codon to fully occupy the mRNA channel. A modest unfolding of the dIV helix would then allow this element to serve as the mRNA template for translation of the HCV polyprotein (Figure 2.5, right).

## 2.9 Conclusion

RNA structure refinement using inferred constraints consistently yields nativelike models for RNAs spanning 49−158 nucleotides. This approach does not require a specific optimized form for the long-range constraints but does require knowledge of through-space tertiary interactions. The success of

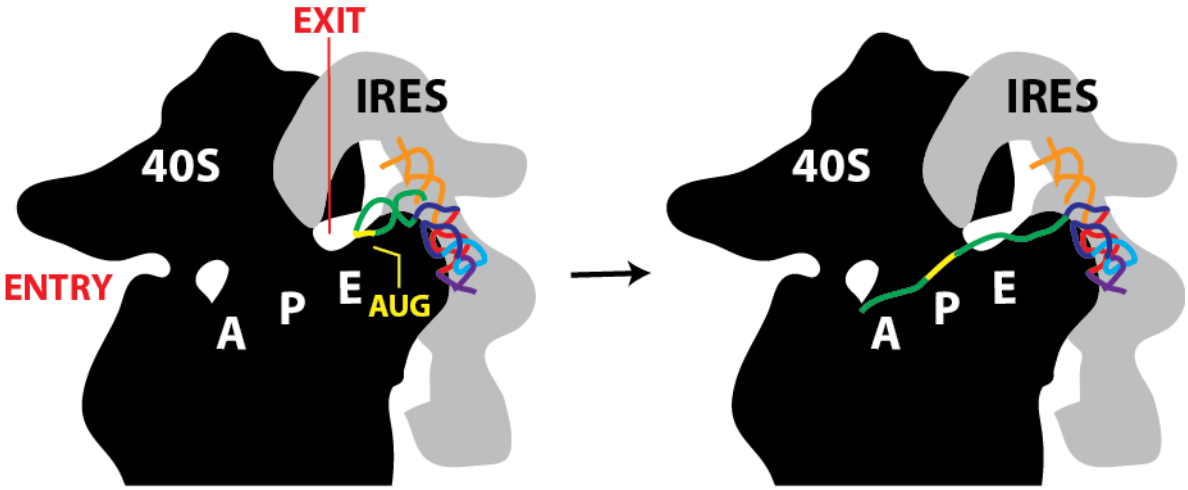



**Figure 2.5** Docking of the HCV IRES RNA into the mRNA channel of the 40S ribosome. Cartoons of the 40S subunit (black) and HCV IRES (gray) are based on cryo-EM studies.[36, 37] The AUG start site codon, mRNA entry and exit sites, and tRNA binding sites are labeled on the 40S subunit. The HCV-PK model is colored and positioned in the same orientation relative to the cryo-EM density as Figure 2.4A.

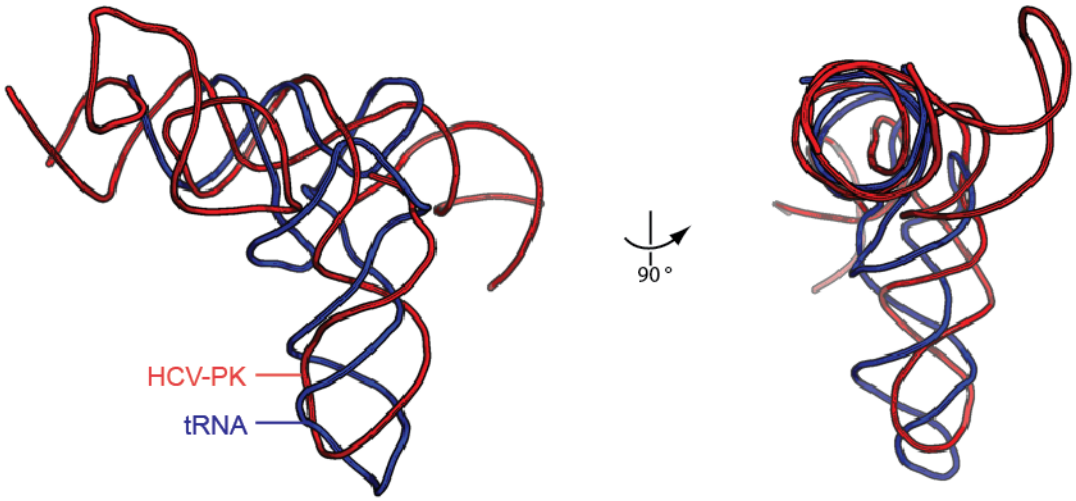



**Figure 2.6** Alignment of the HCV-PK model (red) with tRNA (blue).[12]

this approach implies that knowledge of only a few long-range constraints is sufficient to refine accurate folds for many RNAs with complex structures.

The HCV-PK domain model rationalizes substantial preexisting biochemical information for this RNA and provides specific and novel functional insights useful for guiding future hypotheses and experiments. RNA structure refinement using inferred constraints holds significant promise for understanding the functions of many biologically important RNAs whose analysis is recalcitrant to high-resolution approaches.

**2.10 Experimental methods**

**2.10.1 Sequences of RNA models**

CrPV: 5'-UGCGG UUUUU CAGAU UAGGU AGUCG AAAAA CCUAA GAAAU UUACC UGCU-3'

HHR: 5'-GGAUG UACUA CCAGC UGAUG AGUCC CAAAU AGGAC GAAAC GCCAA AAGGC GUCCU GGUAU CCAAU CC-3'

$tRNA^{Asp}$: 5'-UCCGU GAUAG UUUAA UGGUC AGAAU GGGCG CUUGU CGCGU GCCAG AUCGG GGUUC AAUUC CCCGU CGCGG AGCCA-3'

P546: 5'-GAAUU GCGGG AAAGG GGUCA ACAGC CGUUC AGUAC CAAGU CUCAG GGGAA ACUUU GAGAU GGCCU UGCAA AGGGU AUGGU AAUAA GCUGA CGGAC AUGGU CCUAA CCACG CAGCC AAGUC CUAAG UCAAC AGAUC UUCUG UUGAU AUGGA UGCAG UUC-3'

HCV-PK: 5'-CUCCC CUGUG AGGUU UUCCU CCAGG ACCCC CCCUC CCGGG AGAGC CUUUU GGUAC UGCCU GAUAG GGUGC UUGCG AGUGC CCCGG GAGGU CUCGU AGACC GUGCA UCAUG AGCAC GAAUC-3'

Sequences correspond to those used in prior structural studies.[16, 18, 21, 23] In HHR, where the crystallographic model includes two strands, the strands were connected with a 5'-AAAA-3' linker. The HCV-PK sequence was taken from genotype 1b.[44] The HCV-PK model includes nucleotides 40-52, 111-139, and 285-354, comprising the base of domain II, the four-way junction at the base of domain III, and domain IV. During modeling, these three segments of the HCV IRES sequence were connected using two 5'-UUUU-3' linkers.

**2.10.2 Secondary and tertiary constraint sources**

Sources for the secondary structures and tertiary contact information used to constrain DMD refinement are outlined in detail in Table 2.2. Base pairs in the secondary structures were constrained as described previously.[3, 10] Tertiary contacts were imposed using the generic constraint system created in this work.

### 2.10.3 RNA discrete molecular dynamics engine (DMD)

The DMD engine[3] models each nucleotide as three pseudo-atoms corresponding to the phosphate, sugar, and base moieties. Pair-wise interactions including base pairing, base stacking, packing interactions, and electrostatic repulsion are approximated using square-well potentials. In base-paired regions of the model, distance constraints between base and phosphate pseudo-atoms are used to maintain the rigid structure characteristic of the RNA double helix.[10]

The engine used in this work extends base stacking interactions to both base-paired and single-stranded RNA regions. Single-stranded stacking interactions contribute $-k_BT$ to the RNA free energy, where $k_B$ is the Boltzmann constant and $T$ equals 300 K ($-k_BT$ = -0.6 kcal/mol). This is one-half of the free energy contribution assigned to base-paired stacking interactions ($-2k_BT$).[8]

### 2.10.4 General constraint system

Distance constraints were included between the base pseudo-atoms of residues inferred to participate in pair-wise interactions. A maximum free energy bonus of 2.0 kcal/mol was applied when base pseudo-atoms were within 10.0 Å of each other. For each 0.5 Å beyond the inter-pseudo-atom distance of 10.0 Å, the bonus was reduced by 0.2 kcal/mol, giving the constraint an effective length of 15.0 Å. The attractive potential is described in Figure 1 and in the following potential function, where d is the distance between base pseudo-atoms:

$$E_{\text{constraint}} = \begin{cases} -2.0\,\text{kcal/mol}, 0 \le d < 10.0\,\text{Å} \\ -1.8\,\text{kcal/mol}, 10.0 \le d < 10.5\,\text{Å} \\ \vdots \\ -0.2\,\text{kcal/mol}, 14.5 \le d < 15.0\,\text{Å} \\ 0, 15.0\,\text{Å} \le d \end{cases}$$

### 2.10.5 General refinement protocol

Simulations begin with the RNA strand in an extended linear conformation at a high temperature. The RNA sequence and constraints based on secondary structure were the initial input to the DMD algorithm. The RNA was first subjected to a folding phase designed to allow base pairs and local helical structure to form. In this phase, the RNA was cooled through the following automated steps, where $T_i$ and $T_f$ are the initial and final reduced temperatures [in kcal/(mol × $k_B$)]: (1) $T_i$, $T_f$ = 30, for $10^5$ DMD time units (tu); (2) $T_i$ = 0.30, $T_f$ = 0.28, 2 × $10^4$ tu; (3) $T_i$, $T_f$ =0.28, $10^5$ tu; (4) $T_i$ =0. 28, $T_f$ = 0.26, 2 × $10^4$ tu; (5) $T_i$, $T_f$ =

0.26, $10^5$ tu; (6) $T_i$ = 0.26, $T_f$ = 0.24, 2 × $10^4$ tu; (7) $T_i$, $T_f$ = 0.24, $10^5$ tu; (8) $T_i$ = 0.24, $T_f$ = 0.22, 2 × $10^4$ tu; (9) $T_i$, $T_f$ = 0.22, $10^5$ tu; (10) $T_i$, $T_f$ = 0.22, 2 × $10^4$ tu. After step 10, constraints describing tertiary contacts were added. The RNA model was then cooled to a target reduced temperature through the following steps: (11) $T_i$, $T_f$ = 0.22, $10^5$ tu; (12) $T_i$ = 0.22, $T_f$ = 0.20, 2 × $10^4$ tu; (13) $T_i$, $T_f$ = 0.20, $10^5$ tu; (14) $T_i$ = 0.20, $T_f$ = 0.18, 2 × $10^4$ tu; (15) $T_i$, $T_f$ = 0.18, $10^5$ tu; (16) $T_i$ = 0.18, $T_f$ = 0.16, 2 × $10^4$ tu; (17) $T_i$, $T_f$ = 0.16, $10^5$ tu; (18) $T_i$ = 0.16, $T_f$ = 0.14, 2 × $10^4$ tu; (19) $T_i$, $T_f$ = 0.14, $10^5$ tu; (20) $T_i$ = 0.14, $T_f$ = 0.12, 2 × $10^4$ tu; (21) $T_i$, $T_f$ = 0.12, $10^5$ tu; (22) $T_i$ = 0.12, $T_f$ = 0.10, 2 × $10^4$ tu; (23) $T_i$, $T_f$ = 0.10, $10^5$ tu 100,000 structures are generated at this final refinement step. To select a representative structure for each refinement, structures from step 23 were subjected to hierarchical clustering as described.[10] Structures were first filtered on the basis of energy and simulation distance. Clustered structures were required to have an energy less than the median energy from step 23 and were required to be at least 1000 tu apart from other clustered structures (to prevent analysis of consecutive structures). 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 workstation (Intel Pentium 4 processor, 3.2 GHz) running Fedora Core 4. Refinement times ranged from 18 (CrPV, 49 nts) to 42 hrs (P546, 158 nts).

In some cases, the local structure that forms before incorporation of tertiary contact constraints restricts conformational space such that residues implicated in a tertiary interaction may not come into contact during refinement. In cases where imposed tertiary contacts were not present in the final structure (HHR, P546 and HCV-PK RNAs), the refinement was repeated with temporary strong constraints during step 10; these constraints simply function to promote initial collapse of the RNA molecule. For a given long-range pair-wise interaction, square well potentials were included between each pair of phosphate, sugar, and base pseudo-atoms, providing a graduated potential well that extends from an inter-residue distance of 10.0 to 155.0 Å. The maximum energy bonus for this constraint set was 20.0 kcal/mol when phosphate pseudo-atoms are within 85 Å of each other, 15.0 kcal/mol when sugar atoms are within 30 Å, and 10.0 kcal/mol when bases are within 10 Å.

**2.10.6 Fitting of cryo-EM models**

HCV-PK models were fit into cryo-EM models[37] manually using UCSF Chimera.[45]

### 2.10.7 Software used

RMSD calculations were performed using lsqman[46] and consider phosphate and phosphate pseudo-atom positions. RNA model images were composed using Pymol[47] with the exception of Figure 2.4A, which was generated using UCSF Chimera[45]. Clustering was performed using OC[48].

## 2.12 References


1. Gesteland, R. F., Cech, T., and Atkins, J. F. (2006) *The RNA World*, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

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

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

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

5. Ziehler, W. A., and Engelke, D. R. (2001) Probing RNA structure with chemical reagents and enzymes, *Curr Protoc Nucleic Acid Chem Chapter 6*, Unit 6 1.

6. Juzumiene, D.*, et al.* (2001) Short-range RNA-RNA crosslinking methods to determine rRNA structure and interactions, *Methods 25*, 333-343.

7. Gutell, R. R.*, et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods, *Nucleic Acids Res 20*, 5785-5795.

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

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

10. Gherghe, C. M.*, et al.* (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics, *J Am Chem Soc 131*, 2541-2546.

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

12. Westhof, E., Dumas, P., and Moras, D. (1988) Restrained Refinement of 2 Crystalline Forms of Yeast Aspartic-Acid and Phenylalanine Transfer-RNA Crystals, *Acta Crystal. A 44*, 112-123.

13. Cate, J. H.*, et al.* (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing, *Science 273*, 1678-1685.

14. Martick, M., and Scott, W. G. (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis, *Cell 126*, 309-320.

15. Costantino, D. A.*, et al.* (2008) tRNA-mRNA mimicry drives translation initiation from a viral IRES, *Nat Struct Mol Biol 15*, 57-64.

16. Jan, E., and Sarnow, P. (2002) Factorless ribosome assembly on the internal ribosome entry site of cricket paralysis virus, *J Mol Biol 324*, 889-902.

17. Kanamori, Y., and Nakashima, N. (2001) A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation, *RNA 7*, 266-274.

18. Canny, M. D., *et al.* (2004) Fast cleavage kinetics of a natural hammerhead ribozyme, *J Am Chem Soc 126*, 10848-10849.

19. Khvorova, A., *et al.* (2003) Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity, *Nat Struct Biol 10*, 708-712.

20. De la Pena, M., Gago, S., and Flores, R. (2003) Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity, *EMBO J 22*, 5561-5570.

21. Levitt, M. (1969) Detailed molecular model for transfer ribonucleic acid, *Nature 224*, 759-763.

22. Cannone, J., *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs., *BMC Bioinformatics 3*, e2.

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

24. Murphy, F. L., and Cech, T. R. (1994) GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain, *J Mol Biol 236*, 49-63.

25. Flor, P. J., Flanegan, J. B., and Cech, T. R. (1989) A conserved base pair within helix P4 of the Tetrahymena ribozyme helps to form the tertiary structure required for self-splicing, *EMBO J 8*, 3391-3399.

26. Honda, M., Brown, E. A., and Lemon, S. M. (1996) Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA, *RNA 2*, 955-968.

27. Wang, C., *et al.* (1995) An RNA pseudoknot is an essential structural element of the internal ribosome entry site located within the hepatitis C virus 5' noncoding region, *RNA 1*, 526-537.

28. Hajdin, C. E., *et al.* (2010) On the significance of an RNA tertiary structure prediction, *RNA 16*, 1340-1349.

29. Pisarev, A. V., Shirokikh, N. E., and Hellen, C. U. (2005) Translation initiation by factor-independent binding of eukaryotic ribosomes to internal ribosomal entry sites, *C R Biol 328*, 589-605.

30. Kieft, J. S. (2008) Viral IRES RNA structures and ribosome interactions, *Trends Biochem Sci 33*, 274-283.

31. Pfingsten, J. S., Costantino, D. A., and Kieft, J. S. (2006) Structural basis for ribosome recruitment and manipulation by a viral IRES RNA, *Science 314*, 1450-1454.

32. Lukavsky, P. J. (2008) Structure and function of HCV IRES domains, *Virus Res*.

33. Filbin, M. E., and Kieft, J. S. (2009) Toward a structural understanding of IRES RNA function, *Curr Opin Struct Biol 19*, 267-276.

34. Kieft, J. S., *et al.* (2001) Mechanism of ribosome recruitment by hepatitis C IRES RNA, *RNA 7*, 194-206.

35. Kieft, J. S., *et al.* (1999) The hepatitis C virus internal ribosome entry site adopts an ion-dependent tertiary fold, *J Mol Biol 292*, 513-529.

36. Spahn, C. M.*, et al.* (2001) Hepatitis C virus IRES RNA-induced changes in the conformation of the 40s ribosomal subunit, *Science 291*, 1959-1962.

37. Boehringer, D.*, et al.* (2005) Structure of the hepatitis C virus IRES bound to the human 80S ribosome: remodeling of the HCV IRES, *Structure 13*, 1695-1706.

38. Pestova, T. V.*, et al.* (1998) A prokaryotic-like mode of cytoplasmic eukaryotic ribosome binding to the initiation codon during internal translation initiation of hepatitis C and classical swine fever virus RNAs, *Genes Dev 12*, 67-83.

39. Otto, G. A., and Puglisi, J. D. (2004) The pathway of HCV IRES-mediated translation initiation, *Cell 119*, 369-380.

40. Fraser, C. S., and Doudna, J. A. (2007) Structural and mechanistic insights into hepatitis C viral translation initiation, *Nat Rev Microbiol 5*, 29-38.

41. Lyons, A. J., and Robertson, H. D. (2003) Detection of tRNA-like structure through RNase P cleavage of viral internal ribosome entry site RNAs near the AUG start triplet, *J Biol Chem 278*, 26844-26850.

42. Nadal, A.*, et al.* (2002) Specific cleavage of hepatitis C virus RNA genome by human RNase P, *J Biol Chem 277*, 30606-30613.

43. Hammond, J. A.*, et al.* (2009) Comparison and functional implications of the 3D architectures of viral tRNA-like structures, *RNA 15*, 294-307.

44. Brown, E. A.*, et al.* (1992) Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs, *Nucleic Acids Res 20*, 5041-5045.

45. Pettersen, E. F.*, et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput Chem 25*, 1605-1612.

46. Kleywegt, G. J. (1996) Use of non-crystallographic symmetry in protein structure refinement, *Acta Crystallogr D Biol Crystallogr 52*, 842-857.

47. Schrodinger, LLC. (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

48. Barton, G. (1993) OC - a cluster analysis program.

## CHAPTER 3: ACCURATE PREDICTION OF RNA TERTIARY FOLDS BY RING-MAP DIRECTED MOLECULAR MODELING

### 3.1 Introduction

Previously considered the passive intermediate in the transfer of genetic information from DNA to protein, RNA is now known to be active participant in a number of biological processes.[1] RNA actively regulates translation and transcription, controlling levels of gene expression.[2] Many RNA functions depend upon the formation of well-defined three-dimensional structures.[3] Investigation of this structure-function relationship is critical to the understanding of RNA and its place in biology.

RNA interaction group detection by mutational profiling (RING-MaP) is a recently developed method for analysis of RNA structure (Homan *et al.*, in preparation; Figure 3.1). RING-Map combines a number of chemical and enzymatic processes to reveal interactions within an RNA molecule. In this method, multiple structure-selective chemical modification events are detected on a single RNA molecule through next-generation sequencing approaches. By comparing chemical modifications across thousands of single molecules, nucleotide pairs with highly correlated structure-selective modifications are identified. These correlations imply that these nucleotides interact in the three-dimensional structure (Figure 3.2).

An attractive application of RING-MaP is to use these implicated interactions to bias molecular dynamics simulations in order to predict RNA structures.[4] RING-MaP describes a great variety of RNA interactions, however, and there is no clear way to distinguish interactions of great use to three-dimensional modeling, such as pair-wise tertiary contacts, from less useful interactions, such as long-distance interactions associated through coordinated folding. The challenge of modeling RNA using RING-MaP-determined RNA interactions is to create a method that accepts a wide-variety of structural information but still yields a precise RNA fold.

Here, we describe a technique for using RING-MaP-determined RNA interactions to bias RNA molecular dynamics simulations in order to produce precise and accurate three-dimensional RNA folds. The method utilizes free energy potentials dependent on the through-space distances between RING-

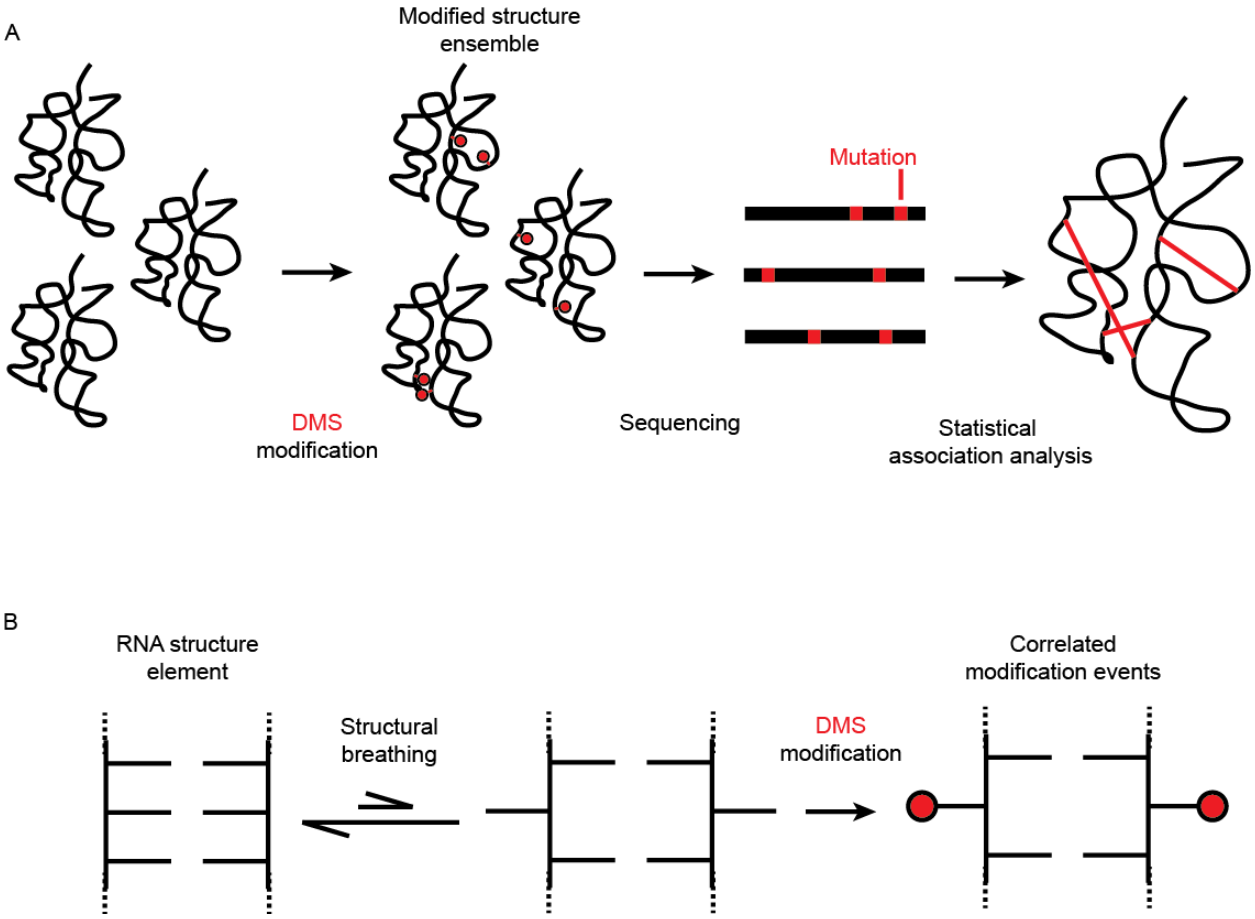



**Figure 3.1** Characterization of RNA by RING-MaP. (**A**) Schematic of detection of DMS modifications by mutational profiling. Following folding, RNA is modified by DMS, creating an ensemble of chemically-modified RNA molecules. Mutations are incorporated at modified positions during reverse transcription and are detected by NGS techniques. Statistical association analysis is then used to find correlated mutation events, implying interactions between nucleotides. (**B**) Structural breathing results in coordinated structure-specific DMS modification.

MaP-determined nucleotide pairs. When such pairs are close, an energy bonus is applied such that RING-MaP-determined pairs remain in contact when sampled. The quality of RNA structures is comparable to structures predicted using constraints identified by other structural characterization techniques such as small-angle X-ray scattering.[5]

**3.2 RING-MaP: statistical association of single-molecule chemical modification**

Chemical modification is a commonly used means of characterizing RNA structure.[6] In chemical modification experiments, a structure-selective chemical is allowed to react with a natively folded RNA. By assaying positions of modification, structural inferences may be made about the RNA studied. For instance, dimethyl sulfate (DMS) preferentially modifies adenosine and cytosine bases at unstructured RNA nucleotides.[7] Adenosine and cytidine residues that are not modified by DMS are inferred to participate in structural interactions.

Recently, approaches have been developed that allow RNA chemical modification experiments to be resolved by sequencing.[8] One such approach is mutational profiling. Mutational profiling exploits a recently described phenomenon in which reverse transcriptase incorporates a mutation at the cDNA position corresponding to the modified residue (Siegfried et al., in preparation). Through mutational profiling, positions of chemical modification may be determined by sequencing.

Chemical modification by DMS was measured using next generation sequencing (Figure 3.1A). In next-generation sequencing approaches, $>10^6$ sequences are rapidly characterized in parallel.[9] Each sequence analyzed is derived from a single DNA molecule in the sequencing library. When a solution-based ensemble of DMS-modified RNA molecules is analyzed by next generation sequencing, an array of single-molecule DMS experiments is effectively performed in parallel.

Rates of mutation were analyzed by statistical association analysis. To identify correlated modifications, all reads where a selected position in the RNA was mutated were selected, and the mutation rates for each other position were determined for those reads. These mutation rates were then compared to mutation rates across all reads. If a position was found to be mutated at a significantly greater or lesser rate when another position was modified, those two positions were considered to be correlated. The significance of this correlation was determined by chi-square analysis.

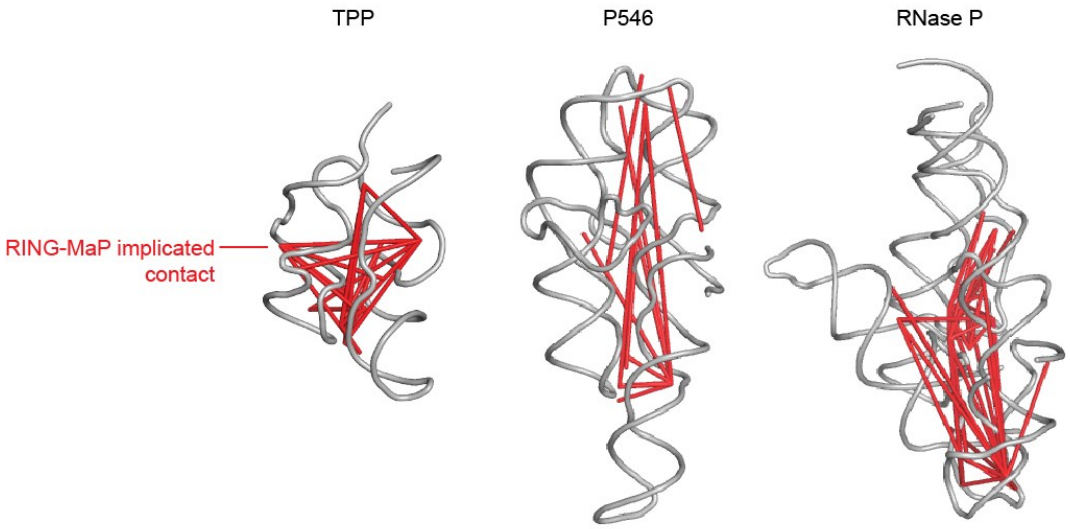



**Figure 3.2** Nucleotide pairs identified by RING-MaP displayed on high-resolution structures. Red lines are drawn between nucleotide pairs implicated in RNA interactions.

Pairs of nucleotides with correlated DMS modifications likely participate in common intramolecular interactions. These correlations are thought to reflect coordinated structural breathing that increases the chance of associated modifications at interacting RNA nucleotides (Figure 3.2B). The collection of interactions found by this analysis is defined as an RNA interaction group, or RING.

**3.3 Characterization of RING-MaP RNA interactions**

In order to incorporate RING-MaP interactions into molecular dynamics simulations, the interacting pairs must be related to some structural metric such as through-space distance between correlated residues. Therefore, the first step in incorporating RING-MaP data was structural characterization of RING-MaP interaction networks. In previous work, RING-MaP RNA interactions were determined for three RNA molecules: the thymine pyrophosphate-dependent RNA riboswitch (TPP), the P5-P4-P6 domain of the *Tetrahymena thermophila* Group I intron (P546), and the catalytic domain of *Bacillus stearothermophilus* ribonuclease P (RNase P) (Homan *et al.*, in preparation). High-resolution structures derived from X-ray crystallography are available for each of these RNAs.[9-11] Through-space distance distributions for correlated nucleotide pairs with absolute correlations greater than 0.025 were determined for each RNA based on corresponding high-resolution structures (Figure 3.3A).

Pair-wise through-space distances for identified interactions were found to follow a normal distribution (Figure 3.3B). The average pair-wise through-space distance for RING-MaP pairs was 23.2 Å. The variability in through-space distances is reflected in a standard deviation of 13.0 Å. Despite this variability, the average distance considering the three RNAs individually was fairly consistent. TPP, P546, and RNase P have average through-space distances of 21.0 Å, 25.1 Å, and 23.5 Å, respectively.

**3.4 Incorporation of RING-MaP information into molecular dynamics simulations**

Energy potentials used to restrain molecular dynamics simulations were based upon pair-wise distance statistics. For a given RING-MaP nucleotide pair, if the distance between the two constituent nucleotides was less than 36 Å (the sum of the average and standard deviation distances) or 23 Å (the average distance), an energy bonus of -0.3 kcal/mol or -0.6 kcal/mol, respectively, was applied (Figure 3.3C). The maximum energy bonus of -0.6 kcal/mol is equal to the energy potential applied for a single nucleotide stack in the modeling engine. Molecular dynamics simulations were also constrained by RNA secondary structure. The secondary structure for each RNA was taken from the corresponding crystal

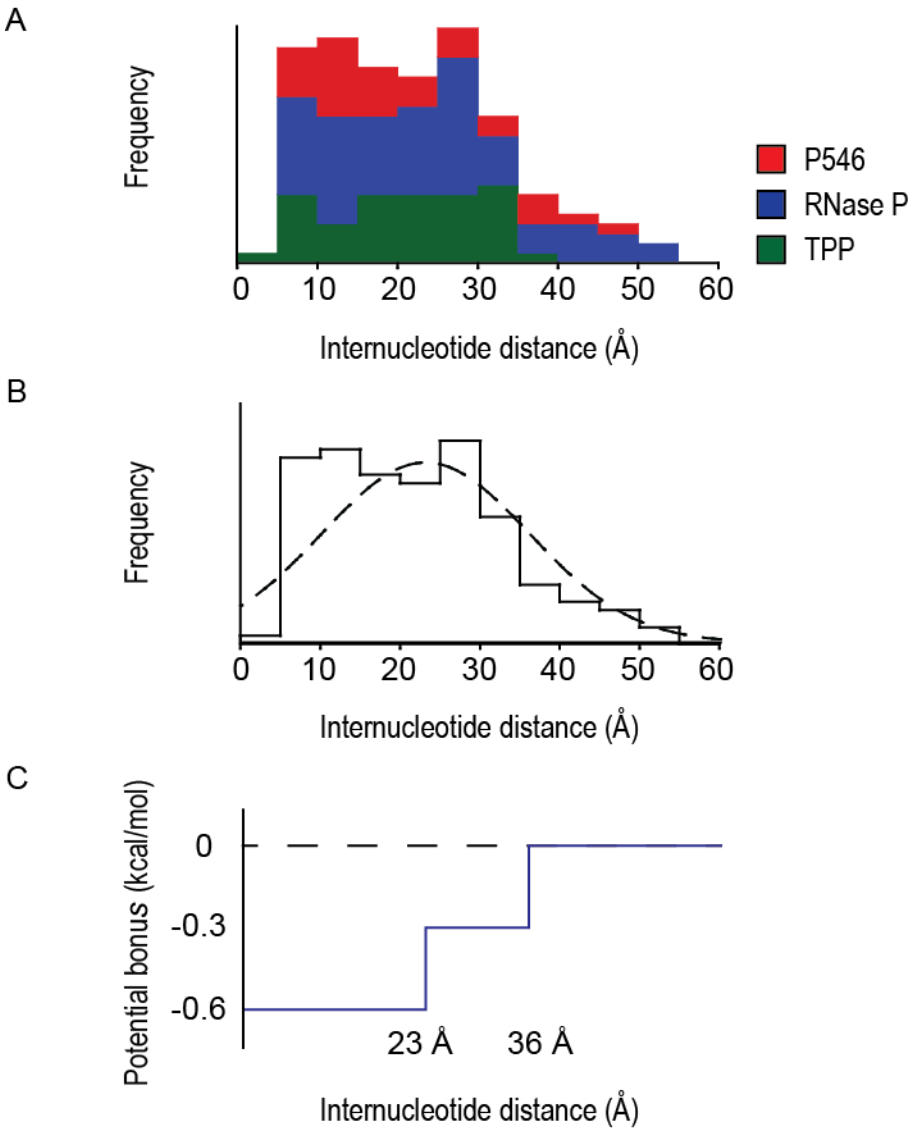



**Figure 3.3** Characterization of through-space distances in high-resolution structures for RING-MaP pairs and subsequent energy potential design. (**A**) Stacked column histogram of C1'-C1' through-space distances for TPP (green), P546 (red), and RNase P (blue). (**B**) Histogram considering C1'-C1' distances for all three RNAs. The described normal distribution is shown with a dotted line. (**C**) Energy potential well used to bias molecular dynamics simulations. Internucleotide distances of 23 Å corresponds to the average distance, and 36 Å corresponds to the sum of the average and standard deviation distances.

structure. Secondary structure may be accurately predicted by comparative analysis or chemical modification-directed algorithms,[12, 13] and constraint by a known secondary structure is common in knowledge-directed modeling approaches.[14]

Energy potentials derived from RING-MaP analysis were incorporated into RNA discrete molecular dynamics (RNA DMD) simulations.[15] RNA DMD is distinguished from other molecular dynamics approaches by two characteristics. First, an RNA molecule is modeled using a reduced atom representation. Each RNA nucleotide is represented using only three pseudo-atoms that correspond to the sugar, base, and phosphate groups. Second, force field interactions are approximated using square-well potentials based on experimentally-derived values. These two approximations allow for efficient and robust sampling of RNA conformational space.

RNA DMD simulations were performed using replica exchange; that is, multiple simulations were run in parallel. Over the course of the simulations, the temperature factor that governs energy transfer during simulations was varied. In RING-MaP-biased modeling, a total of eight replicas were run in parallel, with each replica run for 1,000,000 time units. The temperature factors used were 0.1000, 0.1375, 0.1750, 0.2125, 0.2500, 0.2875, 0.3250, and 0.3625, representing a broad range of temperature factors commonly used in RNA DMD refinement. Prior to initiating simulations, energy potentials based on RING-MaP analysis and constraints based on known secondary structures were incorporated. Simulations were performed for all RING-Map characterized RNAs.

**3.5 Filtering generated structures by radius of gyration**

Following molecular dynamics simulation, structural snapshots were generated at every 1,000 time units for each replica. Structures generated from RING-MaP-biased simulations were compared with structures generated using no constraints. In comparisons of radius of gyration values for biased and unbiased simulations, a bias-dependent collapsed state was apparent (Figure 3.4). In order to ensure that predicted structures reflected this collapsed state, structures were filtered on the basis of radius of gyration.

To do this, radius of gyration histograms were created for both biased simulations and unbiased controls. The unbiased histogram was scaled such that its difference from the biased histogram was minimized. A difference histogram was then generated, yielding a histogram describing the RING pair-

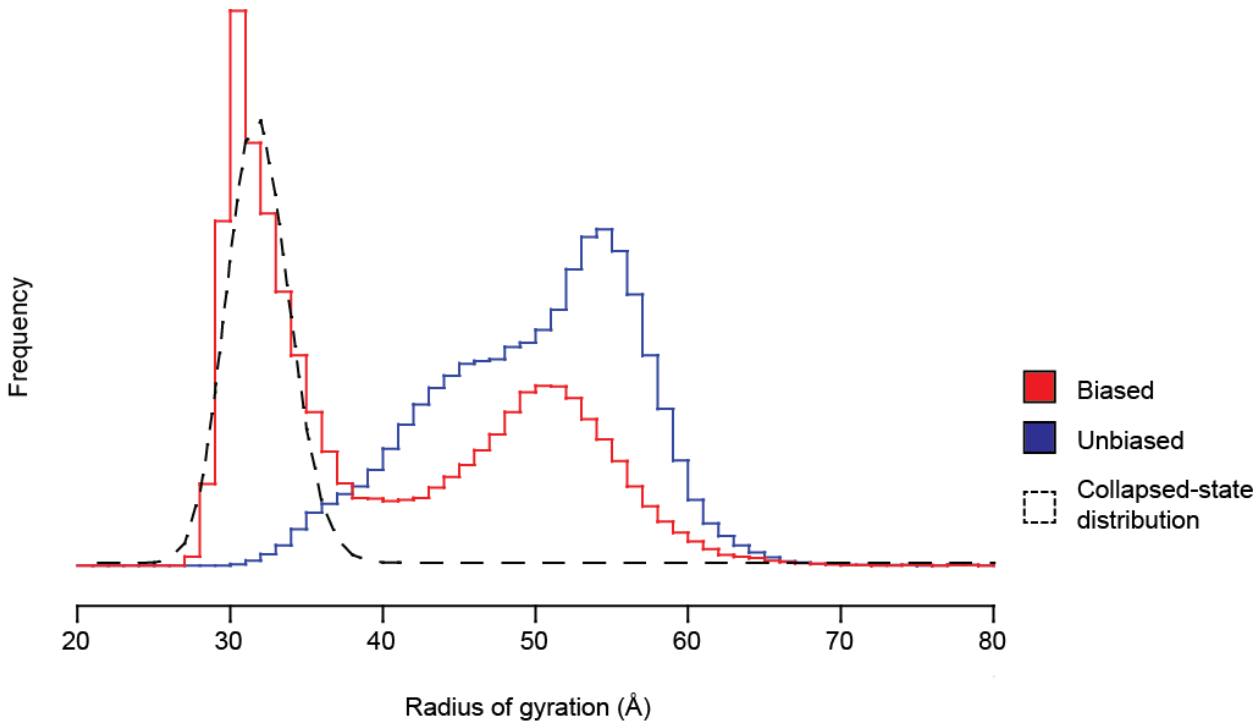



**Figure 3.4** Radius of gyration distributions of biased (red) and unbiased (blue) simulations of P546. The derived log-normal distribution for the bias-dependent collapsed state is shown with a dotted line.

dependent collapsed state. A log-normal distribution was then fit to this difference in order to characterize the collapsed state (Figure 3.4). Only structures with radii of gyration within one geometric standard deviation of the geometric mean were considered in further analyses. Radius of gyration filtering resulted in an ensemble with higher predictive accuracy as measured by root mean square deviation (RMSD) from accepted structures (Figure 3.6).

**3.6 Selection of a final predicted structure by hierarchical clustering**

Hierarchical clustering was used to characterize the structural ensemble and generate a final predicted structure.[16] For those RNA models selected by radius of gyration, the 250 models with the lowest free energies were analyzed by clustering (structural ensembles with 10 representative models are shown for each RNA in Figure 3.5). Clustering was performed based on RMSD values between models in the structure ensemble. Clustering was performed such that the greatest RMSD value allowed in a given cluster was less than the sum of the predicted average and standard deviation RMSDs.[17]

For each RNA, clustering resulted in only a few clusters or a single cluster. For P546, all structures were within a single cluster. TPP and RNase P each had a prominent primary cluster. In analyses of 250-structures ensembles, the most populated cluster for TPP contained 249 structures and that for RNase P contained 210 structures. The clustering results show that RING-biased molecular dynamics simulations are convergent on relatively well-determined RNA folds, indicating a high level of precision.

From each simulation, the centroid of the most populated cluster was taken to be the final predicted structure (Figure 3.5). To evaluate the final predicted models, RMSD values were calculated using the phosphate backbones of the predicted models and accepted high-resolution structures. For TPP, P546, and RNase P, the predicted structures had RMSDs of 9.6 Å, 17.1 Å, and 22.4 Å, respectively. TPP, P546, and RNase P predictions had *p*-values of 0.002, $1.1 \times 10^{-5}$, and $1.4 \times 10^{-5}$, respectively, indicating that the predictions have a high degree of statistical significance.[17]

Of the many contacts found in RING-MaP analysis of RNase P, few were found between nucleotides making up the P3-P2-P19 helical stack and the rest of the RNA. Based on this observation, the area excluding the P3-P2-P19 helical stack was determined by RING analysis to be the structural

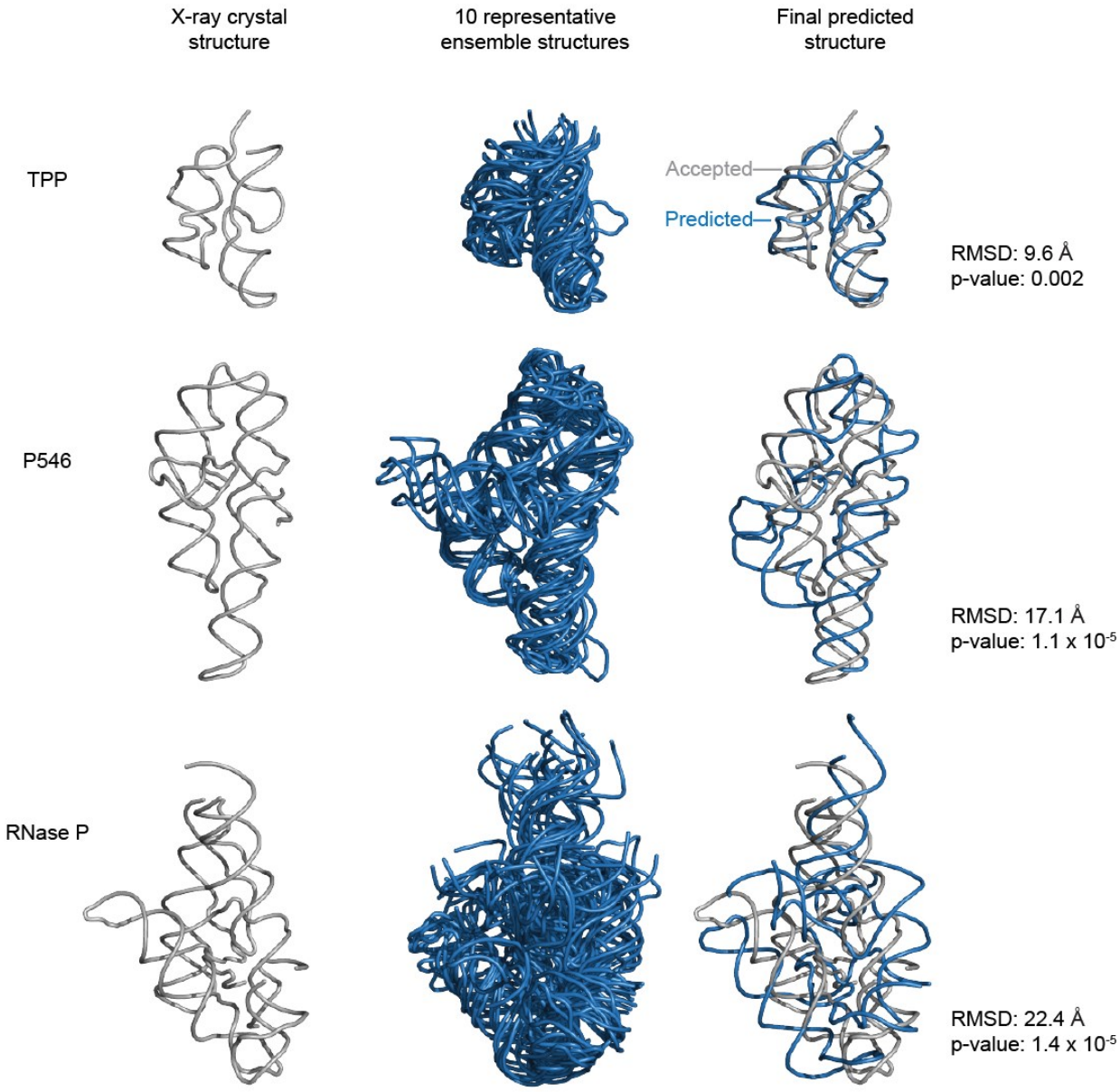



**Figure 3.5** Reference models, limited structure ensembles, and final predicted structures for TPP, P546, and RNase P. Ten representative models (blue) for each RNA molecule were found by hierarchical clustering of RING MaP-biased simulations. The final predicted structure (blue) for each RNA is shown aligned to the reference X-ray crystal structure (grey).

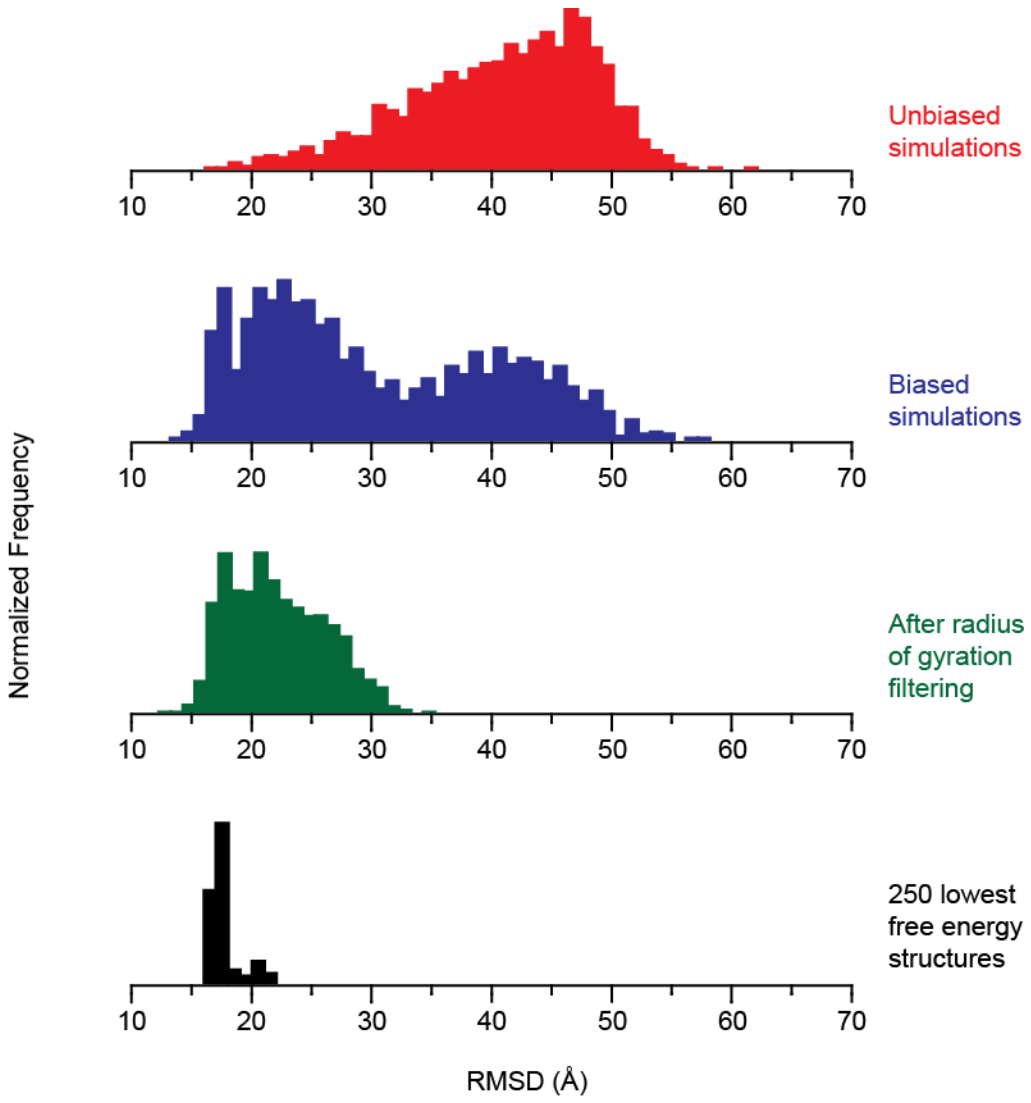



**Figure 3.6** Sequential improvements in RMSDs during refinement. RMSD histograms were calculated for representative structures for both biased (red) and unbiased (blue) simulations. RMSD histograms were also calculated for representative structures after filtering by radius of gyration (green) and after selection by lowest free energy (black).

core of the molecule. Within this RING-determined structural core, the predicted structure has an RMSD of 14.4 Å with a *p*-value less than $1.0 \times 10^{-6}$ (Figure 3.7).

**3.7 Discussion**

With RMSD values ranging from 9.6 to 22.4 Å for RNA molecules as large as 268 nucleotides, RING-MaP-directed modeling successfully recapitulated known RNA folds (Figure 3.5). RING-MaP-directed modeling correctly modeled fold-defining tertiary motifs, such as the bent-hinge motif of P546 that allows for two-helix packing. The degree of accuracy of RING-MaP-directed modeling was comparable to other low-resolution structural determination techniques such as modeling based on small angle X-ray scattering data.[5]

A distinct advantage of RING MaP-dependent modeling is that RING-MaP analysis may be performed under various conditions, allowing generation of solution-specific structures. This is most readily shown by the success modeling TPP, a riboswitch that binds thiamine pyrophosphate. The high-resolution structure of TPP was determined in the ligand-bound conformation. Our solution chemical probing of TPP RNA was performed in the presence of ligand, and RING-directed modeling successfully recapitulated the ligand-bound fold (Figure 3.5). In future experiments, RING-directed modeling may be used to discover and define structural changes in an RNA molecule due to external factors such as the presence of RNA-binding proteins or ligands.

The methods developed in this work may also be used with other RNA characterization techniques. As a whole, this refinement strategy incorporates pair-wise data on the basis of statistical guidelines, allowing incorporation of data from other sources, such as crosslinking studies or fluorescence resonance energy transfer (FRET) analysis. Each step in the methodology may see application individually, as each leads to greater model accuracy as measured by RMSD (Figure 3.6).

**3.8 Conclusion**

RING MaP-dependent modeling results in accurate structure prediction for RNAs ranging in length from 80 to 268 nucleotides, with TPP, P546, and RNase P predictions having RMSDs relative to high-resolution structures of 9.6 Å, 17.1 Å, and 22.4 Å, respectively. As application of RING-MaP is generic, characterization by RING-MaP and subsequent structure prediction may be applied to other

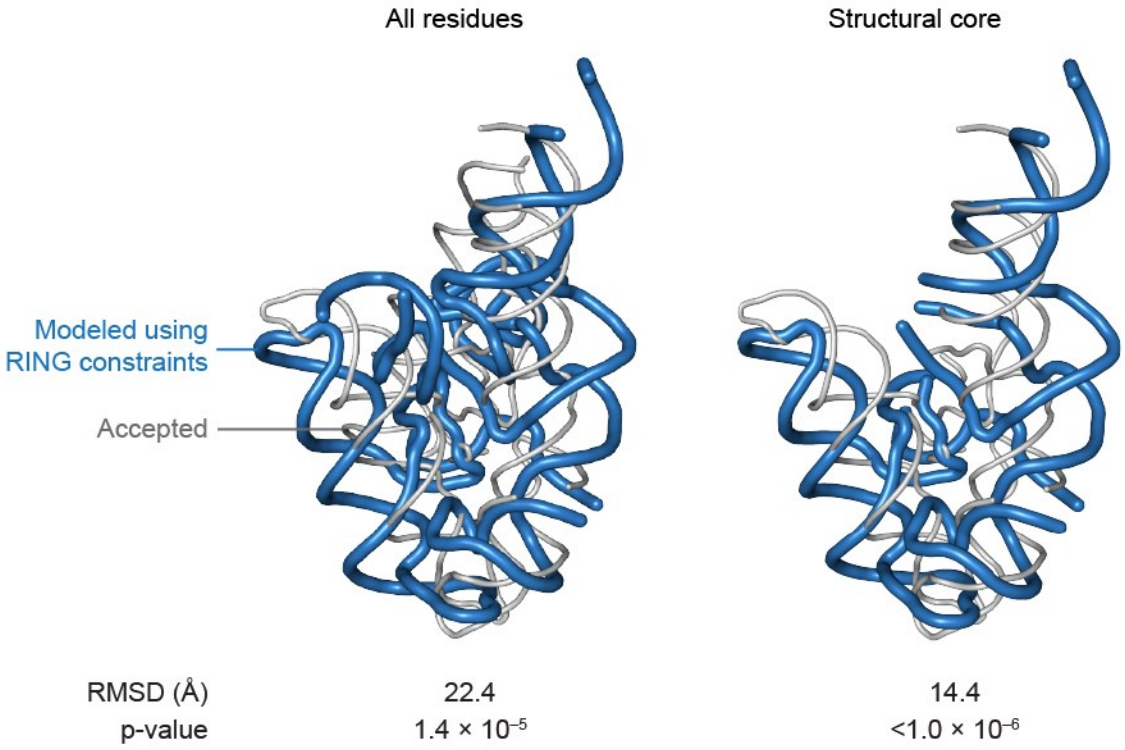



**Figure 3.7** RING-MaP-determined structural core of RNase P. The quality of alignment of the predicted model (blue) to the reference high-resolution structure (grey) for the whole model is shown on the left; alignment based only on the structural core is shown on the right.

biologically relevant RNA molecules. The modeling methodologies developed here may be applied to other RNA structure characterization techniques.

### 3.9 Experimental methods

#### 3.9.1 Selection of pairwise contacts implicated by RING-MaP

RING-MaP nucleotide pairs were selected on the basis of absolute correlation coefficient. Only nucleotide pairs with an absolute correlation coefficient greater than 0.025 were considered in analysis. RING-MaP nucleotide pairs were excluded if the positions of the two constituent nucleotides were less than or equal to 11 nucleotides apart. RING-MaP pairs also excluded if they were considered already implicated in a secondary structure contact: Considering positions $n_i$ and $n_j$ in a RING-MaP nucleotide pair and positions $m_i$ and $m_j$ in any given base pair in constraining secondary structure, if $|n_i - m_i| + |n_j - m_j| <= 11$, the nucleotide pair was considered implicated by secondary structure contacts.

#### 3.9.2 Energy potential system used for implicated contacts

A free energy bonus *E* was applied based on the distance *d* between the two constituent sugar pseudo-atoms in a RING-MaP nucleotide pair (Figure 3.3C):

$$E = \begin{cases} -0.6\,kcal/mol\,, 0 \leq d < 23.0\ \text{Å} \\ -0.3\,kcal/mol\,, 23.0\ \text{Å} \leq d < 36.0\ \text{Å} \\ 0, 36.0\ \text{Å}\ \leq d \end{cases}$$

#### 3.9.3 Replica exchange molecular dynamics simulations

Molecular dynamics simulations were performed using the RNA DMD engine.[15] Simulations were performed using replica exchange with eight replicas run in parallel, with each being run for 1,000,000 time units. The eight replicas were run with temperature factors of 0.1000, 0.1375, 0.1750, 0.2125, 0.2500, 0.2875, 0.3250, and 0.3625, where the temperature factors used described heat exchange by the Andersen thermostat. An unbiased control simulation with identical parameters was run in parallel.

#### 3.9.4 Filtering by radius of gyration

Structures were generated at every 1000 time units. The ensemble of structures for the biased simulation was compared to the unbiased control in order to determine a constraint-dependent collapsed state. Radius of gyration histograms were created for both the biased and unbiased simulations (Figure 3.4). The unbiased histogram was then normalized to the biased histogram such that the difference between the two was minimized by least squares. The normalized unbiased histogram was then

subtracted from the biased histogram. A log-normal distribution was then fit to this difference using least squares, giving a distribution describing the collapsed state. The log-normal is described by the following probability distribution function, where $x$ is a given radius of gyration, $\sigma$ is the location parameter, and $\mu$ is the scale parameter:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2 / 2\sigma^2}$$

The geometric mean was taken as $e^{\mu}$, and the geometric standard deviation was taken as $e^{\sigma}$. The structural ensemble was filtered such that each structure taken forward in the analysis had a radius of gyration within one geometric standard deviation of the geometric mean.

**3.9.5 Hierarchical clustering and final structure selection**

From the ensemble of structures selected by radius of gyration, the 250 structures with the lowest free energy were analyzed by hierarchical clustering based on pair-wise RMSD.[16] In previous work, Hajdin et al. showed that the RMSDs relative to a known high-resolution structure for molecular dynamics simulation-generated RNA decoys follow a length-dependence based on that RNA's chain length.[17] They were able to develop an equation that relates RNA chain length to a predicted mean RMSD value. Clustering was performed such that within each cluster the greatest RMSD value between any given pair was less than the sum of the standard deviation and mean RMSD values predicted for a given RNA based on chain length.

The final predicted structure was taken as the centroid of the most populated cluster. To find ten structures representative of the most populated cluster, the constituent structures of the most populated cluster were themselves clustered with the number of clusters constrained to be ten. The centroids of those ten clusters were considered as a limited representation of the whole structure ensemble.

**3.9.6 Software used**

Least-squares fitting and statistical analysis were performed using the SciPy and NumPy modules of Python.[18] Clustering analysis was performed using OC.[19] Images of three-dimensional RNA structures were generated using Pymol.[20]

### 3.10 References


1. Gesteland, R. F., Cech, T., and Atkins, J. F. (2006) *The RNA World*, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

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

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

4. Lavender, C. A.*, et al.* (2010) Robust and generic RNA modeling using inferred constraints: a structure for the hepatitis C virus IRES pseudoknot domain, *Biochemistry 49*, 4931-4933.

5. Yang, S.*, et al.* (2010) RNA structure determination using SAXS data, *J Phys Chem B 114*, 10039-10048.

6. Weeks, K. M. (2010) Advances in RNA structure analysis by chemical probing, *Curr Opin Struct Biol 20*, 295-304.

7. Peattie, D. A., and Gilbert, W. (1980) Chemical probes for higher-order structure in RNA, *Proc Natl Acad Sci U S A 77*, 4679-4682.

8. Lucks, J. B.*, et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), *Proc Natl Acad Sci U S A 108*, 11063-11068.

9. Serganov, A.*, et al.* (2006) Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch, *Nature 441*, 1167-1171.

10. Cate, J. H.*, et al.* (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing, *Science 273*, 1678-1685.

11. Kazantsev, A. V., Krivenko, A. A., and Pace, N. R. (2009) Mapping metal-binding sites in the catalytic domain of bacterial RNase P RNA, *RNA 15*, 266-276.

12. Gutell, R. R.*, et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods, *Nucleic Acids Res 20*, 5785-5795.

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

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

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

16. Gherghe, C. M.*, et al.* (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics, *J Am Chem Soc 131*, 2541-2546.

17. Hajdin, C. E.*, et al.* (2010) On the significance of an RNA tertiary structure prediction, *RNA 16*, 1340-1349.

18. Oliphant, T. E. (2007) Python for scientific computing, *Computing in Science & Engineering 9*, 10-20.

19. Barton, G. (1993) OC - a cluster analysis program.

20. Schrodinger, LLC. (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

# CHAPTER 4: MODEL-FREE RNA STRUCTURE ALIGNMENT INCORPORATING CHEMICAL PROBING DATA

## 4.1 Introduction

In recent decades, RNA has been found to be an active participant in gene expression and regulation. As discoveries related to the function of RNA were being made, the fundamental understanding of the human genome was changing. Initially, most of the human genome was believed to be translated into protein. In 1964, it was estimated that the human genome contained 6.7 million genes based on the assumption that nearly the entire genome encoded for protein.[1] With publication of the human genome sequence in 2001, the number of predicted genes dropped to between 30,000 and 40,000.[2] In this same report, it was noted that a majority of the human genome was transcribed into RNA although many regions do not appear to encode proteins. Recent estimates suggest that only 2.94% of the human genome is translated into protein, but 62.1% and 74.7% of the genome is transcribed into processed and primary transcripts, respectively.[3, 4]

A current challenge is the characterization and categorization of the RNA elements transcribed across the human genome. For a vast majority of RNA transcripts, no structures or functions are known. To address this problem, a number of sequence comparison techniques are being used to find related RNA elements and to annotate the transcriptome. A number of these comparison approaches consider secondary structures of RNA molecules. Structure-informed approaches such as Infernal and Foldalign perform alignments based on both sequence and local secondary structure.[5, 6] These approaches are necessarily limited by our current understanding of RNA structure. Though great strides have been made in *ab initio* prediction of RNA secondary structures, the best current approaches only correctly predict 40-70% of known base pairs.[7] Secondary structure prediction approaches also do not consider RNA tertiary structures like pseudoknots. Moreover, optimization and benchmarking of these approaches are confined to known RNA structure motifs, themselves limited to structures that have been studied by high-resolution structure characterization techniques such as X-ray crystallography and nuclear magnetic resonance

(NMR) spectroscopy. RNA structure prediction techniques are therefore likely biased by a small number of well-characterized elements.

Sequence comparison considering structure-dependent chemical modification is attractive alternative to comparisons considering *ab initio* or concurrent structure prediction. RNA chemical modification is robust and is not limited by the current understanding of RNA structure. Current chemical modification approaches, such as SHAPE, are not limited to specific bases and allow characterization of virtually all nucleotides of any RNA target. The adaptation of RNA chemical modification approaches to next-generation sequencing (NGS) platforms allows for high-throughput analysis with approaches rapidly advancing towards transcriptome-scale assays.[8]

In this work, a sequence comparison approach that considers chemical modification data is introduced and evaluated. We found that a SHAPE-dependent alignment that was blind to base identity had comparable accuracy to traditional sequence comparison techniques. Approaches that consider both SHAPE modification and sequence identity showed improved accuracy relative to approaches considering only base identity. Chemical modifications were compared using a simple pair-wise scoring function that may be applied to a variety of current and future sequence comparison techniques.

**4.2 Selection of test-case RNA molecules and subsequent data generation**

In order to develop a method for RNA sequence comparison by chemical modification, target RNAs for chemical modification must first be selected. For development and evaluation of SHAPE-dependent RNA structure alignment, ribosomal RNA was used. Ribosomal RNA was selected based on two criteria. First, ribosomal RNA has been extensively characterized. The majority of the nucleotides described by high-resolution structures in the RCSB Protein Data Bank are found in ribosomal RNAs; currently 83% of nucleotides in RNA structures belong to rRNA. Thousands of ribosomal RNA sequences have been curated and aligned in the Comparative RNA Web site (CRW), with secondary and tertiary structures predicted based on covariation analysis.[9] Second, ribosomal RNA contains a variety of secondary and tertiary RNA structure motifs. Ribosomal RNA has been a primary source for RNA motifs for knowledge-based structure prediction methods such as the MC-Fold | MC-Sym pipeline.[10]

Three ribosomal RNA samples were considered during development of this approach. Ribosomal RNA samples were taken from cell cultures of *Escherichia coli*, *Clostridium difficile*, and

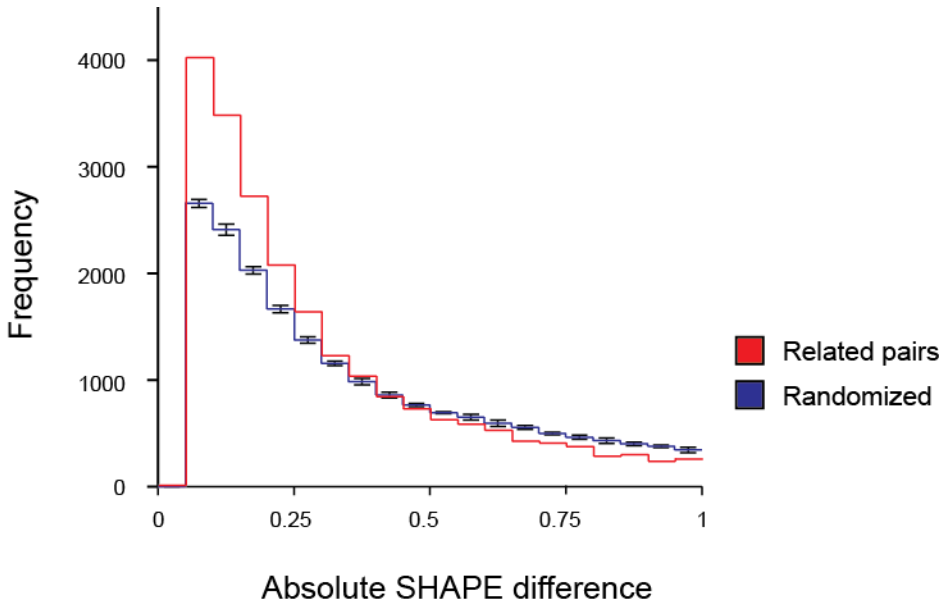



**Figure 4.1** Histogram of the absolute differences in SHAPE reactivities for associated nucleotides in CRW alignments. Differences between related pairs are shown in red, and differences between pairs in a randomized control are shown in blue. Pairs were randomized over eight individual trials; average values are shown and standard deviation values are included as error bars.

*Haloferax volcanii.* These ribosomal RNA samples represent a diverse test set; compared to *E. coli*, *C. difficile* and *H. volcanii* have percent nucleotide identities of 72.6% and 59.7%, respectively. The three ribosomal samples were analyzed using SHAPE chemistry, in which structural flexibility at a given position is determined by the extent of modification by chemical probe.[11] In the immediate future, large-scale RNA chemical modification experiments will almost certainly be detected using NGS approaches. To accommodate this, all data for the ribosomal samples were generated using SHAPE-MaP, a new approach for determining SHAPE reactivity values using NGS platforms (Siegfried *et al*, in preparation). Ribosomal RNA data for *E. coli* were obtained in previous experiments (Siegfried *et al*, in preparation), and data for *C. difficile* and *H. volcanii* were determined as part of this work.

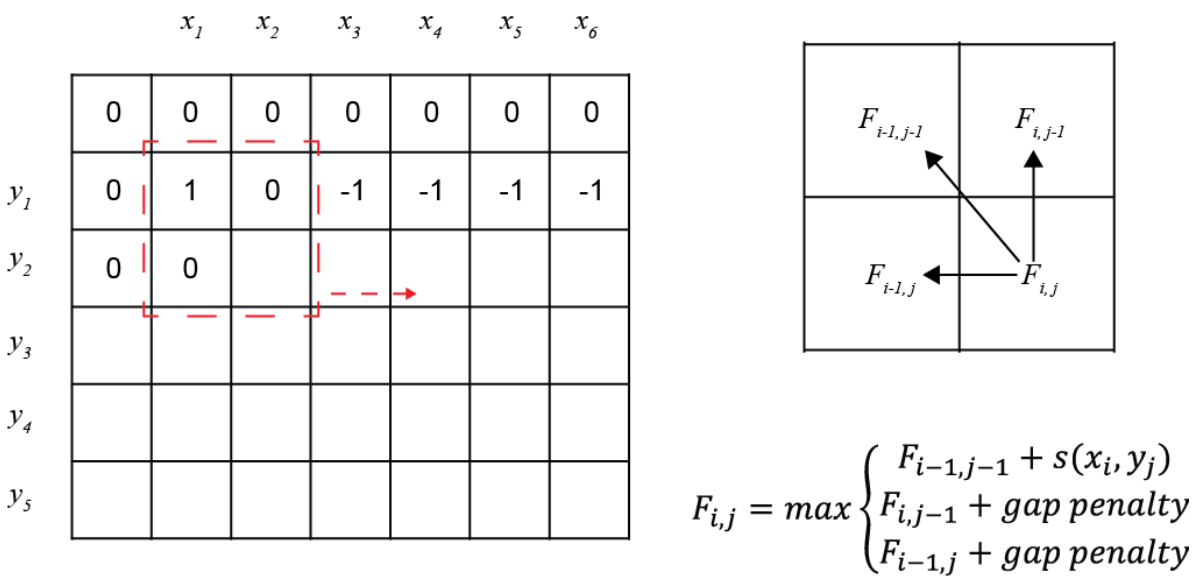
**4.3 Comparison of SHAPE data for related RNA sequences**

Following characterization of ribosomal RNA samples, SHAPE data were compared for related RNA nucleotides. Related nucleotides were taken from annotated sequence comparisons from the Comparative RNA Project web site.[9] The absolute differences in SHAPE reactivities were found for each related nucleotide pair. These differences were then plotted as a histogram (Figure 4.1, in red). The distribution of the SHAPE reactivity differences in related RNA nucleotides follows an exponential decay distribution. In order to gauge the significance of the association of SHAPE reactivity values, SHAPE data were randomly resorted, and absolute differences were calculated for randomly assigned pairs (Figure 4.1, in blue). The distributions indicate that the difference in SHAPE reactivities between two related nucleotides is likely to be smaller than the difference between two unrelated nucleotides. Based on Student's t-test, these two distributions were found to be significantly different from one another with a $p$ value $< 10^{-6}$.

**4.4 SHAPE-based scoring function and alignment approach**

Global SHAPE-dependent sequence comparisons were made using a pair-wise dynamic programming algorithm illustrated schematically in Figure 4.2.[12] The algorithm utilizes recursion to optimally align two sequences based on a pair-wise scoring function between individual nucleotides. The algorithm also incorporates penalties based on gap openings and gap extensions, where gaps are unaligned regions of sequence.

A




B

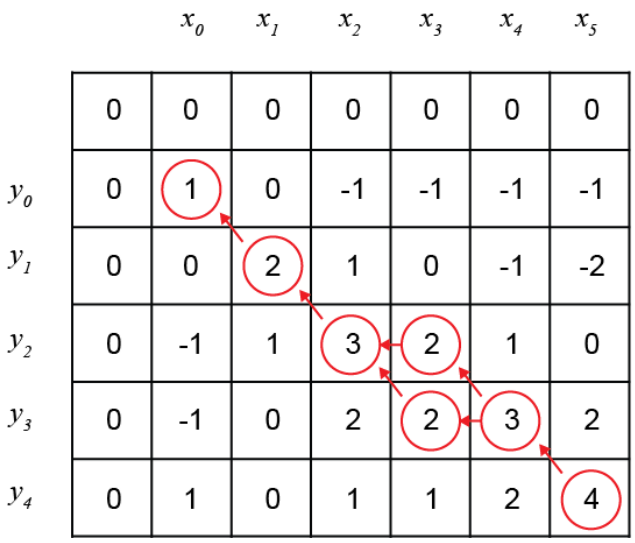



**Figure 4.2** Schematic of a general dynamic programming approach. (**A**) First, a comparison matrix *F* for sequences *x* and *y* is generated and subsequently filled using a scoring function. Scoring proceeds across a given row until that row ends, and then scoring proceeds on the next row. The scoring function considers previously determined scores in the matrix. (**B**) Following matrix generation, a trace-back function follows the highest scores in the comparison matrix, generating the optimal global alignment.

$$Score(x_i, y_j) = N_0 e^{-\lambda |x_i - y_j|} + b$$

$$where \quad N_0 = 4$$
$$\lambda = 1$$
$$b = -1$$

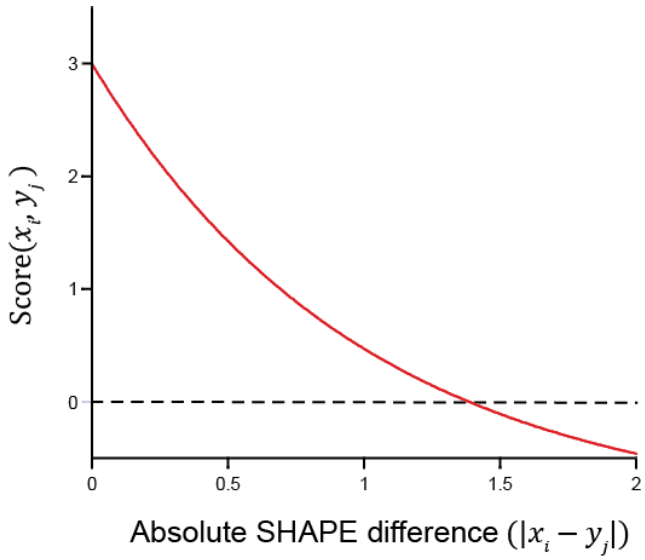



**Figure 4.3** Scoring function used to compare SHAPE values at positions $i$ and $j$ in sequences $x$ and $y$, respectively.

Pair-wise SHAPE comparisons were scored by an exponential decay function (Figure 4.3). This function is described by the following equation, where $i$ and $j$ are SHAPE values for any two given nucleotides and $N_0$, $\lambda$, and $b$ are parameters for the exponential decay function:

$$Score_{i,j} = N_0 e^{-\lambda|i-j|} + b$$

Parameters $N_0$, $\lambda$, and $b$, as well as gap opening and extension penalties, were optimized by grid searches for 16S and 23S ribosomal RNA. The resulting alignment was compared to an accepted pairwise alignment from the CRW. Parameters were selected based on the average sensitivities for pairwise alignments of *E. coli* RNA to *C. difficile* and to *H. volcanii* RNA.

**4.5 Quality of SHAPE-based alignments**

SHAPE-based alignments were performed for all 16S and 23S ribosomal RNA pairs (representative region of alignment shown in Figure 4.4). In order to evaluate the accuracy of the alignment algorithm with optimized parameters, qualities of alignments were compared to global sequence alignments performed using the Needle algorithm on the EMBOSS server using default parameters. Both SHAPE-based and Needle alignments were evaluated by computing the sensitivity relative to accepted CRW alignments.

SHAPE-based alignments were comparable in quality to Needle alignments (Table 4.1). For 16S rRNA, SHAPE-based alignments have sensitivities of 84% and 77% for alignments of *E. coli* to *C. difficile* and to *H. volcanii*, respectively. Alignments with Needle have sensitivities of 84% and 72%, respectively. For 23S rRNA, SHAPE-based alignments have sensitivities of 80% and 43% for alignments to *C. difficile* and to *H. volcanii*, respectively, whereas alignments with Needle have sensitivities of 83% and 58%.

**4.6 Incorporating a base-identity match score into SHAPE-based alignments**

An additional scoring term considering base identity was included in the SHAPE-alignment algorithm. If in a pair-wise comparison the two bases of a given nucleotide pair were identical, that pair was scored as a match; otherwise the pair was scored as a mismatch. The score terms associated with both matches and mismatches were optimized by a grid search. Gap opening and gap extension penalties were re-optimized given this new scoring system.

Alignments considering both base identity and SHAPE data showed significant improvements relative to alignments considering only SHAPE data (Table 4.1). For 16S rRNA, alignments of *E. coli* to

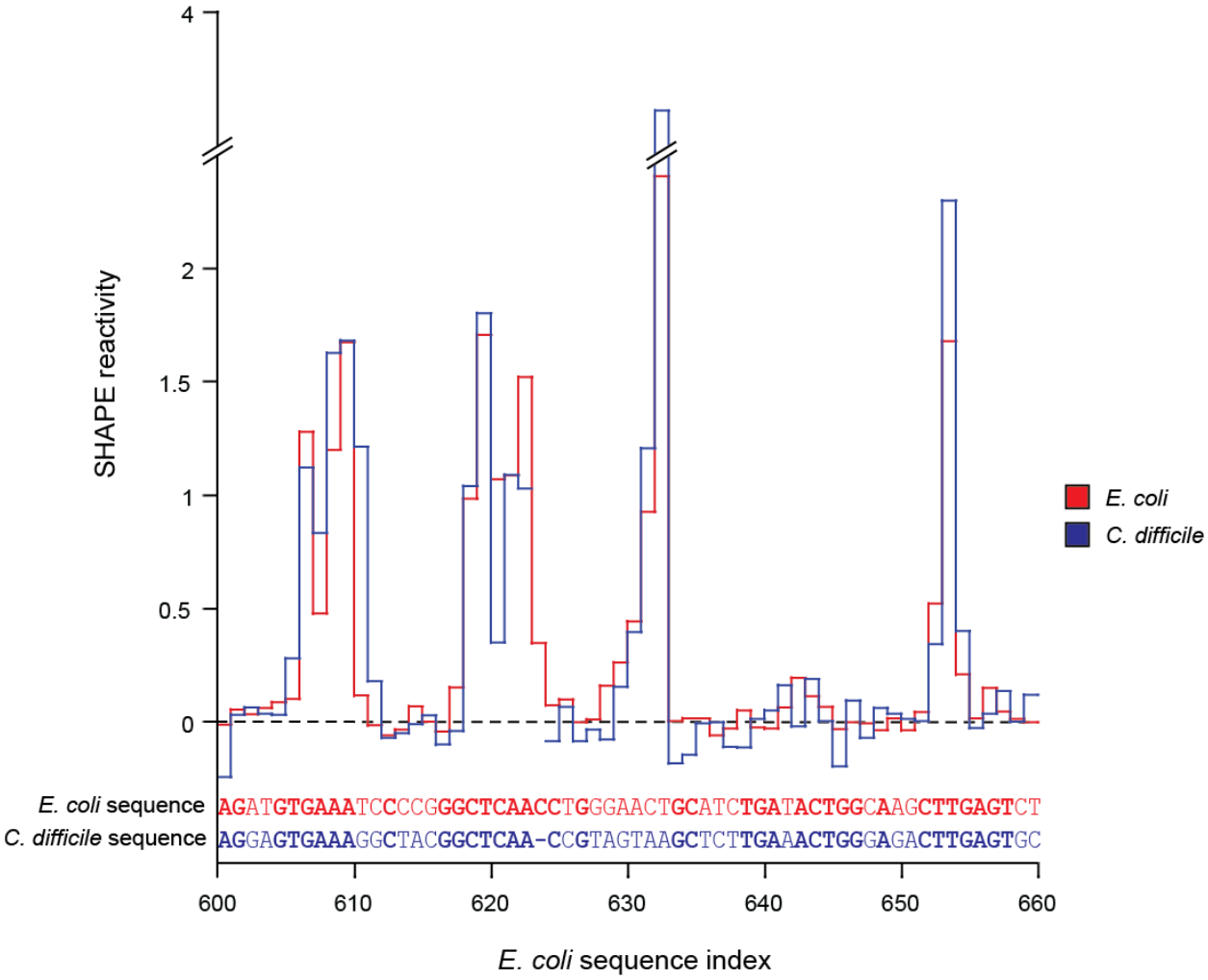



**Figure 4.4** Representative section of the SHAPE-dependent global alignment between *E. coli* and *C. difficile* 16S ribosomal RNAs. Only SHAPE values (red and blue lines) were used in the generation of this alignment. Alignment is shown as a function of *E. coli* sequence numbering.

**Table 4.1** Sensitivities of pairwise SHAPE-dependent alignments relative to accepted CRW alignments. Pairwise sensitivities were also determined from multiple sequencing alignments generated using T-Coffee.

| Sequence 1 | Sequence 2 | Sensitivity (sens) relative to CRW alignments (%) | | | |
|---|---|---|---|---|---|
| | | Needle alignment | SHAPE-only alignment | SHAPE and base identity (pairwise) | SHAPE and base identity (MSA) |
| *E. coli* 16S | *C. difficile* 16S | 84 | 84 | 96 | 96 |
| | *H. volcanii* 16S | 72 | 77 | 90 | 90 |
| *E. coli* 23S | *C. difficile* 23S | 83 | 80 | 96 | 96 |
| | *H. volcanii* 23S | 58 | 43 | 75 | 75 |

*C. difficile* and to *H. volcanii* have sensitivities of 96% and 90%, respectively, when both base identity and SHAPE data are considered. For 23S rRNA, alignments to *C. difficile* and to *H. volcanii* have sensitivities of 96% and 75%, respectively.

**4.7 Generation of multiple sequence alignments with T-Coffee**

The SHAPE-based alignments were in turn be applied to T-Coffee[13] in order to generate multiple sequence alignments (MSAs) for both 16S and 23S rRNA samples. With generation of MSAs, alignment quality did not significantly change (Table 4.1). Had a larger number of sequences been compared, the MSA may have resulted in an increase in alignment accuracy relative to CRW MSAs.

**4.8 Secondary structure prediction with SHAPE-directed alignments**

Both sequence comparisons and SHAPE data have been used to direct secondary structure prediction.[14] Given that SHAPE-based alignments effectively combine both these sets of information, we sought to predict secondary structures using SHAPE-based alignments. Sequence comparison-based secondary structure predictions are highly dependent on alignment quality.[15, 16] Therefore, success in secondary structure prediction would offer further validation of the SHAPE-based alignment approach.

The generated alignments were used as arguments in secondary structure prediction with RNAalifold and RNAfold in the Vienna package.[17-19] RNAalifold uses a pseudo-free energy potential to bias predictions based on covariation. Base pairs supported by covariation are given a free-energy bonus. RNAalifold and RNAfold were modified to accept SHAPE data as an argument. As has been used in past approaches, a pseudo free energy term was incorporated based on SHAPE reactivity.[20] Secondary structure prediction was performed in two steps. In the first step, a consensus fold considering all RNA sequences was determined using RNAalifold. Base pairs with a pairing probability greater than 95% were taken forward into the second step, in which additional base pairs were found for each individual sequence using SHAPE-directed RNAfold.

Structures were predicted for 16S and 23S ribosomal RNAs (16S *E. coli* prediction shown in Figure 4.5). The predicted structures were compared to those based on covariation models found in the CRW. All SHAPE reactivity-based predicted structures had sensitivities greater than 85%, indicating structure predictions of high accuracy (Table 4.2). Predictions considering both sequence alignment and SHAPE reactivities had higher sensitivity and positive predictive values (ppvs) than RNAfold predictions

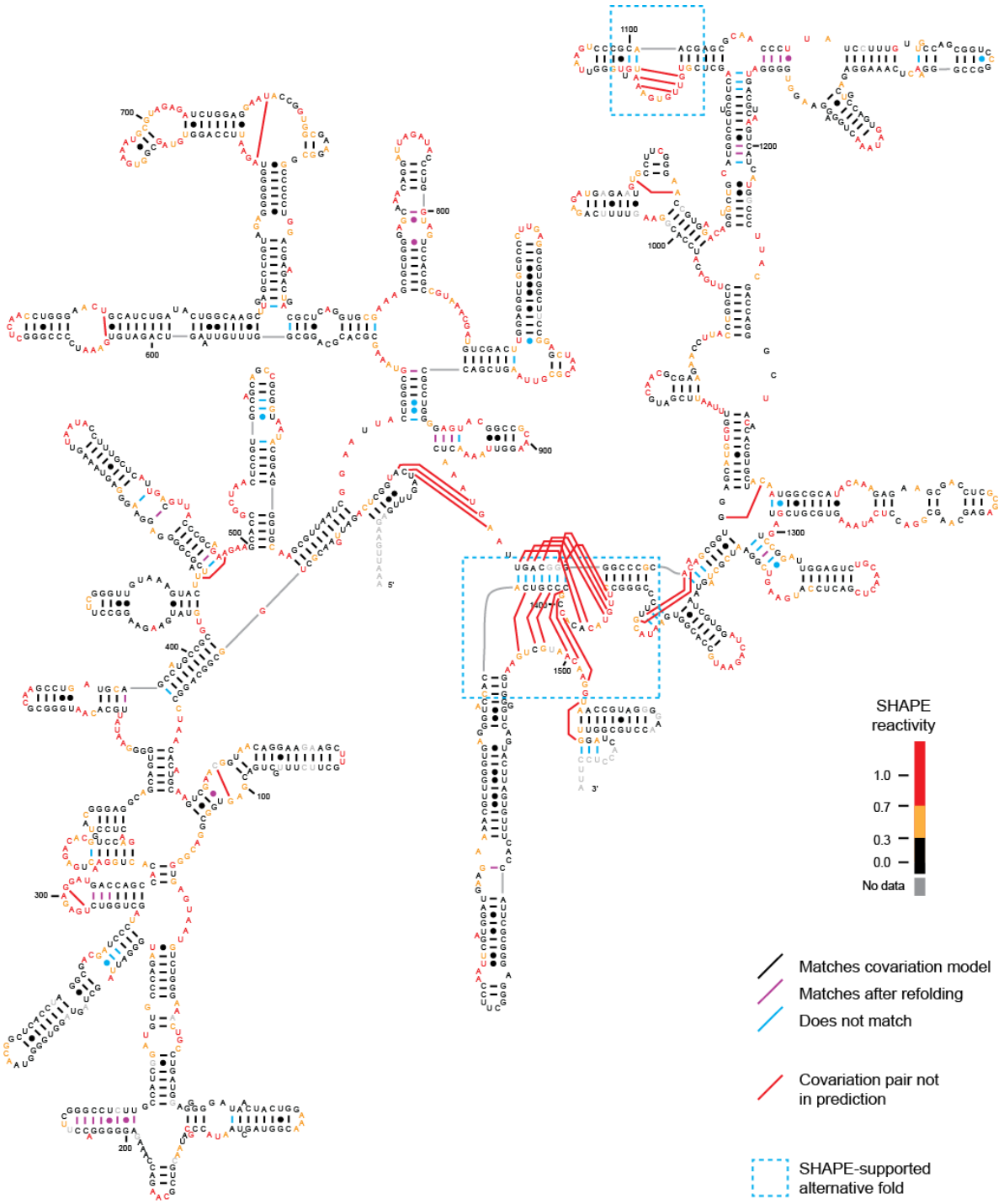



**Figure 4.5** Predicted secondary structure for *E. coli* 16S rRNA following constrained RNAfold prediction. Predicted pairs that exactly match the covariation model are shown in black, and predicted pairs that match after modest refolding are shown in purple. Predicted pairs not in the covariation model are shown in blue. Covariation pairs not in the predicted structure are shown using red lines. *E. coli* SHAPE values are shown by coloring of individual residues. Areas with SHAPE-supported alternative folds are shown boxed in blue.

**Table 4.2** Sensitivities (sens) and positive predictive values (ppvs) for secondary structure predictions.

| RNA structure predicted | RNAfold (SHAPE only) | | RNAalifold consensus, >95% pairing probability | | RNAfold constrained by consensus base pairs | | Constrained RNAfold without alternative structure regions | |
|---|---|---|---|---|---|---|---|---|
| | sens (%) | ppv (%) | sens (%) | ppv (%) | sens (%) | ppv (%) | sens (%) | ppv (%) |
| *E. coli* 16S | 90.7 | 83.5 | 79.2 | 91.6 | 95.5 | 88.5 | 98.1 | 89.8 |
| *C. difficile* 16S | 90.9 | 83.8 | 84.3 | 90.3 | 95.0 | 86.6 | 97.6 | 87.9 |
| *H. volcanii* 16S | 90.1 | 81.3 | 74.8 | 89.3 | 90.6 | 84.1 | 90.1 | 85.3 |
| | | | | | | | | |
| *E. coli* 23S | 84.2 | 78.2 | 78.0 | 89.9 | 85.3 | 79.0 | | |

considering SHAPE data alone. Given the sensitivity of covariation-based approaches to alignment quality, the successful modeling further validates the utility of the SHAPE-based alignment.

## 4.9 Discussion

### 4.9.1 Success and further application of SHAPE-based comparisons

SHAPE-based alignments have accuracies similar to traditional approaches that consider base identity. When both SHAPE data and base identity are considered, alignment quality increased significantly, with alignment sensitivities greater than 95% for alignments of *E. coli* ribosomal RNA to *C. difficile* sequences.

SHAPE-based alignments were performed with a dynamic programming algorithm in which a pair-wise scoring system was used. The pair-wise scoring system is analogous to the substitution matrices commonly used in alignment approaches. SHAPE-based sequence comparisons may also be readily applied to other existing sequencing systems. For methods using a heuristic scoring system, such as BLAST, a SHAPE-based scoring system may be applied following optimization.[21] SHAPE-based sequence alignments may also be used to train Markov-based alignment methods.[5]

### 4.9.2 Alternative base-pairing arrangements in the 16S rRNA

Areas of disagreement between SHAPE-based prediction and established covariation models (Figures 4.6 and 4.7) may be indicative of alternative structures adopted by the ribosome under the experimental conditions used for SHAPE probing. In various regions of each of the three rRNAs, the SHAPE-based models are in better agreement with SHAPE reactivities than the covariation models. For instance, the stem-loop beginning at *E. coli* 16S residue 1074 does not form in SHAPE-based predicted structures and is not present in the RNAalifold consensus fold when SHAPE data is included (Figure 4.6). The predicted models also differ from the covariation models in the 16S structure near the decoding site (Figure 4.7). These predicted models may be representative of states adopted by the ribosome during the dynamic process of translation. If these regions are excluded from prediction sensitivity calculations, the sensitivity of the *E. coli* 16S prediction increases from 95.5% to 98.1%.

## 4.10 Conclusion

SHAPE-dependent alignment is highly accurate and allows for consideration of RNA structure during sequence comparisons. SHAPE-based alignments had accuracies comparable to traditional

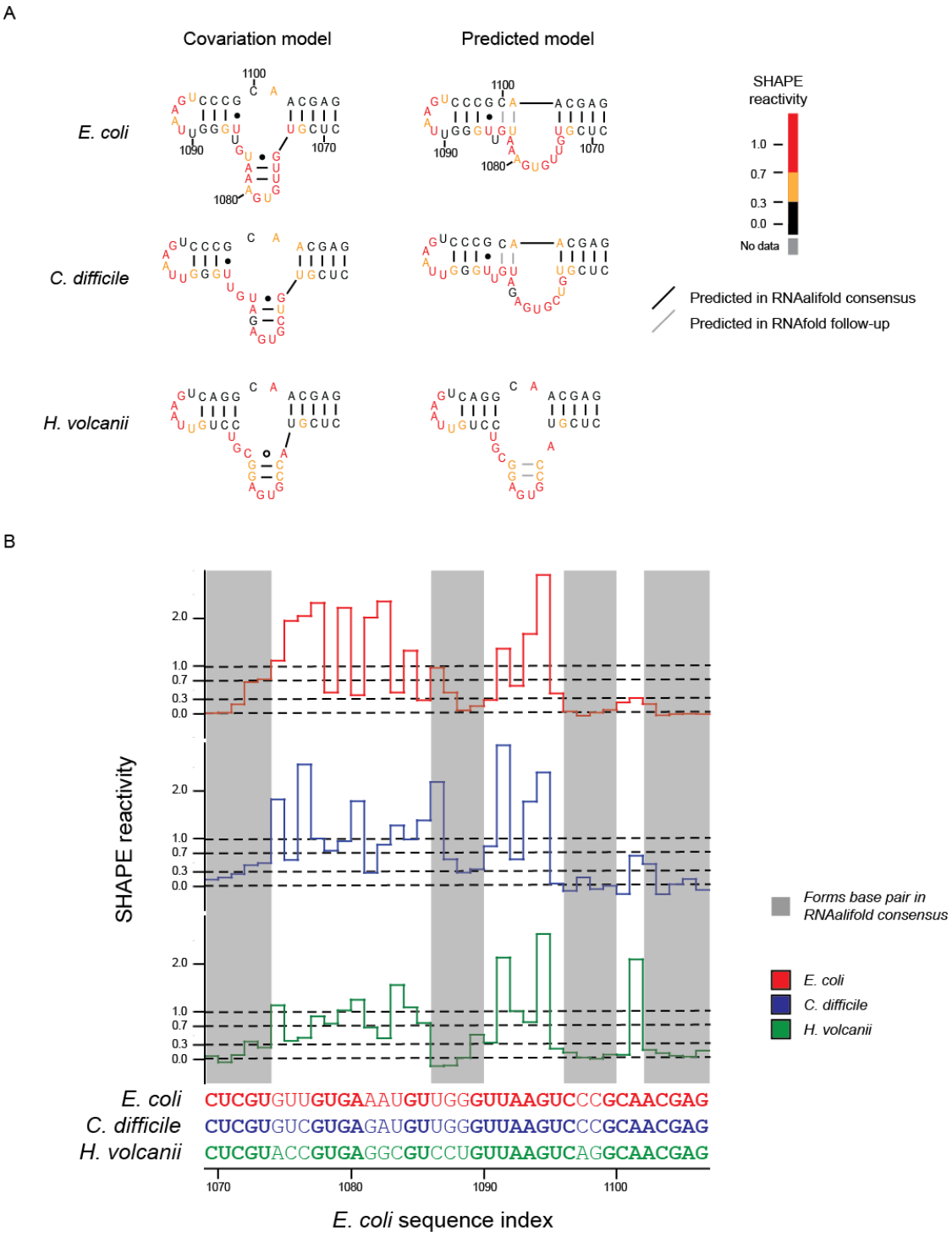



**Figure 4.6** Consensus alternative structures for the stem-loop at *E. coli* 16S rRNA residues 1069 to 1106. (**A**) Structures for the covariation and predicted models are shown with base pairs predicted in the RNAalifold consensus shown in black and base pair predicted in the constrained RNAfold prediction shown in gray. (**B**) SHAPE reactivities are shown for this region of the alignment with areas participating in RNAalifold consensus base pairs highlighted in gray.

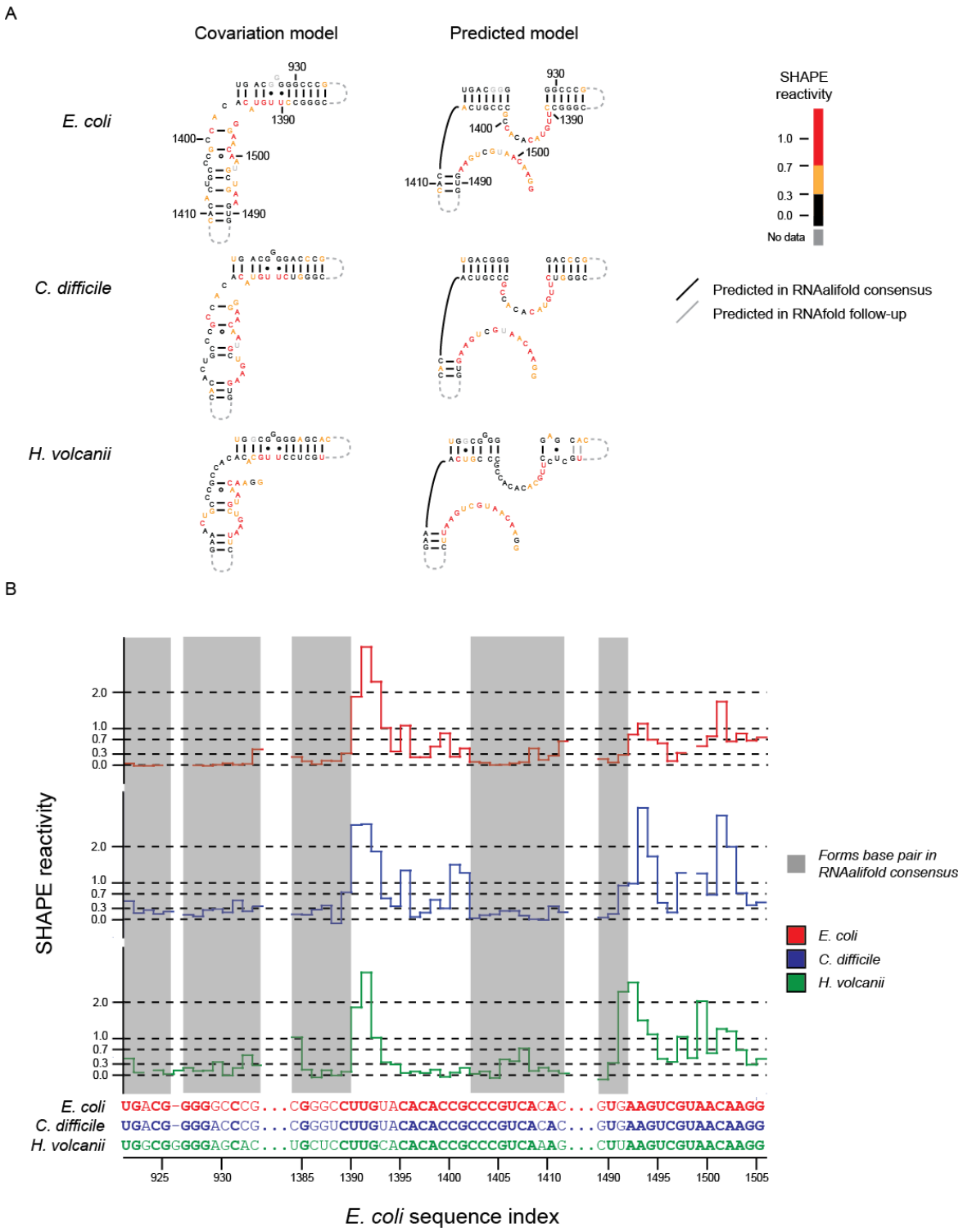



**Figure 4.7** Consensus alternative structures for the structural element at *E. coli* 16S rRNA residues 921 to 933, 1384 to 1411, and 1489 to 1505. (**A**) Structures for the covariation and predicted models are shown with base pairs predicted in the RNAalifold consensus shown in black and base pair predicted in the constrained RNAfold prediction shown in gray. (**B**) SHAPE reactivities are shown for this region of the alignment with areas participating in RNAalifold consensus base pairs highlighted in gray.

sequence identity-based approaches. The alignments generated by this technique were used to provide information for secondary structure prediction techniques. Predictions were highly accurate when compared to accepted models. Differences from the accepted models are supported by SHAPE reactivities and suggest the presence alternative structures not necessarily consistent with covariation. Based upon success with a dataset derived from NGS approaches, SHAPE-based alignments may see utility in future high-throughput structure-based RNA motif searches and in structure characterization.

### 4.11 Experimental Methods

#### 4.11.1 *E. coli* ribosomal RNA SHAPE-MaP data

*E. coli* ribosomal RNA SHAPE-MaP data were taken from previous studies (Siegfried *et al*., in preparation).

#### 4.11.2 Preparation of *C. difficile* ribosomal RNA

*C. difficile* was grown in BHIS medium[22] at 37 °C under anaerobic conditions (90% $N_2$, 5% $CO_2$, and 5% $H_2$).[23] Cells were grown to an $OD_{600}$ of 1.0. Cells were collected by centrifugation (10 min, 4 °C, 4,000 □g). The pellet was washed with 1x TE [10 mM Tris (pH 8.0), 1 mM EDTA]. The supernatant was discarded, and the pellet was allowed to air-dry for 5 minutes.

To lyse the cells, 1 mL TRIsure (Bioline) was added to the pellet. The resultant mixture was incubated at room temperature for 5 minutes. The mixture was then transferred to a vial containing 250 µL 0.1 mm glass beads. Cells were lysed by bead beater over two 90 second pulses, with cells held on ice in between pulses. The resultant mixture was extracted with 200 µL chloroform, with the aqueous phase taken forward.

Following lysis, this solution was extracted three times with phenol [(pH 8.0): chloroform:isoamyl alcohol; 25:24:1], followed by three times with chloroform. The RNA-containing solution was exchanged for folding buffer [50 mM Hepes (pH 8.0), 200 mM potassium acetate (pH 8.0), and 5 mM $MgCl_2$] using a pre-equilibrated gel filtration column (G-25 column, GE). The quantity of RNA was found using absorption spectroscopy.

#### 4.11.3 Preparation of *H. volcanii* ribosomal RNA

Growth medium was prepared by bringing 600 ml 30% salt solution [4 M sodium chloride, 150 mM magnesium chloride hexahydrate, 150 mM magnesium sulfate heptahydrate, 100 mM potassium

chloride, 5 mM Tris (pH 7.5)], 5 g bacteriological peptone (LP37; Oxoid), and 1 g yeast extract (LP21; Oxoid) to 1 L with deionized water. Cells were grown to an $OD_{600}$ of 0.8. Cells were collected by centrifugation (5 min, 4 °C, 14,000 □g).

Cells were lysed by incubation in low salt solution [220 µL 50 mM Hepes (pH 8.0) and 5 mM $MgCl_2$; incubation at 22 °C for 5 min, followed by incubation on ice for 5 min]. Following lysis, this solution was extracted three times with phenol [(pH 8.0): chloroform:isoamyl alcohol; 25:24:1], followed by three times with chloroform. The RNA-containing solution was exchanged for folding buffer [50 mM Hepes (pH 8.0), 200 mM potassium acetate (pH 8.0), and 5 mM $MgCl_2$] using a pre-equilibrated gel filtration column (G-25 column, GE). The quantity of RNA was found using absorption spectroscopy.

**4.11.4 SHAPE-MaP characterization of ribosomal RNA**

Determination of SHAPE reactivity by SHAPE-MaP requires three different experiments: chemical modification of native RNA, chemical modification of denatured RNA, and a no-modification control. All chemical modifications were performed using 1-methyl-7-nitroisatoic anhydride (1M7). Chemical modification of native RNA and the no-modification control were performed in parallel. To 1x folding buffer [50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), and 5 mM $MgCl_2$] was added to a concentrated RNA solution (280 ng *H. volcanii* total RNA or 70 ng *C. difficile* total RNA) to a final volume of 90 µL. The RNA solution was incubated at 37° C for 30 minutes. Following incubation, 10 µL DMSO (no-modification control) or 10 µL 100 mM 1M7 in DMSO (native 1M7-modified sample) was added to the RNA solution. The RNA solution was then held at 37 °C for 3 minutes.

For the denatured control, 25 µL 4x denatured control buffer [200 mM HEPES (pH 8.0), 16 mM EDTA] and 50 µL deionized formamide were added to a concentrated RNA solution (280 ng *H. volcanii* total RNA or 70 ng *C. difficile* total RNA), and deionized water was added to a final volume of 90 µL. This solution was held at 95 °C for 1 minute, and then 10 µL 100 mM 1M7 in DMSO was added. The combined solution was held at 95 °C for 1 minute.

After modification, all three samples were purified using an RNeasy Min-Elute kit (Qiagen), eluting into 22 µL Qiagen elution buffer. To prepare sequencing libraries, the purified RNA samples were first fragmented using divalent cations: 20 µL of RNA solution was combined with 30 µL fragmentation buffer [250 mM Tris (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$], held at 94 °C for 4 minutes, and then transferred

immediately to ice. Fragmented RNA was then purified using a G-25 column (GE), eluting into 1x TE [10 mM Tris (pH 8.0), 1 mM EDTA].

Following fragmentation, reverse transcription (RT) was performed with random primers. A 20-µL aliquot of fragmented RNA was combined with 2 µL random DNA nonamers (200 ng/µL). The solution was incubated at 65 °C for 5 minutes and then moved to ice. To this solution, 7 µL reaction buffer [286 mM Tris (pH 8.0), 429 mM KCl, 57 mM DTT, 2.9 mM dNTP mix (dATP, dCTP, dGTP, and dTTP, 2.9 mM each)], 4 µL 60 mM $MnCl_2$, and 5 µL water were added. The solution was pre-incubated at 25 °C for 2 minutes prior to adding 2 µL Superscript II (Invitrogen). The reaction was incubated at 25 °C for 10 minutes, 42 °C for 180 minutes, and 70 °C for 15 minutes. Following reverse transcription, the RNA was purified using a G-25 column (GE), eluting into 1x TE.

The cDNA was converted to a double-stranded DNA library with Illumina platform-specific sequence tags. First, 40 µL of the purified RT product was used in a 80-µL second-strand synthesis reaction using standard conditions (NEBNext Second Strand Synthesis Module, New England Biolabs). The second strand synthesis reaction was applied to a PureLink PCR Micro Kit (Life Technologies), eluting into 12 µL elution buffer. A 10-µL aliquot of the purified DNA solution was then used in a 50-µL end repair reaction using standard conditions (NEBNext End Repair Module, New England Biolabs). Following end repair, the DNA was purified using a 1.6x Ampure XP Bead clean-up (Agencourt, Beckman Coutler), eluting into a final volume of 30 µL 1x TE.

To incorporate Illumina platform-specific sequence tags, a dA-tailing reaction was first used to incorporate a single-nucleotide overhang on the 3' ends of the double-stranded DNA. A 15-µL aliquot of purified DNA from the end repair step was used in a 20-µL dA-tailing reaction (NEBNext dA-Tailing Module, New England Biolabs). Standard manufacturer-recommended conditions were used. Illumina sequences were incorporated using a ligation step with Illumina iAdapters (prepared in house). Immediately following completion of the dA-tailing reaction, 7.5 µL 5x reaction buffer (NEBNext Quick Ligation Module, New England Biolabs), 2.5 µL 125 nM DNA adapter, 3.75 µL Quick T4 DNA Ligase (New England Biolabs), and 3.75 µL water were added to the dA-tailing reaction mix. The ligation reaction was then incubated at 20 °C for 15 minutes. The ligation reaction was purified using a 1.0x Ampure XP

Bead clean-up (Agencourt, Beckman Coutler). This bead cleanup was performed twice with final elution into 20 µL 10 mM Tris, pH 8.0.

Illumina libraries were prepared using emulsion PCR [REF]. The aqueous phase was composed of 5 µL of double-stranded DNA, 10 µL 10 µM Illumina-specific forward strand primer, 10 µL 10 µM Illumina-specific reverse strand primer, 40 Q5 5x reaction buffer (New England Biolabs), 100 µL 20 g/L bovine serum albumin, 4 µL dNTP mix (10 mM each, dATP, dCTP, dGTP, dTTP), 2 µL Q5 high-fidelity polymerase (New England Biolabs), and 29 µL water. The DNA was amplified in a 35-cycle PCR reaction (denaturation: 94 °C for 30 seconds; annealing: 67 °C for 30 seconds; extension: 72 °C for 30 seconds). To purify the PCR product, the reaction was first applied to a PureLink PCR cleanup column (Life Technologies). The column eluent was then purified using a 1.0x Ampure XP Bead clean-up (Agencourt, Beckman Coutler). This bead cleanup was performed twice with elution into 12 µL 10 mM Tris, pH 8.0.

The concentrations of sequencing samples were determined by Qubit High Sensitivity DNA fluorescence assays (Life Technologies) and High Sensitivity DNA Bioanalyzer assays (Agilent). Each sample was diluted to 2 nM and pooled. The pooled library was sequenced using an Illumina MiSeq (300 cycles - PE kit).

Sequences were aligned and mutation events counted using the SHAPE-MaP analysis pipeline (Siegfried *et al*, in preparation). SHAPE reactivities were generated considering mutation rates in the native 1M7-modified sample, the denatured 1M7-modified sample, and the background control.

**4.11.5 SHAPE-based RNA alignment**

SHAPE-based alignment of two sequences $x$ and $y$ began with declaration of an empty score matrix $F$ and an empty pointer matrix $P$. Each matrix had dimensions $m$ by $n$, where $m$ and $n$ are the lengths of sequences $x$ and $y$ plus 1, respectively. The zeroth row and zeroth column of each matrix were set to 0. Every other cell in the matrix was populated by recursion with sequence values $x_i$ and $y_j$ corresponding to $F_{i,j}$. The value at each cell in score matrix $F$ was determined by functions $s(x_i, y_j)$ and $g(i, j)$, which describe a pair-wise comparison score and associated gap penalty, respectively. This recursion took the following form, where $i$ and $j$ describe the position of a given cell and $x_i$ and $y_j$ are the SHAPE values of each sequence at $ith$ and $jth$ positions:

$$F_{i,j} = max\begin{cases} F_{i-1,j-1} + s(x_i, y_j) \\ F_{i,j-1} + g(i, j-1) \\ F_{i-1,j} + g(i-1, j) \end{cases}$$

The pointer matrix records the direction of the recursion and whether or not a gap was included in the alignment. A cell $P_{i,j}$ was assigned based on the maximum value found in the calculation of cell $F_{i,j}$:

$$P_{i,j} = \begin{cases} 1, max = F_{i-1,j-1} + s(x_i, y_j) \\ -1, max = F_{i,j-1} + g(i, j-1) \\ -2, max = F_{i-1,j} + g(i-1, j) \end{cases}$$

The scoring function is described by the following equation, with parameters $N_0$, $\lambda$, and $b$:

$$s(x_i, y_j) = N_0 e^{-\lambda |x_i - y_j|} + b$$

The gap penalty function $g$ is dependent on values in the pointer matrix $P$, with parameters $GOP$ (gap open penalty) and $GEP$ (gap extension penalty):

$$g(i,j) = \begin{cases} GOP, P_{i,j} = 1 \\ GEP, P_{i,j} = -1 \\ GEP, P_{i,j} = -2 \end{cases}$$

If base identity was considered during alignment, it was added as an additional scoring term $b$ in the recursion, where $x'_i$ and $y'_j$ are the base identities at positions $i$ and $j$ in sequences $x$ and $y$, respectively:

$$F_{i,j} = max\begin{cases} F_{i-1,j-1} + s(x_i, y_j) + b(x'_i, y'_j) \\ F_{i,j-1} + g(i, j-1) \\ F_{i-1,j} + g(i-1, j) \end{cases}$$

The scoring function $b$ is described by the following equation with parameters $MATCH$ and $MISMATCH$.

$$b(x'_i, y'_j) = \begin{cases} MATCH, x'_i = y'_j \\ MISMATCH, x'_i \neq y'_j \end{cases}$$

When considering base identity, values in matrix $P$ also considered the scoring function $b$:

$$P_{i,j} = \begin{cases} 1, max = F_{i-1,j-1} + s(x_i, y_j) + b(x'_i, y'_j) \\ -1, max = F_{i,j-1} + g(i, j-1) \\ -1, max = F_{i-1,j} + g(i-1, j) \end{cases}$$

Following population of $F$ and $P$ matrices, a trace-back operation was used to find the optimal alignment. The trace-back operation began at the maximum value in the $F$ matrix. This represents the 3'-most position of the alignment. At any given position $i, j$, the next position was selected based on pointer matrix $P$. If $P_{i,j}$ = 1, the next position in the alignment was a match, and the next position considered in the $F$ matrix was $F_{i-1,j-1}$. If $P_{i,j}$ = -1 or $P_{i,j}$ = -2, the next position in the alignment was a gap, and the next

position considered in the $F$ matrix was $F_{i,j-1}$ or $F_{i-1,j}$, respectively. The trace-back operation was finished when a position was encountered where $i$ = 0 or $j$ = 0.

The SHAPE scoring parameters $N_0$, $b$, and $\lambda$, the base-identity scoring parameters $MATCH$ and $MISMATCH$, and the gap penalty parameters $GOP$ and $GEP$ were optimized using grid search. In experiments considering only SHAPE values, $N_0$ = 4, $b$ =-1, $\lambda$ = 1, $GOP$ = 5, and $GEP$ = 0. When both SHAPE and base identity were considered, $N_0$ = 4, $b$ = -1, $\lambda$ = 1, $GOP$ = 9, $GEP$ = -0.5, $MATCH$ = 1, and $MISMATCH$ = -1.5.

**4.11.6 Evaluation of RNA sequence alignments**

From multiple sequence alignments on the CRW, pairwise alignments between *E. coli* and *C. difficile* and *E. coli* and *H. volcanii* were taken for both 16S and 23S rRNA. RNA sequence alignments generated in this work were evaluated by comparison to these alignments. Sensitivities were calculated as the percentage of matched nucleotides in the CRW alignments found in a given alignment.

**4.11.7 Generation of multiple sequence alignments**

Multiple sequence alignments were generated using T-Coffee.[13] First, pairwise alignments were generated for all possible pairs between considered sequences. These pair-wise alignments were then used as arguments for T-Coffee. Default T-Coffee parameters were used.

**4.11.8 Secondary structure prediction by SHAPE-based RNA alignments**

Multiple sequence alignments were used as input for RNAalifold of the Vienna RNA software package.[18] Using a new implementation of the RNAalifold algorithm, SHAPE data was used as an additional input to constraint secondary structure prediction. Secondary structure prediction and partition function calculations were performed using the ribosum matrix with a maximum base pairing distance of 600 nucleotides.

Following RNAalifold prediction, all base pairs in the consensus sequence with pairing probabilities greater than 95% were used as constraints in individual follow-up predictions with RNAfold, also of the Vienna RNA package.[17] SHAPE data was also used to constrain secondary structure prediction in this step. A maximum base pairing distance of 600 nucleotides was maintained in this follow-up prediction.

#### 4.11.9 Evaluation of secondary structure predictions

Predicted secondary structures were evaluated by calculating sensitivity as the percentage of base pairs from the CRW covariation model found in predicted structures and by calculating positive predictive values (ppvs) as the percentage of predicted pairs found in the covariation model. It should be noted that these reference structures are themselves experimental models, and base pairs in these models may not necessarily be present in the structure adopted by the ribosomal RNA under experimental conditions. To account for this, base pairs in the covariation model whose constituent nucleotides have high SHAPE reactivities (both greater > 0.7) were not considered in sensitivity calculations. Also, when comparing base pairs between the covariation and predicted models, a modest local refolding allowance of 5 nucleotides was included. To be considered a matched pair, for a base pair in the covariation model at positions $x$ and $y$ and any base pair in the predicted model at positions $x'$ and $y'$ the following must be true:

$$[x = x' and\ |y - y'| \leq 5]\ or\ [y = y' and\ |x - x'| \leq 5]$$

Pseudoknots and non-canonical base pairs (with the exception of G-U pairs) were not considered in sensitivity and positive predictive value (ppv) calculations.

### 4.10 References


1. Vogel, F. (1964) A Preliminary Estimate of the Number of Human Genes, *Nature 201*, 847.

2. Lander, E. S.*, et al.* (2001) Initial sequencing and analysis of the human genome, *Nature 409*, 860-921.

3. Consortium, E. P.*, et al.* (2012) An integrated encyclopedia of DNA elements in the human genome, *Nature 489*, 57-74.

4. Djebali, S.*, et al.* (2012) Landscape of transcription in human cells, *Nature 489*, 101-108.

5. Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009) Infernal 1.0: inference of RNA alignments, *Bioinformatics 25*, 1335-1337.

6. Havgaard, J. H., Lyngso, R. B., and Gorodkin, J. (2005) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search, *Nucleic Acids Res 33*, W650-653.

7. Hofacker, I. L., and Lorenz, R. (2014) Predicting RNA structure: advances and limitations, *Methods Mol Biol 1086*, 1-19.

8. Lucks, J. B.*, et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), *Proc Natl Acad Sci U S A 108*, 11063-11068.

9. Cannone, J. J.*, et al.* (2002) 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.

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

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

12. Needleman, S. B., and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J Mol Biol 48*, 443-453.

13. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol 302*, 205-217.

14. Seetin, M. G., and Mathews, D. H. (2012) RNA structure prediction: an overview of methods, *Methods Mol Biol 905*, 99-122.

15. Gutell, R. R.*, et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods, *Nucleic Acids Res 20*, 5785-5795.

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

17. Schuster, P.*, et al.* (1994) From sequences to shapes and back: a case study in RNA secondary structures, *Proc Biol Sci 255*, 279-284.

18. Hofacker, I. L., Fekete, M., and Stadler, P. F. (2002) Secondary structure prediction for aligned RNA sequences, *J Mol Biol 319*, 1059-1066.

19. Lorenz, R.*, et al.* (2011) ViennaRNA Package 2.0, *Algorithms Mol Biol 6*, 26.

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

21. Altschul, S. F.*, et al.* (1990) Basic local alignment search tool, *J Mol Biol 215*, 403-410.

22. Smith, C. J., Markowitz, S. M., and Macrina, F. L. (1981) Transferable tetracycline resistance in Clostridium difficile, *Antimicrob Agents Chemother 19*, 997-1003.

23. Purcell, E. B.*, et al.* (2012) Cyclic diguanylate inversely regulates motility and aggregation in Clostridium difficile, *J Bacteriol 194*, 3307-3316.

## CHAPTER 5: STRUCTURE ALIGNMENT AND CONSENSUS SECONDARY STRUCTURE PREDICTION FOR THREE HIV-RELATED RNA GENOMES

### 5.1 Introduction

RNA plays an active role in most biological processes.[1] Multiple examples of RNA functions are found in the life cycle of positive-strand RNA lentiviruses. The genomes of RNA viruses function at two different levels: the level of encoded proteins and the level of functional higher-order RNA structures. Constrained by a small genome size, these viruses make use of genomes that are very efficient in terms of sequence allocation, and multiple RNA structures exist in the genomes that regulate replicative processes.

The human immunodeficiency virus (HIV) has well-defined structural RNA elements that play key regulatory roles throughout the replication cycle. During transcription of the integrated viral genome, a stem-loop structure in the 5' untranslated region (UTR) called the TAR hairpin binds the Tat protein to modulate host proteins involved in transcription.[2, 3] Found within the *env* gene, the Rev response element (RRE) binds the viral protein Rev, allowing for export of unspliced and partially spliced viral mRNA out of the nucleus.[4] During translation, the *gag-pol* frameshift element modulates the reading frame of the ribosome, tightly regulating production of the Gag-Pol polypeptide.[5, 6] Stem-loop structures in the Psi packaging element are required for efficient packaging of viral genome into nascent virions.[7] To date, most characterization of the HIV genome has been directed at the 5' and 3' untranslated regions. Based on recent analyses, it is clear that the central coding region of the HIV genome contains extensive base pairing and secondary structure.[8, 9] The question remained as to the functional significance of many of these structures.

A powerful means to characterize RNA secondary structure is based on chemical modification of the RNA. In the SHAPE chemical modification approach, an electrophile with structure-selective reactivity preferentially acylates the 2'-hydroxyl of unstructured RNA nucleotides.[10] The extent of modification is inversely proportional to the tendency of an RNA nucleotide to participate in an RNA base pair or other structural interactions. Recently, next-generation sequencing has been used to detect the sites of SHAPE

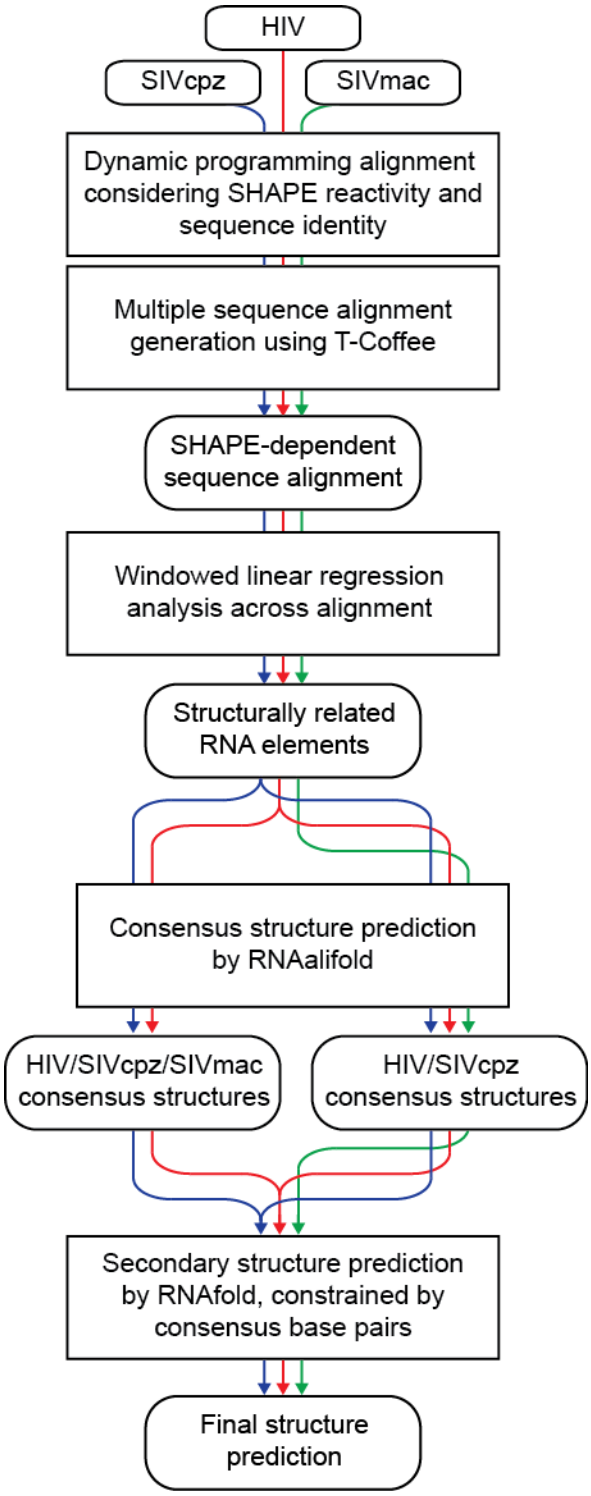



**Figure 5.1** Flow chart indicating steps in genome comparison analysis.

modifications in an approach called SHAPE-MaP (Siegfried *et al.*, in preparation). This approach takes advantage of the tendency of reverse transcriptase to incorporate a mutation in cDNA sequence at corresponding positions of chemically modified RNA residues. Observed mutation rates are related to an absolute scale of SHAPE reactivity.

To address the question of the functional significance of RNA structures in the HIV genome, the conservation of structural features was characterized across the genomes of HIV and two related lentiviruses, SIVcpz MB897 and SIVmac239. A flowchart of the analysis workflow is shown in Figure 5.1. Using chemical modification data, structure-dependent sequence alignments were generated for the three HIV-related strains. Linear regression analysis was used to find areas in the alignment where chemical modification patterns were statistically similar. Finally, secondary structure prediction was performed considering both SHAPE reactivities and covariation. This analysis defined regions where structure was similar across the three HIV-related strains as supported by chemical modification and sequence covariation across three genomes.

**5.2 Selection of virus strains for characterization**

Viral strains were selected on the basis of similarity to reference HIV strain NL4-3, member of HIV-1 group M. One of the selected strains is closely related to HIV-1 NL4-3, and one is distantly related. SIVcpz MB897 (SIVcpz) is a strain found in chimpanzees that is closely related to HIV-1 group M strains.[11, 12] SIVmac239 (SIVmac) is a representative of the SIVsm/HIV-2 lineage that is more distantly related to HIV-1 group M strains.[13] SIVcpz and SIVmac have percent identities of 77.4%and 54.6%, respectively, when compared to NL4-3 using standard codon-based alignments.

**5.3 Characterization by SHAPE-MaP**

SHAPE data for HIV was taken from previous analysis (Siegfried *et al.*, in preparation), and SHAPE data for SIVcpz and SIVmac were collected for this work. Authentic SIVcpz and SIVmac genomic RNAs were purified from mature virions by Dr. Robert Gorelick. In order to preserve the secondary and tertiary of the RNA genome, no heating steps, chelating agents, or chemical denaturants were used during RNA genome purification.

Chemical modification of the viral RNAs with SHAPE reagent 1-methyl-7-nitroisatoic anhydride (1M7) was performed under physiological conditions.[10, 14] Following chemical modification, the extent of

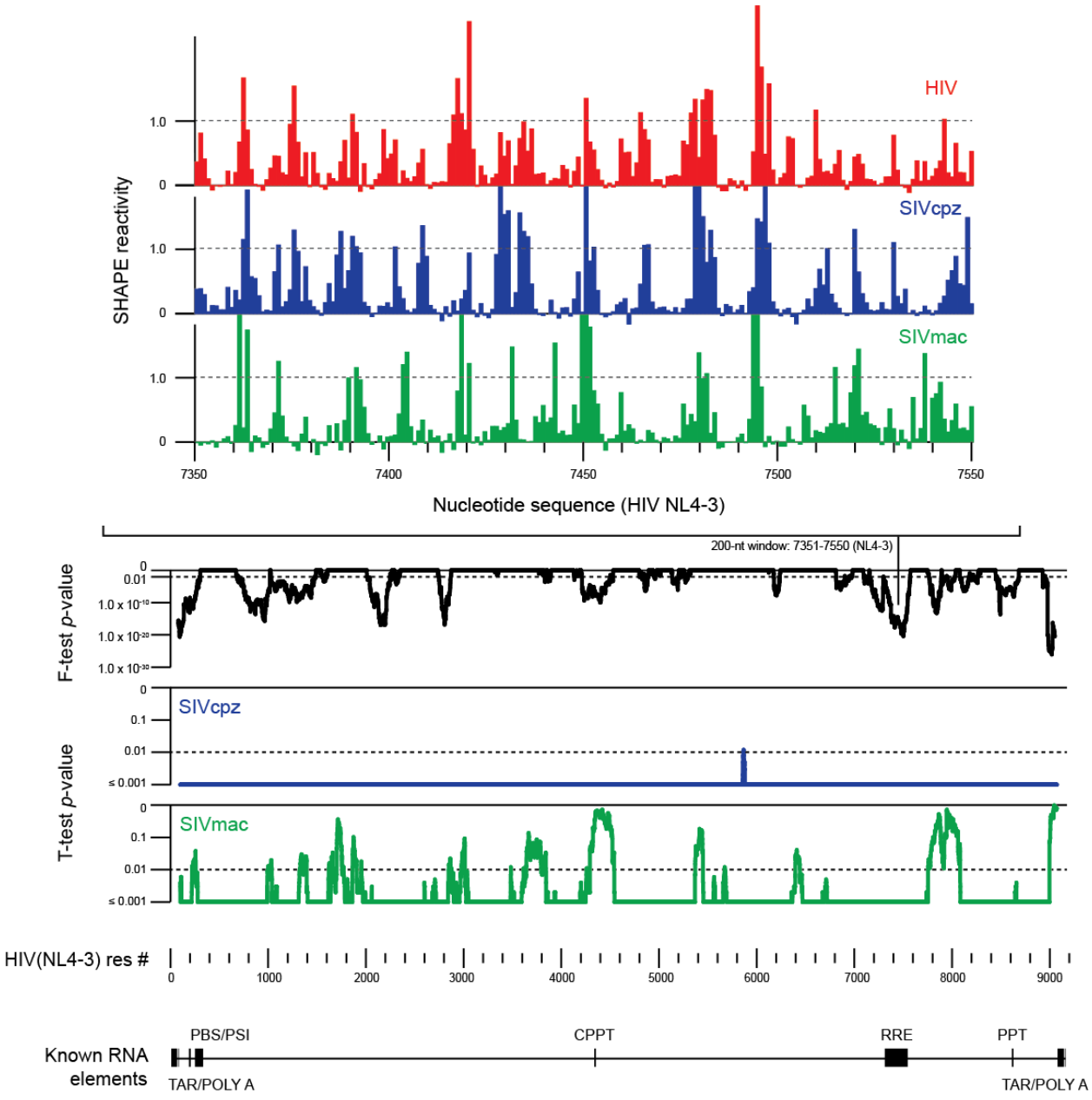



**Figure 5.2** Sample alignment and multi-variable linear regression analysis statistics. At the top, region of the SHAPE-dependent alignment is shown. Windowed linear regression statistics are shown for the same region as a function of the HIV (NL4-3) sequence, indicating the statistical significance of the relationship between SHAPE data sets. F-test *p*-values for the alignment (black), as well as t-test values assay dependence of individual sequences (blue for SIVcpz, green for SIVmac), are shown across the full-length alignment. Known RNA elements are noted at the bottom of the figure.

modification at each nucleotide was determined using mutational profiling (Siegfried *et al.*, in preparation). SHAPE reactivity values were determined for each position with sufficient read depth in the analyzed genomes by comparing mutation rates of a 1M7-modified sample relative to background controls. SHAPE reactivity is correlated with the flexibility of a given nucleotide; nucleotides with low SHAPE reactivity tend to participate in base pairs or other interactions, whereas nucleotides with high SHAPE reactivity tend to be in unstructured regions of the RNA.

**5.4 Generation of a SHAPE-dependent whole-genome alignment**

Pairwise whole-genome alignments of HIV, SIVcpz, and SIVmac RNAs were created using a SHAPE-dependent dynamic programming approach described in Chapter 4. From pairwise comparisons, a multiple sequence alignment was generated using T-Coffee[15] (a representative region of this alignment is shown in Figure 5.2). In these alignments, all regions known to contain functional RNA structures were aligned; these included elements in the untranslated regions (TAR stems, Psi packaging element) and coding regions (*gag-pol* frameshift element, RRE). Additionally, polypurine tracts in the *pol* and *nef* genes (cPPT and PPT, respectively) aligned precisely.[16, 17] Alignment between HIV and SIVcpz also respected protein reading frames as all start codons precisely aligned, with the exception of the *vpu* start codon which aligned with a frameshift of precisely 2 codons.

**5.5 Evaluation of interdependence by multi-variable linear regression**

In order to evaluate the relationships among SHAPE data for the three genomes, multi-variable linear regression was performed over the multiple sequence alignment considering 200-nucleotide windows (Figure 5.2). The dependences of the SHAPE reactivities of HIV on SIVcpz and SIVmac RNAs were evaluated using the F-test. SHAPE data were fit to the following relationship by least squares, where $Y$ represents HIV SHAPE values and $X_1$ and $X_2$ represent SIVcpz and SIVmac SHAPE values, respectively:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

The F-test evaluates the following null hypothesis by the sum of squares due to lack-of-fit for the proposed model:

$$\beta_1 = \beta_2 = 0$$

Based on derived F-statistic measurements, *p*-values evaluating the significance of the interdependence of SHAPE values determined over the entire alignment. To gauge the contributions of SIVcpz and SIVmac individually, t-tests were performed for the fitted model considering only SIVcpz or SIVmac nucleotide reactivities (Figure 5.2).

The F-test revealed a number of regions across the whole-genome alignment with significant interdependences. Statistically similar regions with *p*-values less than 0.01 were found throughout the HIV genome, particularly at the 5' and 3' ends. Statistically similar regions were also found in the coding regions of the genome, particularly in *gag* and *env* genes, where functional elements like the *gag-pol* frameshift element and RRE have been characterized.[4, 5] All known functional elements are located in statistically interdependent regions. Critically, there are a number of regions of similarity where no functional RNA element is known. The t-test results correlate well with F-test results and show that the significance of the relationship between HIV and SIVcpz is greater than the relationship between HIV and SIVmac. This is consistent with relative sequence identity.

**5.6 Prediction of consensus secondary structures**

Using the multiple sequence alignment, consensus secondary structures were predicted considering base identity and SHAPE data across all three genomes (see Chapter 4). Results from F-test analysis were used to select specific regions to evaluate. We considered areas with F-test *p*-values less than 0.01. The ten areas that met this criterion ranged in length from 257 to 2071 nucleotides. Combined, these ten areas cover 72.0% of the HIV NL4-3 genome.

Consensus secondary structures of these regions of high similarity were generated using RNAalifold, which has been updated to incorporate pseudo-energy terms based on covariation and SHAPE data into structure predictions considering Turner free energy rules (see Chapter 4).[18-20] Two consensus structures were predicted: Both incorporated SHAPE data; one considered sequence alignment for all three genomes and the other considered pairwise alignment between HIV and SIVcpz, the two more closely related genomes. Consensus base pairs were then used to constrain a SHAPE-directed secondary structure prediction considering HIV alone.[18, 21] Only consensus base pairs with pairing probabilities greater than 95% that did not disagree between the two consensus structures were used to constrain the HIV-only prediction.

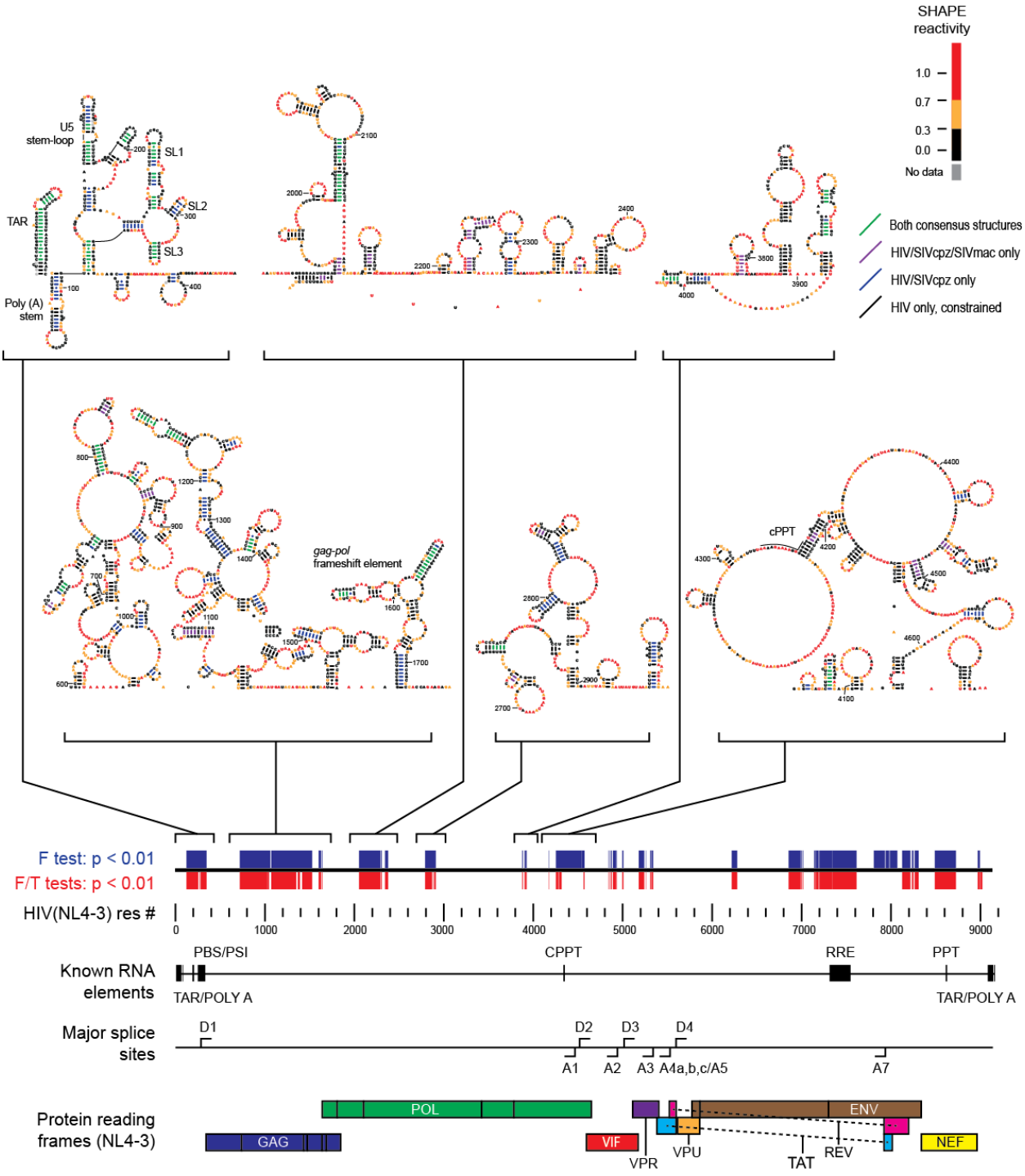



**Figure 5.3** Secondary structure predictions for the first six interdependent elements of the genome-wide alignment as ordered by sequence. Secondary structures shown are for the final constrained HIV secondary structure prediction, with consensus base pairs shown in purple (HIV/SIVcpz/SIVmac consensus), blue (HIV/SIVcpz consensus), and green (present in both consensus structures), and with nucleotides colored by HIV SHAPE reactivity. Predicted elements are shown on the HIV (NL4-3) genome with annotations indicating statistical dependence, known RNA elements, major splice sites, and protein reading frames.

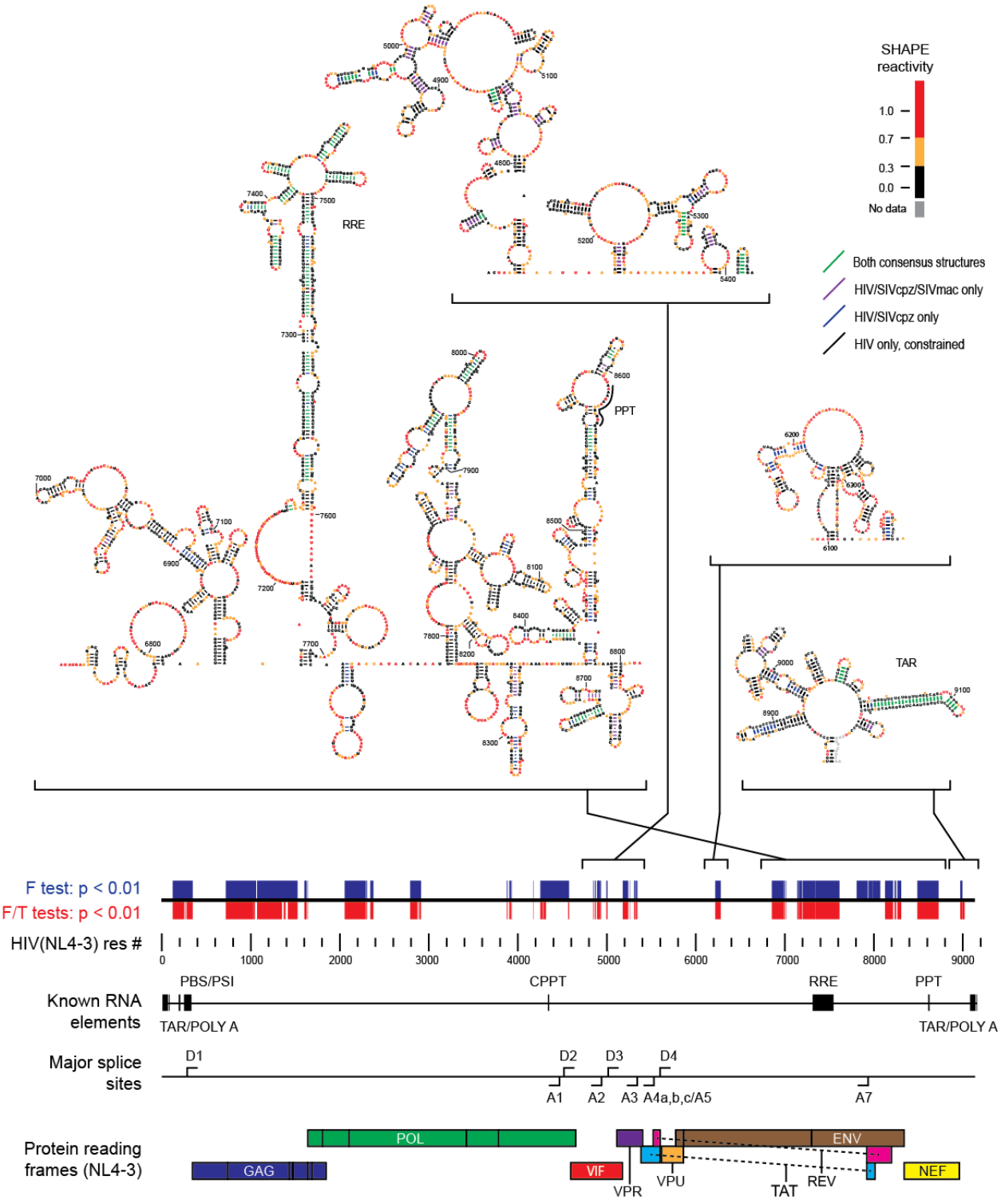



**Figure 5.4** Secondary structure predictions for the last four interdependent elements of the genome-wide alignment as ordered by sequence. Secondary structures shown are for the final constrained HIV secondary structure prediction.

Consensus base pairs are shown on the final constrained HIV model in Figures 5.3 and 5.4. Importantly, all known functional RNA elements are recapitulated in this analysis (Figure 5.5). Consensus base pairs are highly represented in known functional elements, but consensus base pairs are also found in a number of areas with no known function. Generally, consensus base pairs are more represented in the 5' and 3' ends of the genome, though regions with many consensus base pairs exist in the central coding region of genome.

## 5.7 Discussion

### 5.7.1 Relationship with previous sequence comparison analysis

In a previous study by Pollom *et al*., the entire SIVmac (SIVmac239) was characterized by SHAPE-directed structure prediction.[9] This was compared to SHAPE-directed secondary structure predictions for HIV (NL4-3). Based on a strict codon-based alignment, only 71 base pairs were conserved between HIV and SIVmac.[9] Of these 71 base pairs, 22 were found in previously undescribed structures with no known function. In this study, more conserved structures were identified with base pairs. In the predicted consensus for HIV, SIVcpz, and SIVmac genomes, 327 base pairs were found with pairing probabilities greater than 95% that did not conflict with the HIV-SIVcpz consensus prediction. Though these base pairs are highly represented in known functional elements (181 base pairs, 55.4%), many consensus base pairs exist in areas with no known function (146 base pairs, 44.6%).

In this study, which directly considered sequence alignment in structure prediction, a greater number of base pairs in known functional elements were recapitulated, highlighting an increased predictive power. The approach described in this paper does not consider a codon-based alignment. The identification of known functional elements using our strategy indicates that enforcing a strict codon alignment may not necessarily be conducive to RNA structure discovery. Moreover, the occurrence of conserved RNA structures, different from the codon alignment, is consistent with RNA and protein structures evolving independently, even in coding regions.

Despite differences in methodology, similar structural elements were predicted in this work and in work by Pollom *et al*. Pollom *et al.* proposed that a conserved structural element exists at the first splice acceptor site A1 (NL4-3 residue 4458/4459).[9] Perturbation of this stem-loop structure results in modulation of HIV splicing. We recapitulate this predicted structure precisely, and as in the work by

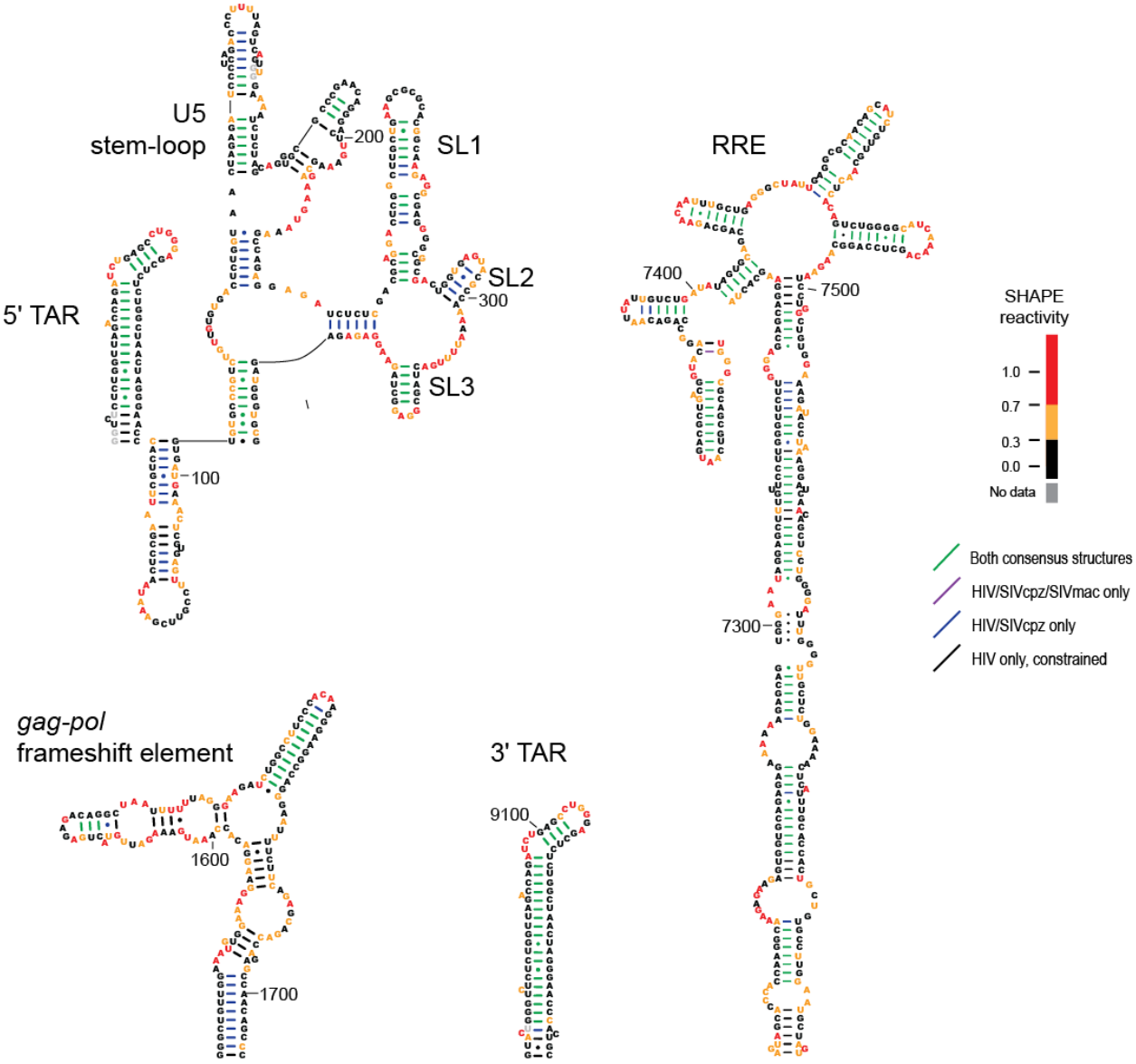



**Figure 5.5** Predicted secondary structures for RNA elements with known function. The secondary structures are from the final constrained HIV prediction with consensus base pairs shown in purple, blue, and green. The structures of all known functional RNA elements are recapitulated in the final HIV secondary structure prediction.

Pollom *et al.*, we see agreement between HIV and SIVmac structures in the form of consensus base pairs at the base of this stem-loop structure (Figure 5.6B).

**5.7.2 Conserved structural elements with no known function**

A number of structural elements are conserved among the three HIV-related strains that have no known function. A connection between protein folding and structure in mRNA has been hypothesized in past studies. Proteins fold co-translationally, and the stability of RNA structure affects ribosomal pausing during translation.[22-24] Changes in local RNA structure have also been shown to modulate protein activity.[25] In past analysis of the HIV genome, a connection was found between the highly structured regions of the RNA genome and protein domains and protein-protein junctions in HIV polyprotein precursors.[8] These highly structured regions were suggested to cause ribosomal pausing, allowing individual protein domains to fold independently.

We predict that a number of structure elements exist at or near protein-protein junctions or at domain boundaries (Figure 5.6A). In *gag*, the junction between p17 (matrix) and p23 (capsid) (residues 731/732) is found within a helical stack in consensus structures (NL4-3 residues 716 to 773). In *gap-pol*, two such structural elements exist. The first is a long helix (starting at NL4-3 residue 2015) in which is nested the junction of protease and reverse transcription proteins (residues 2195/2196); the second contains two helical elements (starting at NL4-3 residue 3753) on either side of the junction between RNase H and integrase proteins (residues 3775/3776). These elements may represent conserved structures that cause ribosomal pausing and allow for the independent folding of protein domains.

A predicted structure element is also found to contain a splice acceptor. Consensus base pairs are highly represented at the junction between the *tat/rev* intron and the second *tat/rev* exon (Figure 5.6B). Splicing at this site is occurs in 87.38% of HIV-1 transcripts in HIV-infected primary $CD4^+$ T cells.[26] The putative structural element is a three-way RNA junction with consensus base pairs prevalent in all three constituent helices (beginning at NL4-3 residue 7879). The tat/rev intron ends (residue 7914) near the junction between first two helices of this element (residues 7915/7916). Though no function in splicing is known for this region, RNA structure has been implicated as a regulator of splicing processes.[27] Combined with the putative element at splice acceptor 1[9] and known structural elements near splice donor 1[28], this structure suggests a regulatory role for RNA structure in HIV splicing.

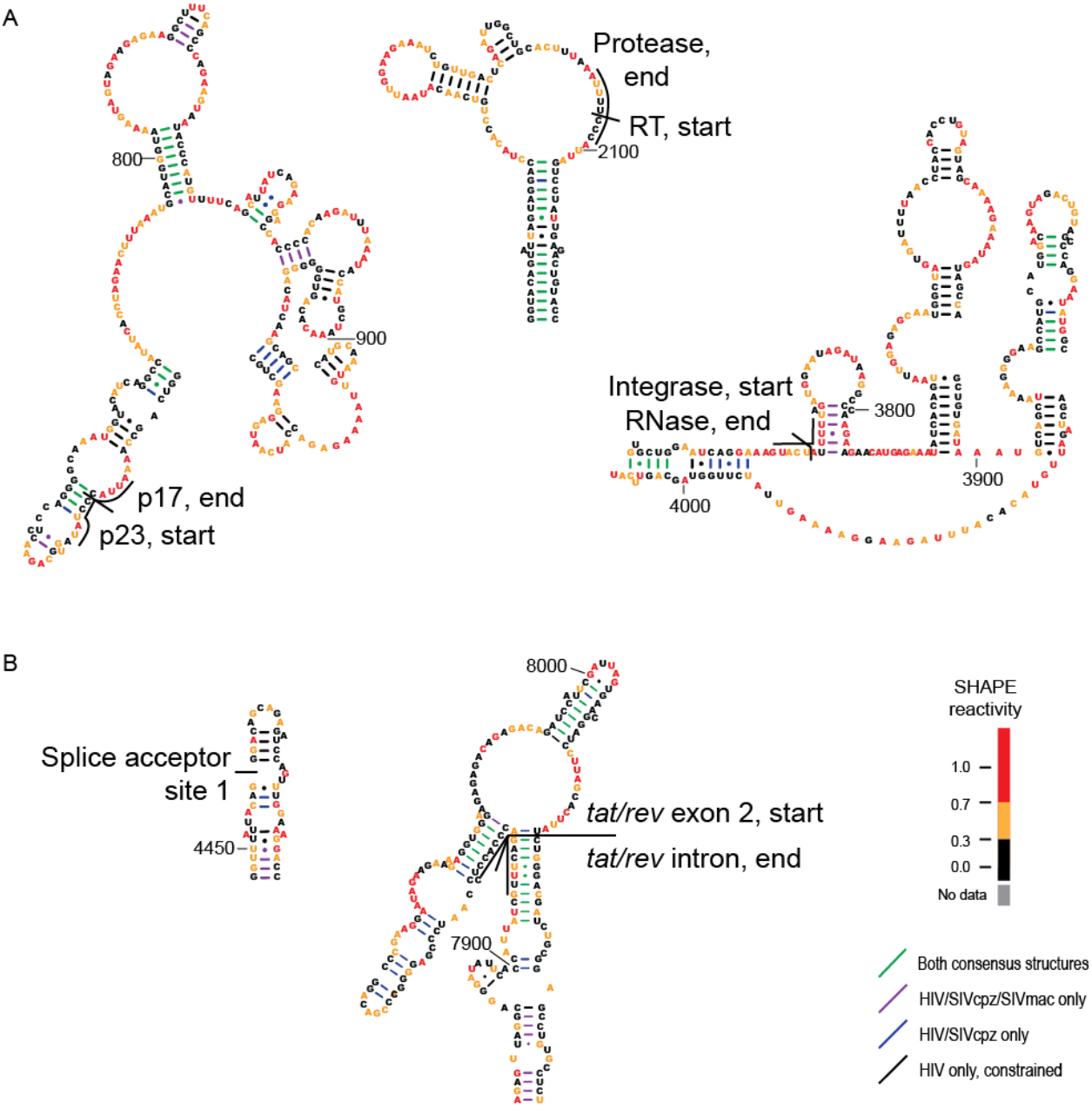



**Figure 5.6** Elements in the final HIV predicted structure with high concentrations of consensus base pairs and no known function. (**A**) Structural elements located in protein domains and at protein-protein junctions. (**B**) Structural elements near splice sites.

### 5.7.3 Structural features common between cPPT- and PPT-containing elements

cPPT- and PPT-containing regions that were statistically interdependent by F-test in whole genome alignments (Figure 5.2). Moreover, both regions contained consensus base pairs at or near the PPT tract: nucleotides 5' of the PPT were predicted to pair with nucleotides immediately 3' of this sequence (Figures 5.3 and 5.4). This result prompted investigation of a possible secondary structure shared by regions containing the cPPT and PPT.

A SHAPE-dependent alignment was first performed for cPPT and PPT sequences from HIV, SIVcpz, and SIVmac, considering six different sequences in total. The boundaries of cPPT- and PPT-containing regions were selected based on the base-pairing arrangement in consensus secondary structures. Because the sequences varied greatly between cPPT and PPT regions, an alignment approach considering only SHAPE reactivities was used. This approach has been found to give precise alignments for structured RNA molecules (see Chapter 4). Despite the fact that base identity was not considered in these alignments, the PPT sequences aligned precisely (Figure 5.7).

RNAalifold was used to predict a consensus secondary structure for the six-sequence alignment. Consensus base-pairs with pairing probabilities above 95% were then used to constrain single-sequence predictions by RNAfold (representative predictions shown in Figure 5.8). cPPT- and PPT-containing regions share two structurally similar elements. The first is a helix comprised of four consensus base pairs. This element is located two nucleotides after the 3' end of the PPT sequence [HIV NL4-3 residues 4347 (cPPT) and 8621 (PPT)]. The second element involves pairing of G residues in the 3' end of the PPT. Though this element does not appear in consensus structures, this may be due to limitations in our secondary structure prediction approach. The approach does not allow for non-canonical base pairs, and non-canonical A-G base pairs in this region are supported by SHAPE reactivities in the HIV PPT structure (denoted by dashed lines in Figure 5.7).

The idea of conserved structures in the cPPT- and PPT-containing regions was proposed in past studies.[9] It is not clear how a conserved intramolecular structure would impact the known function of the PPT tract as an RNA primer for second-strand DNA synthesis. However, elements of the consensus cPPT/PPT structure are conspicuously positioned near the RNase H cleavage site, suggesting a possible

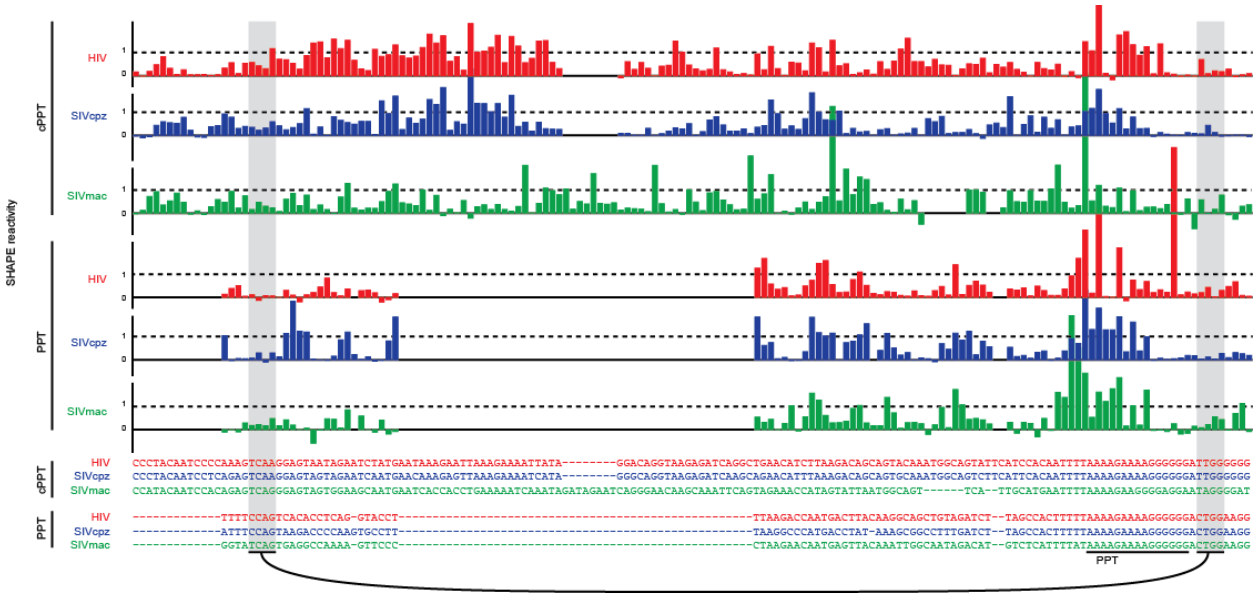



**Figure 5.7** SHAPE-only alignment of cPPT and PPT sequences. SHAPE reactivities and sequence identities are shown for each of the six sequences. Base pairs with pairing probabilities greater than 95% are highlight by gray boxes and shown by connecting lines.

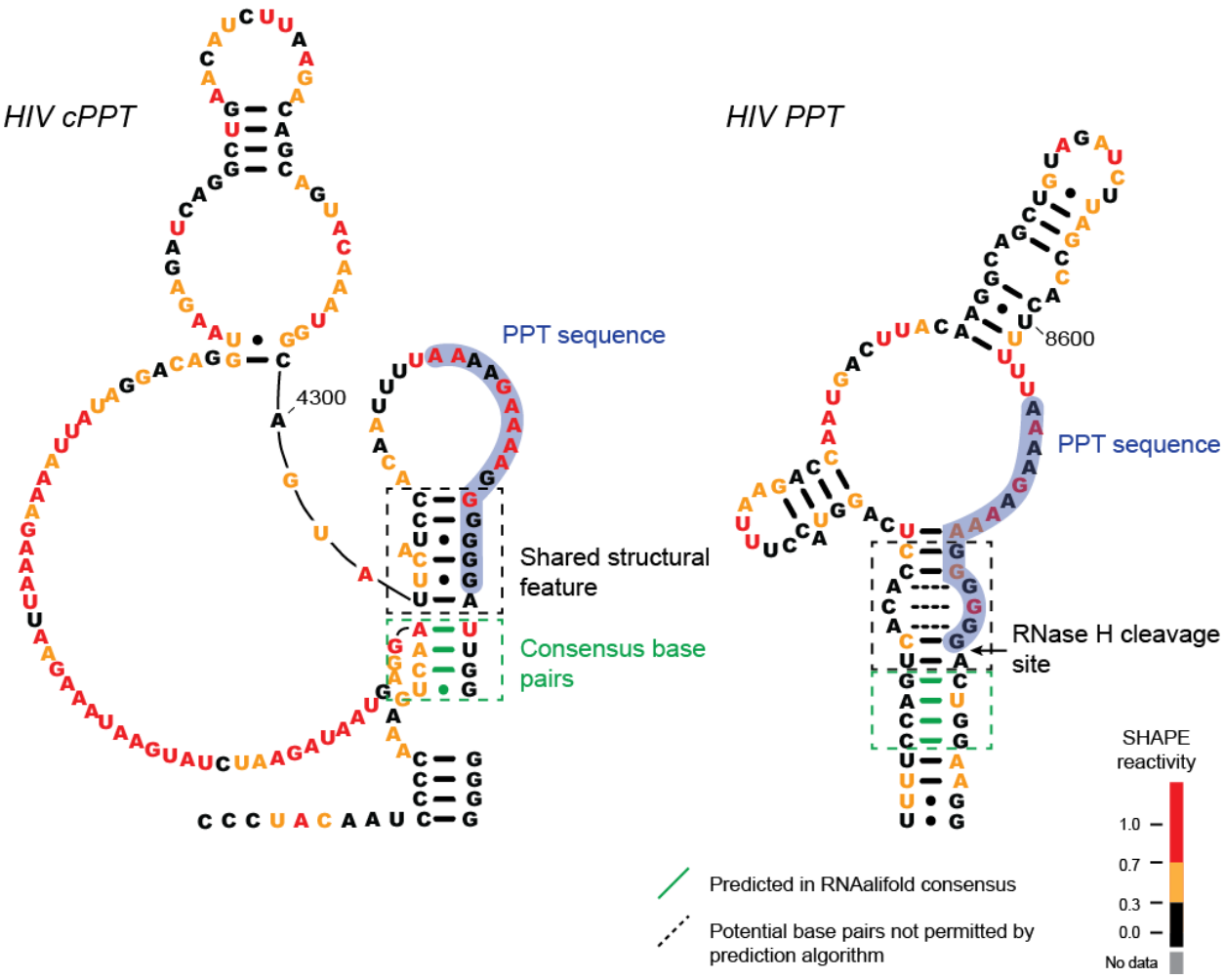



**Figure 5.8** Predicted structures based on cPPT/PPT alignments. Secondary structures shown are for HIV cPPT and PPT regions. Two regions of structural similarity are indicated by boxes: a region with consensus base pairs (green) and a region with similar base pair positions relative to the PPT (black). Both structural elements are present near the RNase H cleavage site (indicated with an arrow).

connection between RNA structure and RNase H recognition. The presence of a conserved RNA structure in these elements may also indicate that PPT-containing regions are multifunctional.

### 5.8 Conclusion

The analysis described in this work revealed regions of HIV genome with structures that are statistically correlated with those in related viral genomes. Secondary structures for these regions were then predicted in consensus-based approaches that considered both sequence covariation and chemical modification data. The resulting structures recapitulated all known functional elements in the HIV RNA genome. Consensus base pairs were also found in structural elements with no known function. Based on the position of these elements in the HIV genome, these new structural elements may regulate co-translational protein folding and RNA splicing.

### 5.9 Experimental Methods

#### 5.9.1 Virus production and genomic RNA purification

Virus was produced and genomic RNA purified as previously described.[8] During genomic RNA extraction, care was taken to avoid denaturation of RNA structure by heat, metal chelation, or chemical denaturants. Following lysis with SDS and proteinase K, viral RNA was extracted using phenol/chloroform. The viral RNA was then precipitated in 70% ethanol with 300 mM KCl and was held at -80 °C until use.

#### 5.9.2 Characterization of genomic RNA by SHAPE-MaP

Tubes containing roughly 10 µg of precipitated SIVcpz or SIVmac RNA in 70% ethanol were spun at 14K RPM at 4 °C for 45 minutes to pellet RNA. Ethanol was removed, and the pellets were held at room temperature for 10 minutes to allow the remaining ethanol to evaporate. The pellets were then resuspended in 20 µL genome resuspension buffer [50 mM HEPES (pH 8.0), 200 mM potassium acetate], and the resulting solution was characterized by absorption spectroscopy to determine RNA concentration.

In order to determine SHAPE reactivities, three different experiments were performed with SIVcpz and SIVmac samples: 1M7 modification of natively-folded RNA, a no-modification background control, and 1M7 modification of denatured RNA. RNA modification followed established protocols (Siegfried, *et al.*, in preparation). For 1M7 modification of natively-folded RNA and no-modification background controls,

aliquots containing 1 µg of SIVcpz or SIVmac RNA were taken from precipitated RNA stocks. To these 1-µg aliquots, 3 µL of 100 mM $MgCl_2$ was added, and the RNA solution was brought up to a volume of 90 µL using genome resuspension buffer. The RNA solution was then incubated at 37 °C for 15 minutes before adding 10 µL of 100 mM 1M7 in DMSO (1M7 modification of natively folded RNA) or 10 µL neat DMSO (background control). The RNA solution was then incubated at 37 °C for 3 minutes, allowing for complete reaction of 1M7. The RNA solution was then held on ice until purification.

For 1M7 modification of denatured RNA, an aliquot containing 1 µg of SIVcpz or SIVmac RNA was taken from precipitated RNA stocks. To this aliquot, 25 µL of 4x denatured control buffer was added [200 mM HEPES (pH 8.0), 16 mM EDTA], and the RNA solution was then brought up to volume of 40 µL using nuclease-free water. To the RNA solution, 50 µL of deionized formamide was added. The RNA solution was held at 95 °C for 1 minute and then added to 10 µL 100 mM 1M7 in DMSO. The reaction was held at 95 °C for 1 minute before transferring the reaction to ice. The RNA solution was held on ice until purification.

RNA from the three SHAPE experiments was then purified using an RNeasy Min-Elute kit (Qiagen). Following purification, sequencing libraries were generated as described previously (see Chapter 4), and sequencing output was analyzed by the SHAPE-MaP pipeline (Siegfried *et al*., in preparation). Background mutation rates were abnormally high for the first 200 nucleotides of the SIVmac genome, resulting in unusual negative peaks for this region. Due to this poor data quality, SHAPE values for the first 200 nucleotides of SIVmac239 were taken from previous work[9]. SHAPE-MaP data for all three genomes was subsequently used in sequence alignments, linear regression analysis, and consensus secondary structure prediction.

**5.9.3 Creation of SHAPE-dependent alignments of genomic RNA**

SHAPE-dependent alignments were performed as described previously (see Chapter 4). Pairwise sequence alignments were generated using a custom dynamic programming-based approach. Subsequently, pairwise sequence alignments were used to create a multiple sequence alignment with T-Coffee.[15]

### 5.9.4 Multi-variable linear regression and statistical analysis

Multi-variable linear regression analysis was performed using NumPy, SciPy, and statsmodels Python modules.[29, 30] Analysis was performed over 200-nt windows. Over each window, only positions with SHAPE values for each genome were considered: No gapped positions were included in analysis. Multi-variable linear models were created using least squares fitting. F-tests and t-tests were performed over each window using the statsmodels module.[30]

### 5.9.5 Selections of areas of interest for secondary structure prediction

Areas of interest for secondary structure prediction were selected based on F-test statistics of multi-variable linear regression models. If a given 200-nt window had an F-test p value less than 0.01, the corresponding 200-nt region was selected as an area of interest. Regions with overlapping areas of interest were considered in the same secondary structure element.

### 5.9.6 Consensus structure prediction using the Vienna software package

The secondary structure of each element selected by linear regression analysis was predicted using RNAalifold and RNAfold, both of the Vienna-RNA software package.[19, 21, 31] The secondary structure prediction was performed on two levels. First, consensus based pairs were generated using RNAalifold. Second, consensus base pairs were used to constrain secondary structure prediction of HIV.

Two consensus secondary structures were predicted for each element. The first consensus considered HIV, SIVcpz, and SIVmac sequences. The second consensus considered HIV and SIVcpz sequences only, with SIVmac removed from the multiple sequence alignment. Consensus structures were generated using RNAalifold, considering the ribosum substitution matrix and a max base pairing distance of 600 nucleotides.[32] Consensus structure prediction incorporated SHAPE reactivities using a pseudo-energy potential.[18]

Following consensus predictions, consensus base pairs were used to constrain an individual secondary structure prediction of HIV. Base pairs from each consensus structure with pairing probabilities greater than 95% were added to a constraint list. The constraint list was curated in order to remove base pairs that disagreed between the two consensus structures. Consensus pairs were excluded if (1) pairs with shared nucleotides contradicted each other in terms of base pairing partners or (2) pairs from two

consensuses were non-nested. In either circumstance, offending pairs were removed from the combined consensus.

HIV structure predictions constrained by consensus pairs were performed with RNAfold. Predictions were constrained such that curated consensus pairs were maintained in the final structure. SHAPE reactivities were incorporated into secondary structure predictions using a pseudo-free energy potential.[18] A maximum base pairing distance of 600 nucleotides was enforced during predictions.

### 5.9.7 SHAPE-only alignment of cPPT and PPT sequences

Regions containing cPPT and PPT sequences were selected based on constrained HIV secondary structure predictions. In predicted cPPT and PPT structures, the 3' end of the PPT was bound in a helix; cPPT and PPT regions of interest were selected using this helical element as a guide. cPPT- and PPT-containing regions of HIV, SIVcpz, and SIVmac were aligned considering only SHAPE reactivities (see Chapter 4).

### 5.9.8 Consensus secondary structure prediction for cPPT/PPT alignment

RNAalifold was used to predict a consensus secondary structure using the resulting cPPT/PPT alignment. RNAalifold predictions were performed considering the ribosum substitution matrix.[32] SHAPE reactivities were incorporated using a pseudo-free energy potential.[18] From this consensus structure, base pairs with pairing probabilities greater than 95% were then used to constrain individual RNAfold structure predictions for the six sequences in the alignment. SHAPE reactivities were incorporated into these individual structure predictions using a pseudo-free energy potential.[18]

## 5.10 References


1. Gesteland, R. F., Cech, T., and Atkins, J. F. (2006) *The RNA World*, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

2. Muesing, M. A., Smith, D. H., and Capon, D. J. (1987) Regulation of mRNA accumulation by a human immunodeficiency virus trans-activator protein, *Cell 48*, 691-701.

3. Hauber, J., and Cullen, B. R. (1988) Mutational analysis of the trans-activation-responsive region of the human immunodeficiency virus type I long terminal repeat, *J Virol 62*, 673-679.

4. Olsen, H. S.*, et al.* (1990) Secondary structure is the major determinant for interaction of HIV rev protein with RNA, *Science 247*, 845-848.

5. Parkin, N. T., Chamorro, M., and Varmus, H. E. (1992) Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo, *J Virol 66*, 5147-5151.

6. Kollmus, H.*, et al.* (1994) The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human T-cell leukemia virus type II in vivo, *J Virol 68*, 6087-6091.

7. Lu, K., Heng, X., and Summers, M. F. (2011) Structural determinants and mechanism of HIV-1 genome packaging, *J Mol Biol 410*, 609-633.

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

9. Pollom, E.*, et al.* (2013) Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs, *PLoS Pathog 9*, e1003294.

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

11. Gao, F.*, et al.* (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes, *Nature 397*, 436-441.

12. Van Heuverswyn, F.*, et al.* (2007) Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon, *Virology 368*, 155-171.

13. Calef, C.*, et al.* (2001) Numbering positions in SIV relative to SIVMM239, In *HIV sequence compendium 2001* (Kuiken, C., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J., Wolinsky, S., and Korber, B., Eds.), Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos.

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

15. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol 302*, 205-217.

16. Charneau, P., Alizon, M., and Clavel, F. (1992) A second origin of DNA plus-strand synthesis is required for optimal human immunodeficiency virus replication, *J Virol 66*, 2814-2820.

17. Swanstrom, R.*, et al.* (1981) Nucleotide sequence of cloned unintegrated avian sarcoma virus DNA: viral DNA contains direct and inverted repeats similar to those in transposable elements, *Proc Natl Acad Sci U S A 78*, 124-128.

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

19. Hofacker, I. L., Fekete, M., and Stadler, P. F. (2002) Secondary structure prediction for aligned RNA sequences, *J Mol Biol 319*, 1059-1066.

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

21. Schuster, P.*, et al.* (1994) From sequences to shapes and back: a case study in RNA secondary structures, *Proc Biol Sci 255*, 279-284.

22. Komar, A. A. (2009) A pause for thought along the co-translational folding pathway, *Trends Biochem Sci 34*, 16-24.

23. Farabaugh, P. J. (1996) Programmed translational frameshifting, *Microbiol Rev 60*, 103-134.

24. Wen, J. D.*, et al.* (2008) Following translation by single ribosomes one codon at a time, *Nature 452*, 598-603.

25. Nackley, A. G.*, et al.* (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure, *Science 314*, 1930-1933.

26. Ocwieja, K. E.*, et al.* (2012) Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing, *Nucleic Acids Res 40*, 10345-10355.

27. McManus, C. J., and Graveley, B. R. (2011) RNA structure and the mechanisms of alternative splicing, *Curr Opin Genet Dev 21*, 373-379.

28. Harrison, G. P., and Lever, A. M. (1992) The human immunodeficiency virus type 1 packaging signal and major splice donor region have a conserved stable secondary structure, *J Virol 66*, 4144-4153.

29. Oliphant, T. E. (2007) Python for scientific computing, *Computing in Science & Engineering 9*, 10-20.

30. Seabold, J. S., and Perktold, J. (2010) Statsmodels: Econometric and Statistical Modeling with Python, *Proceedings of the 9th Python in Science Conference*.

31. Lorenz, R.*, et al.* (2011) ViennaRNA Package 2.0, *Algorithms Mol Biol 6*, 26.

32. Klein, R. J., and Eddy, S. R. (2003) RSEARCH: finding homologs of single structured RNA sequences, *BMC Bioinformatics 4*, 44.