HIGH-THROUGHPUT ANALYSIS OF RNA TERTIARY STRUCTURE AND INTERACTIONS

Philip John Homan

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

Chapel Hill

2013

Approved by:

Kevin Weeks

Marcey Waters

Dorothy Erie

Oleg Favorov

Nikolay Dokholyan

ABSTRACT

Philip John Homan: High-throughput analysis of RNA tertiary structure and interactions
(Under the direction of Kevin M. Weeks)

The many important cellular functions of RNA molecules depend on formation of complex tertiary structures. Knowledge of the specific interactions that stabilize these structures is key to understanding the function of the RNA. Current methods to study RNA tertiary structures are limited. Biophysical methods that measure RNA structure produce high-resolution tertiary structures but are limited by the size of the RNA. Current biochemical methods are difficult to interpret and can only give an average view of all possible structures in solution. In this work I develop two new biochemical techniques that can be used to map tertiary interactions within RNA. First I develop a method in which I blend the principals from modification interference experiments with SHAPE chemistry called 2'-hydroxyl molecular interference (HMX). With this approach I am able to measure structurally crowded regions of an RNA and incorporate this data as experimental constraints in discrete molecular dynamics simulations to yield experimentally informed, three-dimensional models. This method was developed with a small test set of RNA and applied to the *Tetrahymena* group I intron to identify interactions between multiple substructures in the RNA. Second I develop a method termed RING-MaP in which I utilize multi-nucleotide sequencing to detect multiple chemical modifications in a single RNA molecule in order to identify correlated structural interactions between modified nucleotides. With this technique I am able to map through-space interaction networks in RNA of increasing size. Furthermore,

through the use of spectral clustering analysis, I am able to identify multiple structural conformations in a single, in-solution ensemble of RNA.

To my family, friends, and loving wife Stephanie

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Kevin Weeks, for giving me the opportunity to work on this project and who has helped me develop into the scientist I am today.

I'd also like to thank Weeks group members both past and present, for making UNC such a great place to work. I could not ask to be among a better group of scientists and friends, I will truly miss you all.

Finally, I'd like to thank my family and friends. I really appreciate the love, support, and prayers you have given me over the years. I am thankful for my parents, Eric and Nancy as well as my siblings Robert Sarah and Emily for all the support they have given over my many years of schooling. And lastly I would like to thank my loving wife Dr. Stephanie Homan. You have made these last three years of grad school really special with your never-ending love and support, as we have been able to share this adventure together.

TABLE OF CONTENTS

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xix

LIST OF SYMBOLS xxii

CHAPTER 1. Introduction 1

1.1. INTRODUCTION 1

1.1.1. RNA structure and function 1

1.1.2. Probing RNA structure 2

1.1.3. Detection of RNA adducts 4

1.1.4. DMD Modeling with biochemical constraints 5

1.2. RESEARCH OVERVIEW 5

1.3. PERSPECTIVE 7

1.4. REFERENCES 8

CHAPTER 2. RNA tertiary structure analysis and refinement by 2'-hydroxyl molecular interference 11

2.1. INTRODUCTION 11

2.2. RESULTS 13

2.2.1. HMX overview 13

2.2.2. Molecular overlap model for HMX intensities. 21

2.2.3. Three-dimensional RNA structure modeling. 23

2.3. DISCUSSION. 26

2.4. METHODS 28
2.4.1. RNA constructs. 28
2.4.2. 5'-[^{32}P] RNA radiolabeling. 28
2.4.3. RNA modification for molecular interference. 29
2.4.4. RNA folding and structural partitioning. 29
2.4.5. Reverse transcription and adduct detection. 30
2.4.6. Calculation of the HMX score. 30
2.4.7. Modeling of adduct disruption of native RNA tertiary structure. 32
2.4.8. HMX-directed structure refinement by DMD. 32
2.4.9. Replica exchange DMD simulations and consensus structure modeling. 33
2.5. REFERENCES 35
CHAPTER 3. HMX reveals tertiary interactios within multiple stable substructures of the *Tetrahymena* group I intron 38
3.1. INTRODUCTION 38
3.2. RESULTS 39
3.3. DISCUSSION 44
3.4. METHODS 46
3.4.1. RNA constructs 46
3.4.2. 5'-[^{32}P] RNA radiolabeling 46
3.4.3. Denatured RNA modification 47
3.4.4. RNA folding and native gel separation 47
3.4.5. Reverse transcription and adduct detection 48
3.5. REFERENCES 49
CHAPTER 4. tertiary structure revealed by single-molecule correlated chemical probing of rna 51
4.1. INTRODUCTION. 51

4.2. RESULTS 53
4.2.1. Multi-site DMS reactivity with RNA. 53
4.2.2. Through-space RNA interactions detected by statistical association analysis. 58
4.2.3. Through-space RNA interactions detected by statistical association analysis. 59
4.2.4. Multiple RNA conformations detected by spectral clustering. 61
4.3. DISCUSSION. 69
4.3.1. Principles of RNA folding. 69
4.3.2. Three-dimensional RNA structure refinement. 71
4.3.3. Perspective. 75
4.4. METHODS 75
4.4.1. Characterization of reaction between dimethyl sulfate and RNA nucleobases. ... 75
4.4.2. RNA constructs. 76
4.4.3. RNA folding and DMS modification. 77
4.4.4. Reverse transcription and adduct detection. 77
4.4.5. Measurement of inter-nucleotide interactions by statistical association analysis. 78
4.4.6. Spectral clustering of multiple conformations in a single RNA ensemble. 81
4.4.7. Estimating relative fractions of different conformations in RNA sample. 87
4.4.8. Reconstructing modification frequency profiles of individual conformations in an RNA sample. 89
4.4.9. Three-dimensional RNA structure modeling. 91
4.5. GUIDELINES FOR OF RING-MAP DATA ANALYSIS 93
4.5.1. Qualifying data used in RING-MaP calculations 93
4.5.1.1. Number of mutations per sequence. 93
4.5.1.2. Mutation frequency of samples. 94
4.5.1.3. Read depth 97
4.5.2. RING analysis 98

4.5.3. Spectral clustering Analysis 99
4.5.3.1. Eigengap value analysis 100
4.5.3.2. Fraction of the minor cluster 100
4.5.3.3. Difference in cluster profiles 100
4.5.3.4. Analyzing the TPP riboswitch. 101
4.5.3.5. Additional criteria for evaluating data for spectral clustering. 103
4.6. ACKNOWLEDGEMENTS 103
4.7. REFERENCES 104

LIST OF TABLES

Table 2.1 DMD simulation statistics for the four RNA fold refinements. Cluster populations (*n*), mean RMSDs, cluster energies, and *p*-values for each structure are shown. 24

Table 4.1 Summary of spectral clustering analysis for multiple RNA conformations in single ensembles. Clustering analysis is summarized for the TPP riboswitch, P546 domain, P546 mutants, and the RNase P RNA as a function of different levels of structure. The eigengap value measures the structural difference between clusters; samples with eigengaps greater than 0.03 are taken to have two (or more) distinct clusters. The population of each cluster is given in the last column with the most highly structured cluster listed first. An asterisk indicates that smallest cluster population for the TPP riboswitch, at saturating ligand concentration, was too small to accurately generate DMS reactivity profiles. For analysis, the sample was therefore clustered into two conformations with populations of 81% and 19%. 62

Table 4.2. Summary of processed massively parallel sequencing data for RNA nucleotide association analysis and spectral clustering. Clustering analysis requires reads that have two or more mutation events. The fourth column lists the number of nucleotides with mutation frequencies greater than 0.01 used in spectral clustering analysis. 98

LIST OF FIGURES

Figure 2.1 2'-Hydroxyl molecular interference (HMX). (**A**) RNA is modified under denatured conditions such that all nucleotides have a significant probability of being modified. Some 2'-hydroxyl adducts prevent native folding, creating a population of unfolded RNA that can be partitioned from fully folded RNA. In this work, partitioning was performed by non-denaturing gel electrophoresis. (**B**) Structure of the 2'-*O*-ester adduct introduced by reacting RNA with NMIA. (**C**) Partitioned populations were separately subjected to primer extension to detect adducts. Positions with high intensities in the unfolded RNA have low intensities in the folded RNA and indicate positions of adducts that prevent folding. HMX profiles were calculated using a cross-correlation analysis. 14

Figure 2.2 Electropherograms of 2'-*O*-ester modified and unmodified RNAs. Each RNA was modified with NMIA under denaturing conditions. Modifications were detected as stops to reverse transcriptase-mediated primer extension. 15

Figure 2.3 Partitioning of RNA populations by native gel electrophoresis. (**A**) Folded and unfolded populations for modified and unmodified RNAs were separated by non-denaturing gel electrophoresis in the presence of 50 mM NaCl and 5 mM $MgCl_2$. For clarity, gel images were straightened and scaled to show similar representations for each RNA; band profiles and intensities were not altered. (**B**) Band intensities as a function of gel migration distance. 17

Figure 2.4 Calculation RNA HMX scores by normalization and cross-correlation. (**A**) Electropherograms of unfolded RNA (green) scaled to data for folded RNA populations (black). Positions with high intensities in the unfolded RNA have low intensities in the folded RNA. (**B**) Cross-correlation normalization, based on both unfolded and folded 2'-*O*-adduct profiles, was used to create HMX score profiles. Positions with low, medium, and high HMX scores are black, orange, and red, respectively. Experiments were performed in triplicate and error bars are shown with black lines. 19

Figure 2.5 Visualization of HMX interference information on accepted three-dimensional structures.[15-18] The 2'-OH group for each nucleotide is shown as a sphere and the phosphate backbone by a tube. Nucleotides are colored by HMX score; the TPP ligand is green. 20

Figure 2.6 Physical model for 2'-hydroxyl molecular interference. (**A**) Model for interference by molecular overlap in which adducts are represented by a pseudo-atom (grey) at a distance (L) from the O2' position at radius (r). The degrees to which surrounding atoms intersect the pseudo-atom

were estimated by calculating van der Waals radii overlaps. (**B**) Analysis of optimal pseudo-atom bond length and atomic radius. Maximum correlation between Pearson's *r* and pseudo-atom representing 2'-hydroxyl molecular interference is boxed. (**C**) Relationship between pseudo-atom dimensions and 2'-*O*-ester adduct. (**D**) Representative relationships between HMX scores and molecular overlap for the M-Box and P546 domain RNAs. HMX score profiles (red) show a high correlation with calculated molecular overlaps (black) for each RNA. Pearson correlation coefficients are shown. Correlation coefficients for the $tRNA^{Asp}$ and TPP riboswitch RNAs (not shown) were 0.60 and 0.72, respectively. 22

Figure 2.7 HMX-directed RNA fold refinements. RNAare shown as backbone traces. Accepted structures[15-18] and HMX-directed refinements for each RNA are gray and blue, respectively. The cluster populations (*n*), mean RMSD, and *p*-values[25] are shown. For $tRNA^{Asp}$, both the largest cluster (large image) and lowest RMSD structures (inset) are shown. 25

Figure 3.1 Identification of multiple domains with in the group I intron by ShapClash. (**A**) Folded and unfolded along with intermediate fold populations for modified and unmodified RNA were separated on a 0.5x TB polyacrylamide gel containing 50 mM NaCl and 5 mM $MgCl_2$. For clarity, gel images were straightened and scaled to provide similar representations for each RNA. (**B**) Band intensities were measured for each lane. The intensity of the unfolded and intermediate bands was larger in the modified sample compared to the unmodified sample for each RNA. (**C**) Correlation between HMX data and calculated sphere overlap related to RNA structure. Cross correlation normalized HMX score profiles (red) show strong correlation when compared to calculated sphere overlap (black) for the group I intron. Sections of the RNA that are not present in the crystal structure are shown as gas in the calculated sphere overlap trace. 40

Figure 3.2 Analysis of intermediate structures in Group I intron. HMX score profiles were created for the (**A**) first and (**B**) second intermediate bands as well as for the (**C**) unfolded band. Positions with low reactivity below 0.3 are colored in black; positions of medium reactivity (between 0.3 and 0.6) are colored in orange; and positions of high reactivity (greater than 0.6) are colored in red. Reactive positions in the first intermediate profile are marked by green circles and. Positions that become reactive in the second intermediate and unfolded profile are marked in purple and blue respectively. (**D**) Highlighted positions on the HMX profiles are superimposed on the secondary structure representation of the crystal structure.[15] 43

Figure 4.1 Single-molecule RNA structure analysis by massively parallel sequencing. (**A**) RNA molecules experience local structural variations and 'breathing' in which regions of an RNA structure become reactive to a chemical probe in a correlated way. Nucleotides that interact (open red circles) show correlated reactivity. Statistical association analysis is used to detect and quantify the strengths of these interdependencies, ultimately revealing multi-point RNA interaction groups or RINGs. (**B**) In solution, RNAs often adopt multiple conformations. Spectral clustering analysis based on similarity of nucleotide reactivity patterns was used to separate data on individual RNA stands into different conformations. 52

Figure 4.2 Efficient DMS adduct formation at the base-pairing faces of adenosine and cytosine. (**A**) Reaction of radioactively labeled nucleotides with 170 mM DMS in 300 mM cacodylate (pH 7) monitored by gel electrophoresis. (**B**) Time-course of DMS reaction with adenosine and cytosine. Unconstrained nucleotides react to form methyl adducts with ~12% efficiency (arrow). 54

Figure 4.3 Optimization of DMS adduct formation. The pH of the DMS reaction (blue lines) and adduct formation with adenosine (red) were monitored as a function of time. The DMS concentration was 170 mM. (**A**) In 100 mM HEPES (pH 8.0), pH dropped over time, quenching the DMS reaction. These conditions closely resemble those widely employed in conventional DMS experiments. (**B**) Reactions performed in 300 mM HEPES (pH 8.0) limited the pH drop; however, the organic buffer reacted directly with DMS, quenching the reaction with adenosine. (**C**) The pH in reactions performed in 300 mM cacodylate was well-controlled, and the buffer did not react with DMS to quench the reaction. 55

Figure 4.4 RING analysis of RNA structure. (**A**) Number of mutations per transcript detected by reverse transcription with (red) and without (black) DMS modification. (**B**) DMS modification induced mutation frequencies as a function of nucleotide position. Data from DMS-treated samples are shown in red and no-reagent controls are shown in black. (**C**) RINGs for the TPP riboswitch, P546 domain, and RNase P RNAs showing strong (green) and moderate (yellow) correlations. Correlations occur between positions that are reactive in the native structure (filled red circles) or become reactive during 'breathing' motions (open circles), reflecting the structural breathing component of reactivity interdependencies. Correlation coefficients of 0.025 and 0.035 correspond to median increases in correlated mutations of 2.5- and 2.8-fold, respectively (Fig. 4.12C). Secondary structures are drawn to approximate relative helical orientations in three-dimensionalspace based on known structures[9-11]. 57

Figure 4.5 RINGs report the tertiary structure of the P546 domain and mutation-induced structural changes. Strong and medium internucleotide correlations are shown with green and yellow lines, respectively. (**A**)

RINGs in the P546 domain folded in the presence of Mg2+. (**B**) RINGs in the P546 domain folded in the absence of Mg2+. (**C**) RINGs in the P6a mutant. (**D**) RINGs in the J5 hinge mutant. For clarity, panel A is identical to Figure 4.4C..... 60

Figure 4.6 RINGs and clustering analysis of the TPP riboswitch in the presence and absence of TPP ligand. RING analysis in the presence of (**A**) saturating ligand and (**B**) absence of ligand. Strong and moderate internucleotide associations are shown with green and yellow lines, respectively. Nucleotides that are less or more structured in the minor, less populated, cluster are emphasized with open and closed spheres, respectively. Spectral clustering analysis in the (**C**) presence of saturating ligand and (**D**) absence of ligand. There are two clusters in each state. In the presence of saturating ligand, the major cluster (red) corresponds to the fully folded riboswitch. In the absence of ligand, the major cluster (red) reflects an unstructured state with few internucleotide interactions. The minor cluster (blue) in the saturating ligand sample is more unstructured than the major cluster and is similar to the no-ligand structure (gray). The minor cluster (blue) in the no-ligand sample is more highly structured than the major cluster specifically in the region of the thiamine binding pocket (blue closed circles). 64

Figure 4.7 Spectral clustering analysis of the TPP riboswitch at sub-saturating ligand concentration of 200 nM ligand. (**A**) RING analysis of internucleotide association interactions. Interactions are fewer in number and weaker than those for the RNA under saturating ligand conditions (compare with Fig. 4.6A). (**B**) Three clusters were identified with population fractions of 32, 31, and 37% (blue). Each of these clusters corresponds to a state identified in either saturating ligand concentration or in the absence of ligand (red) with nucleotides corresponding to the ligand bound or no ligand structures (gray). 66

Figure 4.8 Clustering analysis of the RNase P domain RNA in the presence and absence of Mg^{2+}. (A) RING analysis of the RNA structure in the presence of Mg^{2+}. (**B**) RING analysis in the absence of Mg^{2+}. (**C**) Separation of the plus- Mg^{2+} data into two clusters. The minor cluster (blue) is characterized by a subset of nucleotides (blue circles) that are more reactive (and thus less structured) than those in the major cluster structure. Positions more reactive in the minor cluster mediate the L5-L15.1 loop-loop tertiary interaction and form the structural core. In most regions, the no- Mg^{2+} state has a RING pattern that is structurally distinct from both of the plus- Mg^{2+} states. In contrast, the P19 element shows the same RING pattern as was observed in the presence of Mg^{2+} suggesting that this region folds independently and is not stabilized by Mg^{2+}..... 68

Figure 4.9 Through-space RNA structural relationships revealed by RINGs. (**A**) Direct, through-hhelix, and global internucleotide interactions are

illustrated on both secondary structures (*top*) and three-dimensional models[9-11] (*bottom*). (**B**) Three-dimensional models determined for each RNA using RING interdependencies as constraints. The *p*-values report the significance of each model; the secondary structure was input during refinement[19]..... 70

Figure 4.10 Long-range constraints for using RING interdependencies to refine three-dimensional RNA structure models. (**A**) Distribution of distances corresponding to nucleotide associations with correlation coefficients greater than 0.025. (**B**) Histogram summed over all observations. Smooth curve corresponds to the normal distribution based on the average and standard deviation. (**C**) Interaction potential for RING-based distance constraints. (**D**) Radius of gyration based filtering of structure models. Representative histograms of radii of gyration for models of the P546 domain RNA generated during unbiased simulations (blue) or simulations biased by RING data (red). The fit log-normal distribution for the bias-dependent collapsed state is shown with a dashed line..... 73

Figure 4.11 Comparison of the RING-directed three-dimensional model for the RNase P RNA with that based on the crystallographically visualized structure[11]. The core accepted structure (right) excludes helices P3-P2-P19 which folds independently (see Fig. 4.4). The p-values report the significance[19] of each model relative to the accepted structure. 74

Figure 4.12 Visualization of internucleotide associations in the presence and absence of ligand for the TPP riboswitch RNA. (**A, B**) Heat map showing positive statistically significant ($\chi^2 > 20$) nucleotide associations for the TPP riboswitch in (**A**) the presence and the (**B**) absence of the ligand. (**C**) Illustration of relationship between correlation coefficient and the effect that mutation of one nucleotide has on the probability of mutation of the other nucleotide in the pair. Such an effect is measured as a ratio of conditional probabilities [$R_p = P_{(A = 1 \mid B = 1)} / P_{(A = 1 \mid B = 0)}$], and it is plotted for statistically significant ($\chi^2 > 20$) positive associations as a function of correlation coefficient. In the RING analysis, we focused on nucleotide pairs with correlation coefficients greater than 0.025, corresponding to a median ratio of conditional mutation probabilities, R_p, greater than 2.50. 80

Figure 4.13 Spectral clustering of RING-MaP data obtained from analysis of the TPP riboswitch RNA. (**A, B**) A synthetic data sample of a known conformational composition was created by combining two sets of sequencing reads obtained after modification of the TPP riboswitch RNA: one set (**A**) of 50,000 reads was obtained in the presence of saturating TPP ligand, and the other set (**B**), also of 50,000 reads, was obtained in the absence of TPP ligand. Note that these two sets have distinctly different modification frequency profiles, reflecting their different RNA conformations. Any conformational variations that

happened to be present within the either set were deliberately destroyed by randomly shuffling the recorded instances of nucleotide modifications among the reads. Such shuffling, performed independently for every nucleotide in each set, preserved the total number of modifications to a given nucleotide in a given set, but made co-occurrences of modifications among nucleotides statistically random. Thus, this data sample has only two conformations, present at 50:50 ratio. This synthetic data sample is representative of data collected in single experiments, since nucleotide modifications occur on each strand independently of other strands in the RNA pool. (**C**) Eigenvalues of the normalized graph Laplacian matrix, L_{NCut}, of the similarity matrix, S, constructed for the 43 reactive adenosine and cytosine nucleotides of the synthetic TPP RNA data sample. The first eigenvalue, λ1, is always zero, whereas the magnitudes of successive eigenvalues reflect inversely the effectiveness of each successive normalized graph cut. If the data points form two distinct clusters, the first graph cut will be most effective, cutting the links between the points lying in different clusters, thus cleanly separating the clusters. Indeed, in the plot the largest difference between successive 84

Figure 4.14 Examples of eigengap plots containing either three (top row) or four (two bottom rows) clusters in varying proportions. These plots were generated using artificial datasets, derived from a hypothetical RNA containing 50 adenosine and cytosine nucleotides with varied frequencies of modification comparable to those observed in real RNAs. For each cluster of reads, nucleotides were assigned their modification probabilities at random. 86

Figure 4.15 Determination of the conformational identity of individual RNA strands. (**A**) The second eigenvector,, of the synthetic data set created for the TPP riboswitch as described in Fig. 4.13, was chosen because the second eigengap indicates that these data are split into two clusters (see Fig. 4.13D). This eigenvector has 43 values, corresponding to 43 adenosine and cytosine nucleotides with high modification rates. For each nucleotide, the eigenvector magnitude specifies how strongly that nucleotide is associated with either of the two detected clusters. (**B**) Distribution of the scores for all strands computed from the second eigenvector. The strands that belong to one cluster are on the left side of the score distribution, whereas the strands that belong to the second cluster are on the right side of the distribution. The classification boundary was determined using *K*-means clustering; red and blue arrows indicate centroids of each cluster. The ratio of strands in the two clusters is 52:48, which is in good agreement with the true ratio of 50:50 for synthetic dataset. (**C**) Estimated relative fractions of the ligand-bound and unstructured conformations plotted as a function of their true relative fractions. The plot was generated by performing spectral and *K*-means clustering on eight different synthetic datasets with different proportions of strands belonging to the two TPP RNA conformations. 88

Figure 4.16 Reconstruction of mutation probability profiles of individual conformations for data samples containing two conformations each. The two conformations were represented by sequencing reads obtained after modification of the TPP riboswitch RNA in the presence of saturating TPP ligand (ligand-bound conformation) or in the absence of TPP ligand (unstructured conformation), respectively. Any conformational variations within the either set were intentionally destroyed, by randomly shuffling the recorded instances of nucleotide modifications among the reads (see Fig. 4.13, legend). To make data samples, the ligand-bound and unconstrained sets of reads were combined at 50:50, 20:80 or 80:20 ratios. (**A**) The true mutation probability profile (blue) and the profile estimated from spectral and *K*-means clustering (red) for the ligand-bound conformation making up 50% of the data sample. (**B**) The true and estimated mutation probability profiles for the unstructured conformation making up 50% of the data sample. (**C**) The true and estimated mutation probability profiles for the ligandbound conformation making up 20% of the data sample. (**D**) The true and estimated mutation probability profiles for the unconstrained conformation making up 20% of the data sample. The 80% fractions are not shown because the true and estimated profiles are nearly identical (as expected). 90

Figure 4.17 Increasing phred score for counting mutations eliminates base call errors from sequencer. (**A**) Reactivity plot of DMS modified RNA (Red) and no modification control (Black) with the phred score set at 10 produces increased average mutation rate at 5' end. (**B**) RING of DMS modified RNA processed with phred score of 20. High mutation rate due to insturment error causes spurious correlations unrealted to RNA structure. (**C**) Increasing phred score to 20 for counding mutations produces reactivity profile with decreased average mutation rate at 5' end of trace. (**D**) RING network of same DMS modified sample with increased phred score eliminates spurious correlations. RINGs now accuratly represent structural interactions in RNA. 96

Figure 4.18 Determining number of clusters in TPP samples using eigengap values and cluster profiles. (**A**) Eigengaps of saturating ligand, no ligand and no Mg^{2+} indicate that samples have 2,1 and 0 clusters respectivly. How well samples are seprated can be determined by how above or below the eigengap threshold egengap values are. Samples with multiple positons above the egiengap threshold determine number of clusters. (**B**) reactivity profiles for culsters derived from TPP samples. The major cluster (Red) is compaired to minor cluser (Blue). The more difinitivly separated clusters have positions in which are preferentially represented in a single cluster (blue circles). The poorley separated, no Mg^{2+} sample produces two clusters with few major differences between reactive positions. 102

LIST OF ABBREVIATIONS

2-ME	2-Mercaptoethanol
2'-OH	2′-hydroxyl
A	adenine
ATP	Adenosine triphosphate
Asp	aspartate
C	cytosine
cDNA	complementary deoxyribonucleic acid
Ci	curie
CTP	cytosine triphosphate
CPM	counts per minute
DMD	discrete molecular dynamics
DMS	dimethyl sulfate
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DTT	dithiotreitol
EDTA	ethylenediaminetetraacetic†acid
FASTQ	sequence file format
G	guanosine
H	hour
H_2O	water
HEPES	N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic†acid

HIV	human immunodeficiency virus
HMX	2'-hydroxyl molecular interference
HRP	hydroxyl radical probing
Hrs	hours
M	molar
Mg^{2+}	magnesium ion
$MgCl_2$	magnesium chloride
Min	minute
mM	millimolar
mRNA	messenger RNA
μg	microgram
μL	microliter
μM	micromolar
NAIM	nucleotide analog interference mapping
NaCl	sodium chloride
NMIA	N-methyl isatoic anhydride
NMR	nuclear magnetic resonance
nt	nucleotide
NTP	nucleoside triphosphate
^{32}P	phosphorus-32
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PDB	protein data bank

pH	potential of hydrogen
Pmol	picomol
PNK	PNK
RING-Map	RNA interaction groups identified by mutational profiling
RMSD	root mean square deviation
RNA	ribonucleic acid
RNase	ribonuclease
RSA	relative solvent accessibility
RT	Reverse Transcriptase
Sec	second
SHAPE	selective 2'-hydroxyl acylation analyzed by primer extension
TBE	90 mM Tris-borate, 2 mM EDTA
TE	10 mM Tris (pH 7.5), 1 mM EDTA
Tris	tris(hydroxymethyl)aminomethane
TPP	thiamine pyrophosphate
tRNA	transfer RNA
$tRNA^{asp}$	aspartate-tRNA
TTP	Thymidine triphophate
U	uridine
UTP	uridine triphosphate
V	volt
W	watt
w/v	weight per volume

LIST OF SYMBOLS

≈	approximately
Å	angstrom
°C	degree Celsius
D	diagonal matrix
H	hit matrix
K_d	dissociation constant
L_{NCut}	Laplacian matrix
ρ	phi coefficient
r	pearson coeeficent
R_p	ratio of conditional mutation probabilities
S	similarity matrix
n	nucleotide
χ_{Yates}	Yates' corrected version of Pearson's chi-squared test
λ	eigenvalue

CHAPTER 1. INTRODUCTION

1.1. INTRODUCTION

1.1.1. RNA structure and function

RNA serves dual roles as both an integral carrier of genetic information at the primary nucleotide sequence level and, through the formation of higher-order structures, can serve as an important bimolecular machine that is involved in almost every biological process in the cell.[1] For example RNA can participate protein translation, mRNA splicing and regulation of gene expression.[2-4] Critical to the function of these RNA is its ability to fold back on itself and form complex three dimensional structures.[5] RNA three dimensional structures can vary both in their complexity and size, ranging from small tRNAs to large ribosomal RNAs.[6-8] Determining the structure of an RNA is key to a greater understanding of its function.[1,5]

Many functional RNAs form multiple stable conformations with different biological activities; it is often the ability to transition between these stable conformations that leads to the biological activity of the RNA.[4,9] The structural diversity of a single RNA thus poses an additional challenge when experimentally determining an RNA structure. Current experiential methods only target a single conformation or must measure an ensample of RNA structures to get an average view of the RNA structure. The ability to accurately separate and analyze distinct conformations in an pool of RNA structures is an experimental objective that to date has not been accomplished by ensemble studies or conventional single-molecule

studies.

1.1.2. Probing RNA structure

Various methods are currently used to probe and understand RNA structures. An important first step towards understanding the function of an RNA is determination of the base-pairing partners involved in the RNA secondary structure. This problem has largely been solved through the use of small molecules that react with specific RNA functional groups. Traditional RNA chemical probing uses reagents that probe RNA by modifying nucleotides in a base specific manner and so require the use of complementary reagents to comprehensively probe RNA structure.[10-12] The relatively recent development of Selective 2'-Hydroyl Acylation analyzed by Primer Extension (SHAPE) chemistry allows all nucleotides in an RNA to be probed in a single experiment through the use of reagents that target the 2'-OH on the ribose sugar.[13,14] SHAPE has been used to probe the secondary structure of a wide range of RNA from small tRNA[15] to the entire HIV genome.[16] Through the development of a suite of reagents SHAPE has also been used to measure RNA dynamics and give insight into the tertiary structure of an RNA.[17-20] While SHAPE reagents are sensitive to both secondary and tertiary interactions, it is difficult to deconvolute the influence of either type of interaction.

In order to understand higher order structures in RNA, much work has been done to develop techniques that more directly probe RNA tertiary structure. The development of these techniques has proven to be more challenging. Biophysical methods such as x-ray crystallography and NMR have been the primary methods for highly accurate tertiary structure determination.[8,21] X-ray crystallography yields high-resolution structures of RNA molecules, which can be very useful for determining stable RNA structure and function.

However, the usefulness of crystallographic approaches to studying RNA is limited because large RNAs can be difficult to crystallize and little information regarding structural dynamics is provided.[8] NMR spectroscopy can also produce high resolution structures of RNA and has been most widely used to measure local nucleotide dynamics.[21,22] However, NMR spectroscopy is also limited by the size and complexity of the RNA.

RNA tertiary interactions can also be interrogated using biochemical approaches. Two useful methods, nucleotide modification interference and analog interference, can identify nucleotides involved in RNA tertiary interactions. In modification interference, an RNA is first treated with reagents that generate chemical modifications on the RNA. The RNA is then subjected to a partitioning experiment to distinguish functional from non-functional molecules. Nucleotides that are absent in the functional RNA identify regions that are involved in the desired RNA function.[23] Comprehensive analysis of specific functional groups in RNA can also be achieved by Nucleotide Analog Interference Mapping (NAIM).[24,25] NAIM generically incorporates nucleotide analogs into an RNA transcript to identify the effect of individual functional groups on RNA interactions. The key step involved in both these approaches is selection of active RNAs from those that are inactivated due to the modification or nucleotide analog. Both modification interference and NAIM require the use of complementary reagents or nucleotide analogs to interrogate every nucleotide in an RNA. However these can identify single nucleotide or single atom interactions in an RNA.[23-25]

In order to more directly probe specific interactions in RNA, functionalized nucleotides can be used to provide information about nucleotide dynamics and through space interactions in three-dimensional space. FRET utilizes the incorporation of tethered

chromophores at various sites in an RNA, which allow for single molecule measurements of through-space interactions between nucleotides in an RNA.[26] Tethered cleaving reagents such as Fe(II)-EDTA^{2-}, which cleave the phosphate backbone of surrounding nucleotides can also be used to determine through-space interactions of RNA nucleotides.[27] While these techniques have been widely used to interrogate RNA tertiary structure, a major limitation is that they require the synthesis of non-native RNA and the addition of bulky functional groups that can alter native RNA folding.

RNA tertiary structure can be more generically mapped through the use of biochemical reagents, which probe native RNA in solution. Current methods are primarily based on the measurement of nucleotide solvent accessibility, measured through the cleavage of the RNA backbone by hydroxyl radicals generated by Fe[II]-EDTA^{2-} free in solution.[28,29] Unfortunately, data from hydroxyl radical probing (HRP) experiments can be technically challenging to interpret and the absolute correlation between hydroxyl radical reactivity and solvent accessibility remains imperfect.

The development of reliable, high-throughput techniques that map RNA tertiary structure is still an important goal in understanding RNA function. Currently, structurally chemical probes that are sensitive to RNA structure provide the most versatile platform to measure the tertiary structure of RNA.

1.1.3. Detection of RNA adducts

Traditionally, RNA adducts are detected as stops in a primer extension reaction using 5′-end labeled primers that are annealed to the 3′-end of the RNA. Reverse transcriptase produces cDNAs that terminate one nucleotide before the adduct position, resulting in a cDNA library in which fragment lengths correspond to the adduct position and fragment

proportions correspond to degree of modification. The cDNA fragments are then separated and analyzed by denaturing polyacrylamide gel or capillary electrophoresis.[11,12]

Recent advances in enzymatic adduct detection have made traditional biochemical probing techniques amenable to next-generation massively parallel sequencing.[30] With this approach, RNA adducts are detected not as stops but as mutations induced during read-through of the reverse transcriptase. Thus, every position in an RNA transcript can be sequenced, and adduct formation measured by rate of induced mutations. This technique has been used with SHAPE adducts to probe the secondary structure of a wide range of RNAs from small riboswitches to the entire HIV genome. While this approach vastly expands the number and type of experiments possible for studying RNA structure, it is currently limited to providing only secondary structure information.

1.1.4. DMD Modeling with biochemical constraints

An important application of experimentally-derived tertiary information is its use in conjunction with molecular modeling. Through the use of discrete molecular dynamics (DMD), a small number of constraints, reflective of through-space RNA structure, can be sufficient to produce high-quality structure models.[31,32] This approach allows for the determination of three-dimensional structures of a wide rage of RNAs not amenable to high-resolution methods. Through the application of HRP data, this approach has been used to produce a number of native-like structures for a variety of RNA.[33,34] While the current DMD algorithm is tolerant of the noise intrinsic to HRP experiments, it is clear that the quality of the derived structures is directly related to the quality of the experimental data.

1.2. RESEARCH OVERVIEW

Consequently, the overarching goal of my research was to develop information-rich

methods that characterize tertiary interactions in RNA of varying sizes and complexity. My focus was to adapt existing chemical probes with alternative processing and detection methods in order to develop new chemical methods that reflect tertiary structures in RNA.

In Chapter 2 I develop a combined biochemical and computational approach for creating high-quality models of RNA tertiary structure. This technique, called HMX, identifies nucleotides in structurally crowded regions of an RNA by exploiting the ability of a bulky adduct at the 2'-hydroxyl position to disrupt the overall RNA structure. These data are then incorporated as experimental constraints in discrete molecular dynamics (DMD) simulations to obtain experimentally informed, three-dimensional models.

In Chapter 3 I present a HMX analysis of the *Tetrahymena* group I intron in which I identify the presence of two stable substructures as well as the tertiary interactions involved in the folding of these separate structures. This analysis allows for the characterization of interactions involved in the docking of the 5' splice, between the P3-P9 and P4-P6 domain, as well as within the independently folded P4-P6 domain. This analysis provides a unique view of RNA structure and serves as guide for future HMX analyses of large RNA

In Chapter 4 I describe a chemical probing technique, which takes advantage of multi-nucleotide sequencing that can be used to detect multiple modifications in a single read. Through the use of massively parallel sequencing, statistical association analysis and spectral clustering, I identify correlated chemical modifications in individual RNA molecules. Through this approach, I have been able to identify higher-order, through-space interactions in an RNA, which have been used to refine three-dimensional structure models as well as distinguish multiple individual conformations in a single RNA ensemble.

1.3. PERSPECTIVE

In this work I use the principles of molecular biology, biochemistry, organic chemistry and physical chemistry to address the problem of accurate RNA structure determination, which can be used for a wide range of RNAs. Through the use of existing biochemical reagents I was able to modify the sample processing and detection of such adducts in order to measure through-space tertiary interactions in RNA. By melding this data with DMD simulations I have created techniques that can be used to predict accurate structural models, which were not yet possible for a wide range of RNA. These techniques also give a unique view of RNA structure that can be used to measure complex dynamics and interactions in RNA.

It is my hope that each of the methods I have developed will be applied to understanding structure-function relationships of RNAs in general, and specifically to predicting the tertiary structures of novel RNAs. I expect that these methods will lend themselves to multi-dimensional experiments in order to study protein-RNA interactions and lead to the further development of other RNA modification agents capable of interrogating RNA structure

1.4. REFERENCES

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

(2) Korostelev, A., and Noller, H. F. (2007) The ribosome in focus: new structures bring new insights. Trends Biochem. Sci. 32, 434–441.

(3) Strobel, S. A., and Cochrane, J. C. (2007) RNA catalysis: ribozymes, ribosomes, and riboswitches. Curr. Opin. Chem. Biol. 11, 636–643.

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

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

(6) Rich, A., and RajBhandary, U. L. (1976) Transfer RNA: molecular structure, sequence, and properties. Annu. Rev. Biochem. 45, 805–860.

(7) Reiter, N. J., Chan, C. W., and Mondragón, A. (2011) Emerging structural themes in large RNA molecules. Curr. Opin. Chem. Biol. 21, 319–326.

(8) Holbrook, S. R. (2008) Structural principles from large RNAs. Annu. Rev. Biophys. 37, 445–464.

(9) Dethoff, E. A., Chugh, J., Mustoe, A. M., and Al-Hashimi, H. M. (2012) Functional complexity and regulation through RNA dynamics. Nature 482, 322–330.

(10) Peattie, D. A., and Gilbert, W. (1980) Chemical probes for higher-order structure in RNA. PNAS 77, 4679–4682.

(11) Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.-P., and Ehresmann, B. (1987) Probing the structure of RNAs in solution. Nucleic Acid Res. 15, 9109–9128.

(12) Stern, B., Moazed, D., and Noller, H. F. (1988) Structural analysis of RNA using chemical and enzymatic probing monitored by primer extension. Methods Enzymol. 164, 481–489.

(13) Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J. Am. Chem. Soc. 127, 4223–4231.

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

(15) Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2005) RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNAAsp transcripts. J. Am. Chem. Soc. 127, 4659–4667.

(16) Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Jr, Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711–716.

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

(18) Mortimer, S. A., and Weeks, K. M. (2009) C2'-endo nucleotides as molecular timers suggested by the folding of an RNA domain. PNAS 106, 15622–15627.

(19) Mortimer, S. A., and Weeks, K. M. (2008) Time-resolved RNA SHAPE chemistry. J. Am. Chem. Soc. 130, 16178–16180.

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

(21) Rinnenthal, J., Buck, J., Ferner, J., Wacker, A., FÜrtig, B., and Schwalbe, H. (2011) Mapping the landscape of RNA dynamics with NMR spectroscopy. Acc. Chem. Res. 44, 1292–1301.

(22) Latham, M. P., Brown, D. J., McCallum, S. A., and Pardi, A. (2005) NMR methods for studying the structure and dynamics of RNA. ChemBioChem 6, 1492–1505.

(23) Conway, L., and Wickens, M. (1989) Modification interference analysis of reactions using RNA substrates, in Methods Enzymol., pp 369–379. Academic Press. Inc.

(24) Strobel, S. A. (1999) A chemogenetic approach to RNA function/structure analysis. Curr. Opin. Struct. Biol. 9, 346–352.

(25) Ryder, S. P., and Strobel, S. A. (1999) Nucleotide Analog interference mapping. Methods 18, 38–50.

(26) Roy, R., Hohng, S., and Ha, T. (2008) A practical guide to single-molecule FRET. Nature Methods 5, 507–516.

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

(28) Tullius, T. D., and Greenbaum, J. A. (2005) Mapping nucleic acid structure by hydroxyl radical cleavage. Curr. Opin. Chem. Biol. 9, 127–134.

(29) Pastor, N., Weinstein, H., Jamison, E., and Brenowitz, M. (2000) A detailed interpretation of OH radical footprints in a TBP-DNA complex reveals the role of dynamics in the mechanism of sequence-specific binding. J. Mol. Biol. 304, 55–68.

(30) N. A. Siegfried, et al. & K. M. Weeks, submitted (2014).

(31) Gherghe, C. M., Leonard, C. W., Ding, F., Dokholyan, N. V., and Weeks, K. M. (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.

(32) Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA 14, 1164–1173.

(33) Ding, F., Lavender, C. A., Weeks, K. M., and Dokholyan, N. V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. Nat. Methods 9, 603–608.

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

CHAPTER 2. RNA TERTIARY STRUCTURE ANALYSIS AND REFINEMENT BY 2'-HYDROXYL MOLECULAR INTERFERENCE*

2.1. INTRODUCTION

RNA plays diverse and central roles in the regulation of gene expression.[1] Information is encoded in the RNA at several levels: the primary sequence, the precise base-pairing pattern that defines the secondary structure, and higher-order RNA structures composed of tightly packed secondary structure elements stabilized by a few key tertiary interactions.[2] The precise formation of higher-order tertiary structures is critical to the function of many RNAs.[3,4] RNA secondary and tertiary interactions can be interrogated using chemical probing approaches that evaluate how a chemical adduct or substitution disrupts structure or that evaluate accessibility of particular functional groups in the RNA to a chemical reagent. In modification interference, an RNA is treated to introduce chemical modifications, usually in the nucleobase moieties, and then the RNA is subjected to a partitioning experiment to distinguish functional from non-functional molecules.[5-7] Comprehensive analysis of specific functional groups in RNA can also be achieved by nucleotide analog interference mapping (NAIM).[8,9] For NAIM, nucleotide analogs are incorporated into an RNA transcript, and active RNAs are partitioned from those that are inactivated due to the nucleotide analog. Both modification interference and NAIM can interrogate most nucleotides in an RNA to identify single nucleotide or single atom interactions, respectively, critical to the tertiary structure.[5,6,8,9] These approaches however, require multiple experiments to interrogate the tertiary environment of every nucleotide in

*This chapter has been submitted for publication in JACS.

the RNA.

Chemical probes are also routinely used to examine both solvent accessibility and dynamics in the RNA backbone. The solvent accessibility of the RNA backbone can be monitored by hydroxyl radical footprinting (HRP).[10,11] Reagents used in the selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) strategy are sensitive to the nucleophilicity of the 2'-OH group, which is dependent on the underlying flexibility of the nucleotide.[12-14] Reactivities of these and other chemical probes, like DMS, CMCT and kethoxal, are inhibited by both secondary and tertiary structure interactions, and it is usually difficult to deconvolute the relative influence of each type of interaction.

Here, we describe a strategy in which 2'-hydroxyl-selective reagents are used in a modificationinterference experiment to simply and directly interrogate RNA tertiary structure. In this approach, which we call 2'-hydroxyl molecular interference or HMX, a hydroxyl-selective reagent is used to create a pool of RNAs with evenly distributed 2'-*O*-ester adducts. A structureselective pressure, such as RNA folding, is placed on the pool of modified RNA. A subset of 2'-*O*-ester groups will interfere with molecular interactions and prevent native RNA folding. By partitioning the sample into folded and unfolded states, nucleotides whose modification disrupts tertiary interactions are identified. This information is used to characterize the internal packing interactions that define higher-order RNA structure and to refine three-dimensional structure models. The HMX strategy yielded accurate and highly statistically significant models for RNAs of known structure.

2.2. RESULTS

2.2.1. HMX overview

In the first step of the HMX strategy, an RNA of interest was modified with a 2'-hydroxyl selective reagent under denaturing conditions such that all nucleotides were modified in some RNAs in the population. Second, the RNA was allowed to fold; and, third, the RNA was subjected to a selection step to partition the RNA into active and inactive components (Fig. 2.1A). An experiment with an unmodified control was performed in parallel. The RNAs in this analysis were modified using N-methylisatoic anhydride (NMIA),[12] which modifies all positions in an RNA at 95 °C at low ion concentrations (Fig. 2.2). Reaction with NMIA yields a bulky ester adduct at the 2'-hydroxyl position in the RNA backbone (Fig. 2.1B). Some adducts will have no or small structural consequences, whereas other adducts will prevent proper folding of the RNA. In this work, we partitioned the natively folded from the unfolded structures based on mobility in a non-denaturing acrylamide gel, although many other selection strategies are compatible with this approach. After partitioning, folded and unfolded populations were analyzed by reverse transcription-mediated primer extension to detect positions of the modified nucleotides. Adducts that disrupted folding were identified by comparing the profiles of the unfolded and folded RNA at each position (Fig. 2.1C). HMX scores for each nucleotide were calculated as the differences between the normalized profiles for the folded and unfolded RNA. Nucleotides with high HMX scores are likely involved in higher-order RNA packing and tertiary interactions.

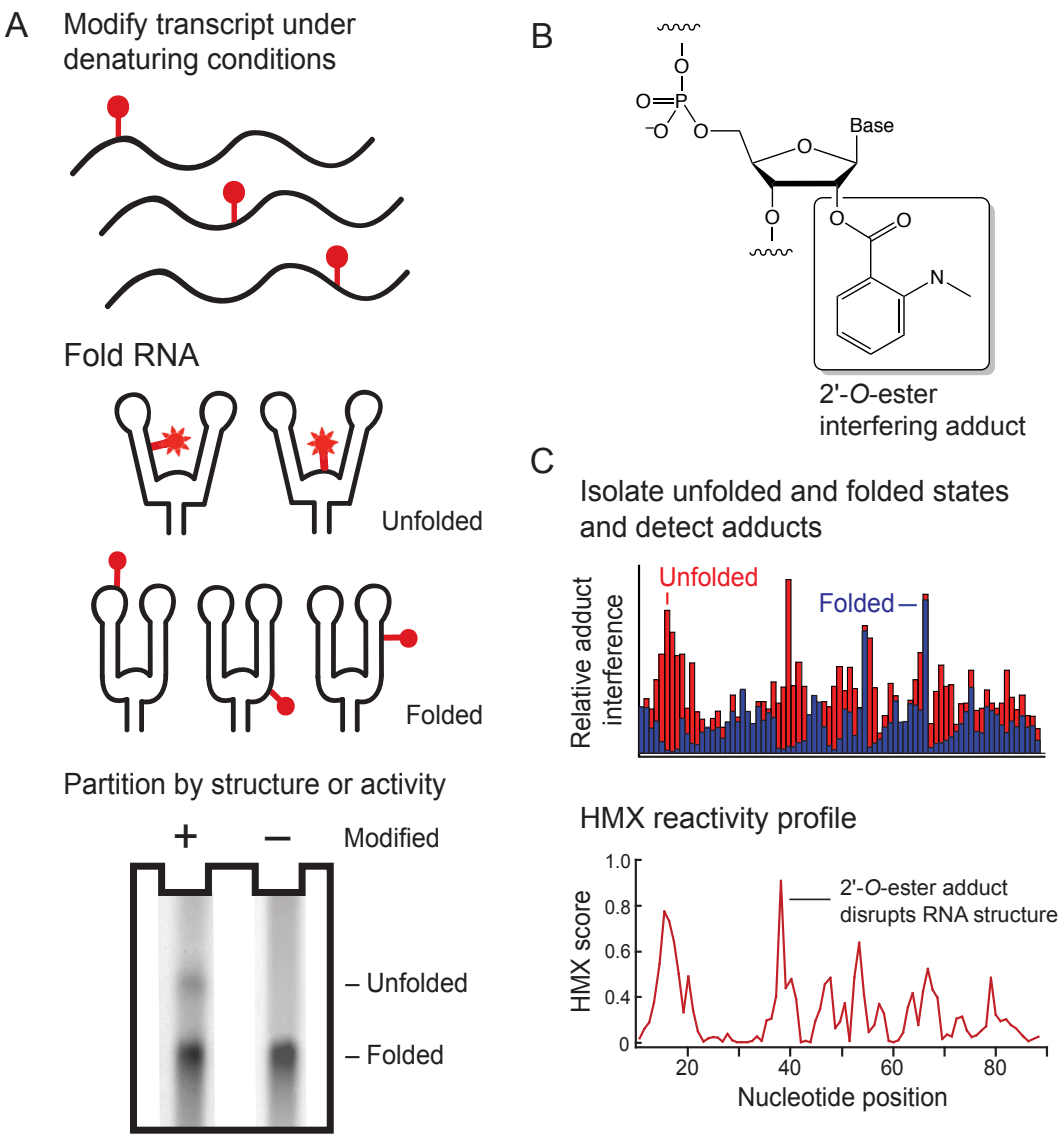


Figure 2.1 2'-Hydroxyl molecular interference (HMX). (**A**) RNA is modified under denatured conditions such that all nucleotides have a significant probability of being modified. Some 2'-hydroxyl adducts prevent native folding, creating a population of unfolded RNA that can be partitioned from fully folded RNA. In this work, partitioning was performed by non-denaturing gel electrophoresis. (**B**) Structure of the 2'-*O*-ester adduct introduced by reacting RNA with NMIA. (**C**) Partitioned populations were separately subjected to primer extension to detect adducts. Positions with high intensities in the unfolded RNA have low intensities in the folded RNA and indicate positions of adducts that prevent folding. HMX profiles were calculated using a cross-correlation analysis.

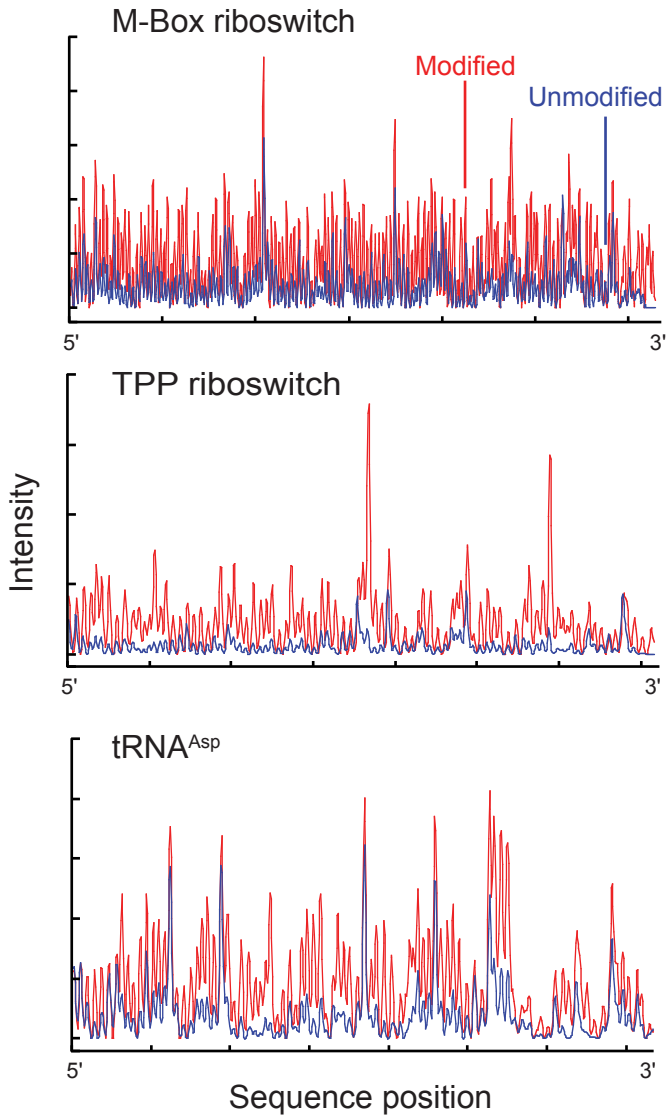


Figure 2.2 Electropherograms of 2'-*O*-ester modified and unmodified RNAs. Each RNA was modified with NMIA under denaturing conditions. Modifications were detected as stops to reverse transcriptase-mediated primer extension.

We examined the HMX approach using structurally diverse RNAs for which high-resolution structures are available: yeast $tRNA^{Asp}$ (75 nts),[15] *E. coli* thiamine pyrophosphate (TPP) riboswitch (79 nts),[16] the *B. subtilis* M-Box riboswitch (156 nucleotides),[17] and the *Tetrahymena* group I intron P546 domain (160 nts).[18] After modification of the RNA with NMIA under denaturing conditions, RNAs were folded under conditions known to stabilize the native and functional tertiary structure of the RNA. The TPP and M-Box riboswitch RNAs were folded in the presence of saturating ligand concentrations.

For three of the RNAs, separate populations of fully folded and unfolded RNA states were readily resolved by non-denaturing polyacrylamide gel electrophoresis (Fig. 2.3A). In these cases, there was a clear shift to a second (unfolded and less compact) state in the modified RNA, relative to the unmodified control RNA (Fig. 2.3B). In the case of $tRNA^{Asp}$, folded and unfolded states were not well separated. The unfolded state for $tRNA^{Asp}$ migrated slightly more rapidly than the folded state, implying that the L-shaped tertiary structure causes the folded RNA to migrate more slowly than does the unfolded structure, consistent with studies on bent RNAs.[19]

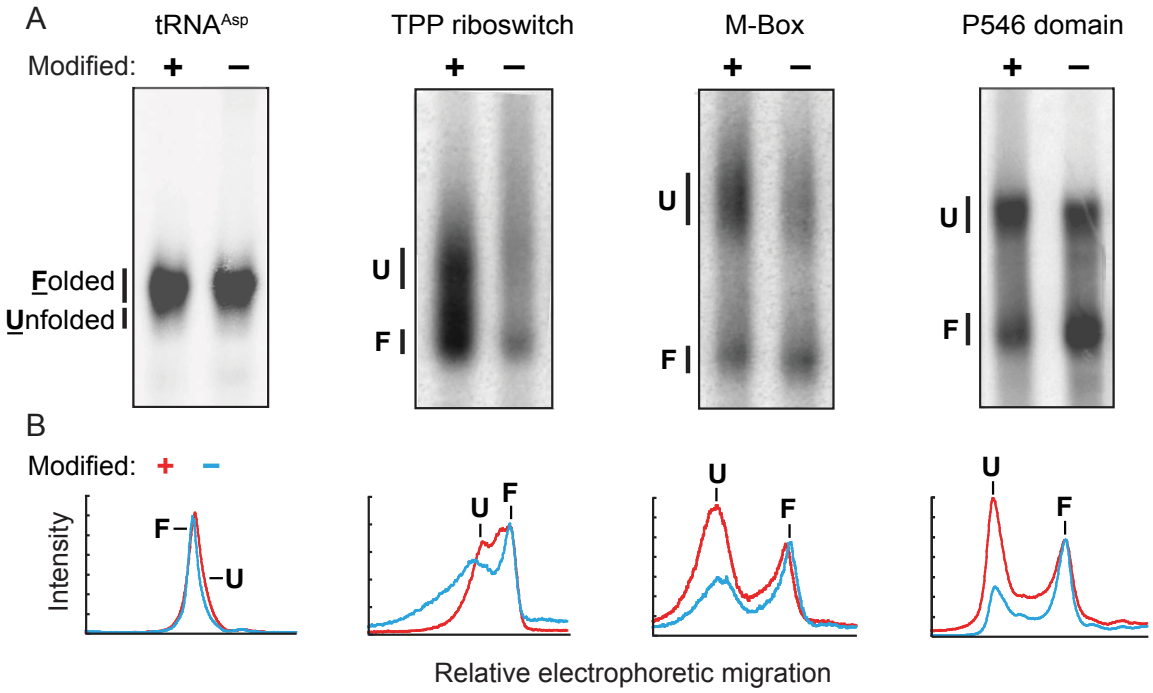


Figure 2.3 Partitioning of RNA populations by native gel electrophoresis. (**A**) Folded and unfolded populations for modified and unmodified RNAs were separated by non-denaturing gel electrophoresis in the presence of 50 mM NaCl and 5 mM $MgCl_2$. For clarity, gel images were straightened and scaled to show similar representations for each RNA; band profiles and intensities were not altered. (**B**) Band intensities as a function of gel migration distance.

After partitioning, adducts that prevent folding are over represented in the unfolded band and underrepresented in the folded band. Site of modification were identified in both populations by reverse transcription-mediated primer extension. The resulting data were normalized using a cross-correlation approach to create an HMX score that allowed identification of nucleotides preferentially modified in the unfolded population relative to the folded population. The HMX score takes into account that 2'-*O*-adduct molecular interference compares two structurally encoded data sets (unfolded and folded populations) partitioned from the same pool of modified RNA and that separation of the unfolded and folded populations is imperfect (Methods and Fig. 2.4). Positions with medium and high interference scores were visualized on the known[15-18] three-dimensional structures of each of the RNAs (Fig. 2.5). Nucleotides with high HMX scores corresponded to nucleotides directly involved in tertiary interactions and to nucleotides within densely packed regions of the RNA. Because the 2'-*O*-ribose modification occurs in the RNA backbone and likely does not significantly destabilize helix formation,[20] interfering positions corresponded almost exclusively to higher-order interactions and not to canonically base-paired nucleotides (Fig. 2.5).

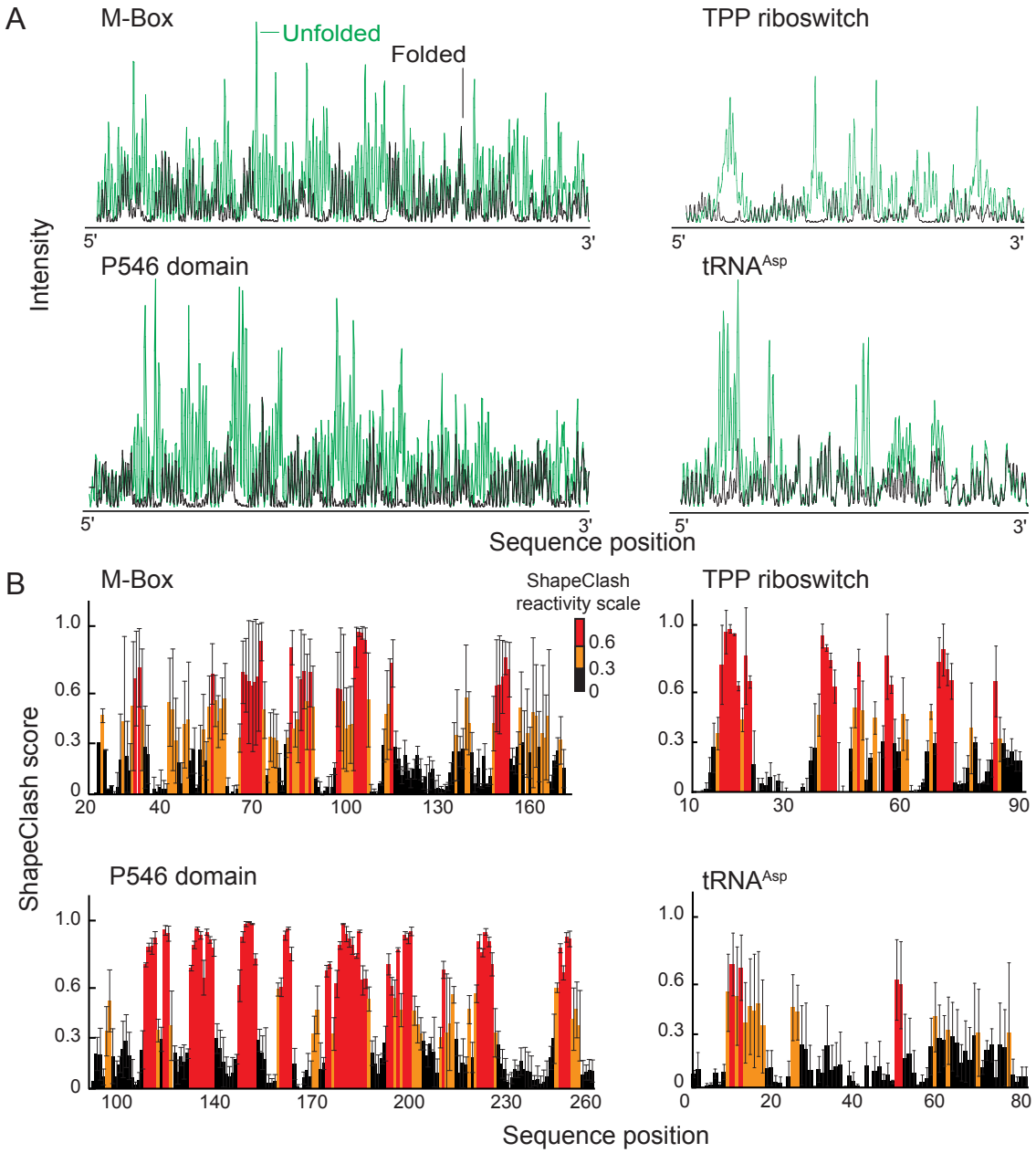


Figure 2.4 Calculation RNA HMX scores by normalization and cross-correlation. (**A**) Electropherograms of unfolded RNA (green) scaled to data for folded RNA populations (black). Positions with high intensities in the unfolded RNA have low intensities in the folded RNA. (**B**) Cross-correlation normalization, based on both unfolded and folded 2'-*O*-adduct profiles, was used to create HMX score profiles. Positions with low, medium, and high HMX scores are black, orange, and red, respectively. Experiments were performed in triplicate and error bars are shown with black lines.

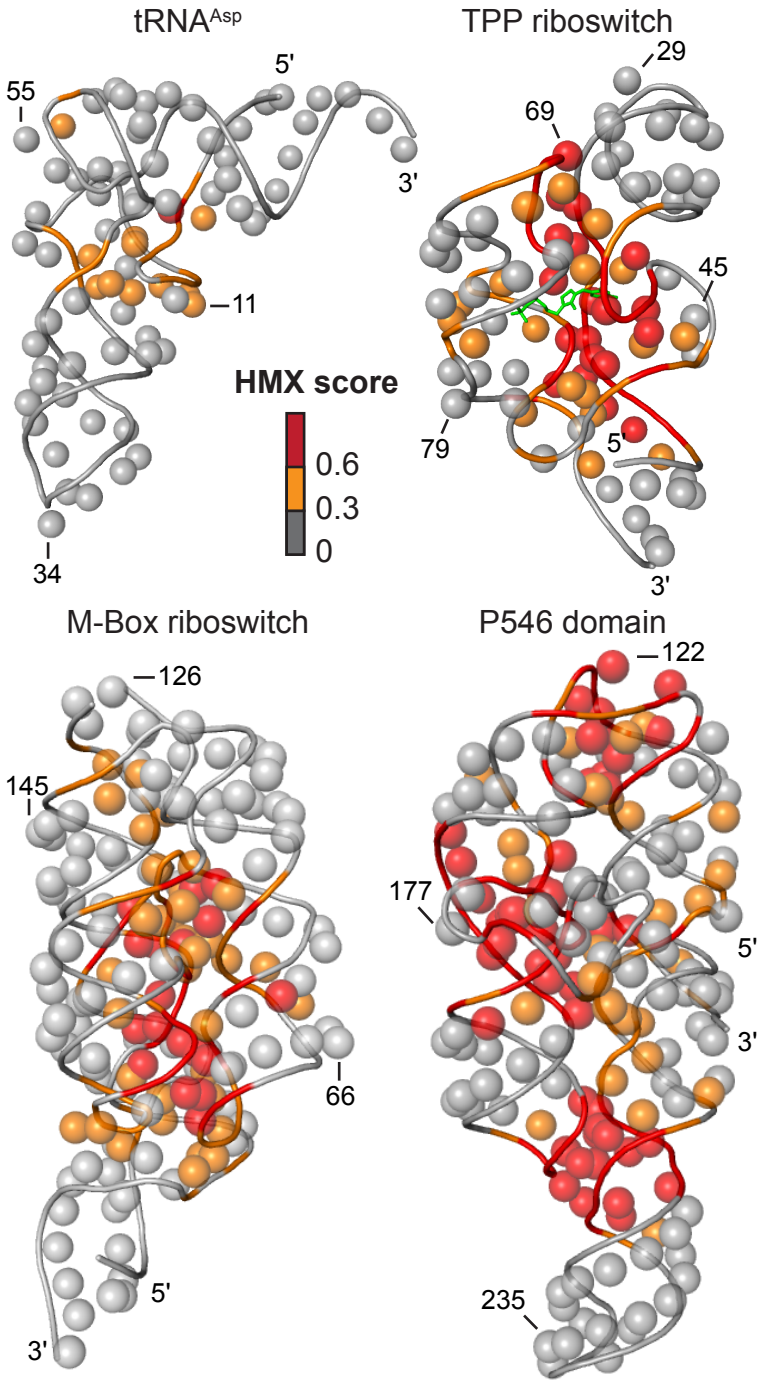


Figure 2.5 Visualization of HMX interference information on accepted three-dimensional structures.[15-18] The 2'-OH group for each nucleotide is shown as a sphere and the phosphate backbone by a tube. Nucleotides are colored by HMX score; the TPP ligand is green.

2.2.2. Molecular overlap model for HMX intensities.

Because molecular interference appeared to be strongly correlated with RNA tertiary interactions, we sought to understand the molecular basis of this correlation. We first defined a pseudo-atom, representing 2'-*O*-ester adduct, described by two parameters: *L*, the distance of the pseudo-atom from the 2'-oxygen atom, and *r,* the radius of the pseudo-atom. We calculated the degree to which surrounding nucleotide atoms intersected the defined pseudo-atom shell, based on the van der Waals radii (Fig. 2.6A). The pseudo-atom bond length and atomic radius were determined by searching over the two pseudo-atom parameters. Pearson correlation coefficients were calculated for ranges of *L* and *r* relative to the experimental interference score (Fig. 2.6B). The pseudo-atom parameters that best fit the experimental data for all RNAs were *L* of 2 Å and *r* of 5 Å. A pseudo-atom with these parameters tightly, and fully, encapsulates the NMIA adduct ester at a ribose ring (Fig. 2.6C). The correlations between the experimental interference scores and the molecular overlap calculations for each RNA are high (Fig. 2.6D), which suggests that the 2'-*O*-ester adduct disrupts RNA structure by sterically blocking RNA interactions in crowded regions of the RNA. For example, in the presence of ligand, nucleotides in the TPP riboswitch interact directly with ligand and other nucleotides form RNA-RNA contacts. The HMX experiment was sensitive to both types of interactions, indicating that HMX will be useful for examining intramolecular and intermolecular RNA contacts and protein and small molecule ligand interactions with RNA.

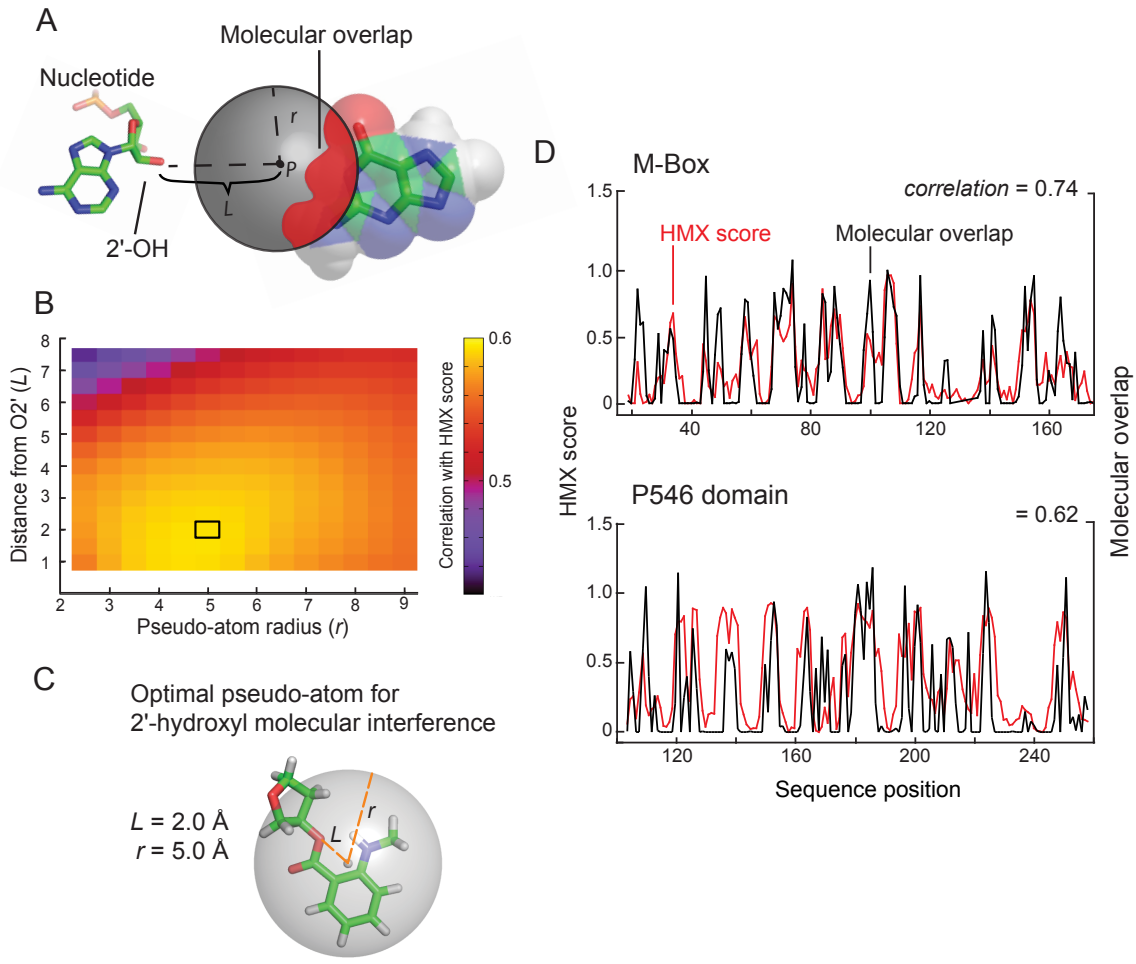


Figure 2.6 Physical model for 2'-hydroxyl molecular interference. (**A**) Model for interference by molecular overlap in which adducts are represented by a pseudo-atom (grey) at a distance (*L*) from the O2' position at radius (*r*). The degrees to which surrounding atoms intersect the pseudo-atom were estimated by calculating van der Waals radii overlaps. (**B**) Analysis of optimal pseudo-atom bond length and atomic radius. Maximum correlation between Pearson's *r* and pseudo-atom representing 2'-hydroxyl molecular interference is boxed. (**C**) Relationship between pseudo-atom dimensions and 2'-*O*-ester adduct. (**D**) Representative relationships between HMX scores and molecular overlap for the M-Box and P546 domain RNAs. HMX score profiles (red) show a high correlation with calculated molecular overlaps (black) for each RNA. Pearson correlation coefficients are shown. Correlation coefficients for the $tRNA^{Asp}$ and TPP riboswitch RNAs (not shown) were 0.60 and 0.72, respectively.

2.2.3. Three-dimensional RNA structure modeling.

The HMX strategy yields a physical measurement related to the extent of RNA tertiary structure packing at the site of an individual nucleotide. We therefore explored whether HMX scores could be used to refine three-dimensional models for RNA. Our labs have previously shown that HRP data, which are roughly related to solvent accessibility, used as structural constraints in discrete molecular dynamics simulations increased the quality of structural models[21,22] relative to unconstrained simulations.[23] Implementation of the HRP-based potential is experimentally and computationally difficult, because HRP data are inherently noisy, especially for smaller RNAs with few solvent-protected nucleotides. We adapted the procedure previously developed for HRP-refined structures[23,24] to develop three-dimensional models for $tRNA^{Asp}$, the TPP and M-Box riboswitches, and the P546 domain based on HMX constraints. DMD simulations were performed using a simplified model in which each nucleotide was represented by three pseudo-atoms corresponding to the base, sugar, and phosphate groups.[22] The energy function used to direct folding included terms for bonded and non-bonded interactions. Additional potential energy terms incorporated information on relative solvent accessibility as estimated from HMX data. Structures obtained from simulations were clustered based on structural similarity to generate a representative model for each RNA.

The predicted structural models for each RNA were evaluated in terms of number and population of clusters as well as the calculated root-mean-square deviation (RMSD) from the accepted crystal structures. HMX-directed simulations produced structurally and statistically significant[25] native-like RNA structure models for the TPP and M-Box riboswitches and for the P546 domain (Fig. 2.7). For each RNA, the central structure in the highest population

cluster contained the structure with the lowest RMSD relative to the accepted structure. The HMX-directed simulations for $tRNA^{Asp}$ were less successful, and the most populated cluster had a poor RMSD relative to the accepted structure while the best RMSD structure was found in the lowest population cluster (Fig. 2.7 insert). This simulation itself revealed that no well-determined model emerged, as no single cluster included a majority of structures consistent with the molecular interference information (Fig. 2.7 and Table 2.1).

Table 2.1 DMD simulation statistics for the four RNA fold refinements. Cluster populations (*n*), mean RMSDs, cluster energies, and *p*-values for each structure are shown.

RNA	Cluster	*n*	RMSD (Å)	Energy	*p*-value
M-Box	1	100	13.7 ± 1.3	-79.9 ± 2.0	$<1 \times 10^{-6}$
P546	1	93	19.8 ± 1.2	-78.5 ± 2.2	5.2×10^{-3}
	2	7	24.2 ± 0.4	-70.6 ± 1.1	4.6×10^{-1}
TPP	1	67	8.3 ± 1.0	-38.1 ± 1.7	8.4×10^{-4}
	2	33	11.6 ± 1.0	-38.6 ± 1.4	3.7×10^{-2}
$tRNA^{Asp}$	1	43	18.7 ± 1.3	-35.0 ± 0.9	1.0
	2	13	14.1 ± 1.1	-34.8 ± 0.7	5.2×10^{-1}
	3	12	15.3 ± 0.9	-35.1 ± 0.9	7.5×10^{-1}
	4	9	15.4 ± 0.5	-34.9 ± 0.5	7.7×10^{-1}
	5	8	17.1 ± 1.3	-34.6 ± 0.8	9.6×10^{-1}
	6	6	12.5 ± 1.1	-34.4 ± 0.6	2.0×10^{-1}
	7	5	13.6 ± 0.9	-34.8 ± 0.7	3.9×10^{-1}
	8	2	15.9 ± 0.2	-35.6 ± 0.1	8.4×10^{-1}
	9	1	18.7 ± na	-35.3 ± na	1.0
	10	1	7.6 ± na	-34.8 ± na	1.7×10^{-4}

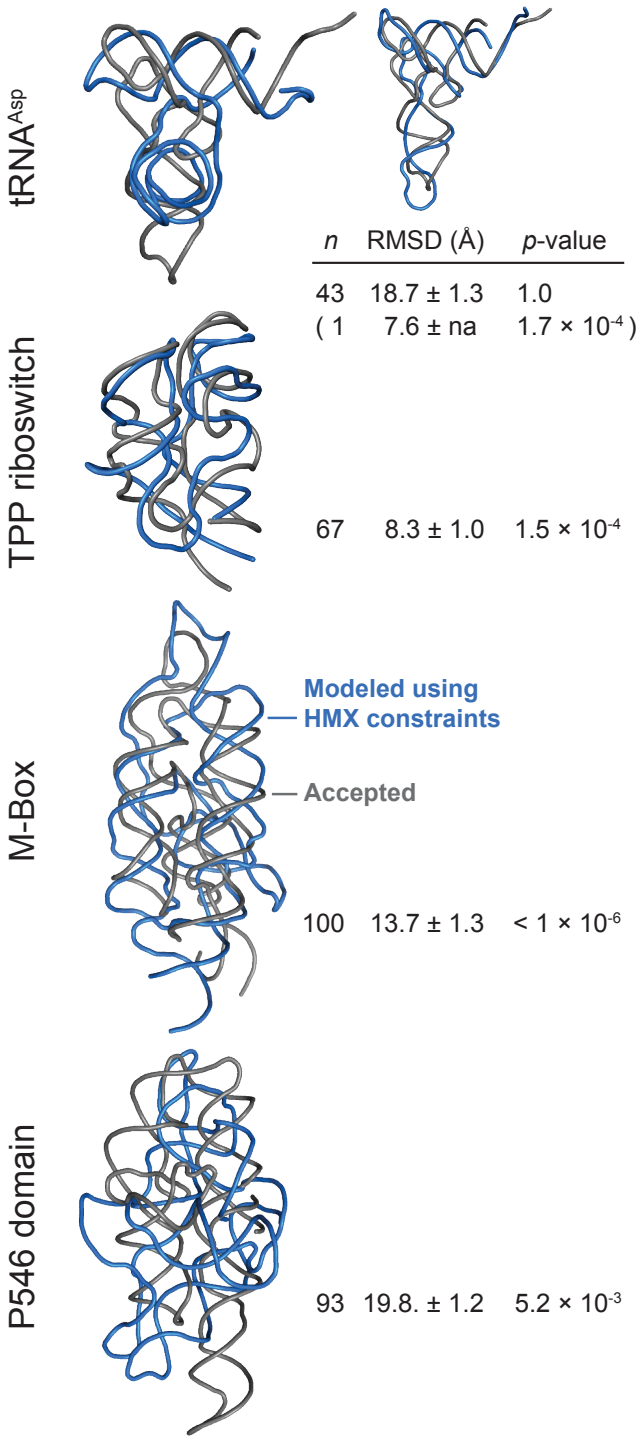


Figure 2.7 HMX-directed RNA fold refinements. RNAare shown as backbone traces. Accepted structures[15-18] and HMX-directed refinements for each RNA are gray and blue, respectively. The cluster populations (*n*), mean RMSD, and *p*-values[25] are shown. For $tRNA^{Asp}$, both the largest cluster (large image) and lowest RMSD structures (inset) are shown.

2.3. DISCUSSION.

HMX measures the effect of introducing a molecular perturbation at the ribose 2'-OH position on RNA folding. Modifications at the 2'-ribose position, which lie on the exterior of an RNA duplex, generally do not substantially destabilize simple RNA secondary structures[20,26]. Thus, the 2'-*O*-ester molecular interference measurement is exquisitely sensitive to interactions that govern RNA tertiary folding. For the four RNA evaluated in this work – $tRNA^{Asp}$, the TPP and M-Box riboswitch aptamer domains, and the P546 domain RNA – the interfering nucleotides identified by HMX correspond to those within the densely packed interior of these structures (Fig. 2.5). This relationship is highly quantitative. Molecular interference by the 2'-*O*-ester group was highly correlated with a sphere of defined location relative to the RNA ribose group for each of the four RNAs analyzed (Fig. 2.6). We anticipate that 2'-*O*-ester mediated molecular interference will prove highly useful in evaluating higher-order RNA packing in the context of large RNAs and RNA-protein complexes.

The gel electrophoresis approach to partitioning approach used here revealed the high level of cooperativity and structural interdependency in RNA tertiary folding. For the three largest RNAs, the introduction of distinct molecular adducts resulted in a single predominant unfolded structural population (Fig. 2.3), rather than multiple populations that would be expected due to the absence of individual tertiary interactions. This result emphasizes the cooperativity of folding of these RNAs of less than 160 nucleotides. In larger RNAs, HMX will likely be extendable to identify stable substructures that partition selectively between the fully folded and unfolded RNA states.

Use of HMX information to direct a DMD-based three-dimensional structure

refinement yielded high-quality structures in agreement with known RNA structures. The molecular dynamics refinement employed was based on a recently developed approach that allows solvent accessibility to be incorporated into the simulation as an energy restraint.[23] In general, use of molecular interference information resulted in final clustered structures that were better defined than those obtained with hydroxyl radical restraints (data not shown). For the TPP and M-Box riboswitches and the P546 domain, HMX-directed refinement yielded a single predominant cluster with a highly statistically significant agreement with the accepted structure (Fig. 2.7). $tRNA^{Asp}$ was poorly modeled, likely reflecting both our inability to cleanly separate folded and unfolded states (Fig. 2.3) and the resulting low magnitude molecular interference data (Fig. 2.5, compare tRNA with other RNAs). Critically, the simulation itself clearly reported that this RNA was not a good target for refinement because ten distinct clusters were observed and no single cluster clearly dominated the simulation. The overall success of HMX-directed refinement (Fig. 2.7) suggests that de novo RNA structure refinement, based on easily obtained high-quality biochemical constraints, holds substantial promise for understanding structure-function interrelationships for RNA.

HMX analysis provides a generic approach for analysis of the internal tertiary structure architecture of functionally important RNAs. The information gained from this experiment provides a unique view of internal and closely packed RNA tertiary structure. Here RNAs were partitioned based on gel electrophoresis; however, techniques that separate functional from non-functional RNAs could be used, allowing HMX analysis to be implemented based on ability of an RNA to interact with proteins or with other RNAs or to perform catalysis.

2.4. METHODS

2.4.1. RNA constructs.

DNA templates for yeast $tRNA^{Asp}$, the *E. coli* TPP and *B. subtilis* M-Box riboswitches, and the *Tetrahymena* group I intron P546 domain included 5' and 3' structure cassette flanking sequences[13] and were generated by PCR. RNAs were transcribed *in vitro* [1 mL; 40 mM Tris (pH 8.0), 10 mM $MgCl_2$, 10 mM dithiothreitol, 2 mM spermidine, 0.01% (v/v) Triton X-100, 4% (w/v) poly(ethylene) glycol 8000, 2 mM each NTP, 50 μL PCR-generated template, 0.1 mg/mL T7 RNA polymerase; 37 °C; 4 h] and purified by denaturing polyacrylamide gel electrophoresis [8% polyacrylamide, 7 M urea, 29:1 acrylamide:bisacrylamide, 0.4 mm × 28.5 cm × 23 cm gel; 32 W, 1.5 h]. RNAs were excised from the gel, recovered by passive elution overnight at 4 °C, and precipitated with ethanol. The purified RNAs were resuspended in 50 μL TE and stored at -20 °C.

2.4.2. 5'-[^{32}P] RNA radiolabeling.

Purified RNAs were dephosphorylated [300 μL; 50 mM Tris (pH 8.5), 0.1 mM EDTA, 10 μM RNA, 300 units SUPERase-In (Ambion), 200 units alkaline phosphatase (Roche); 50 °C; 1 h], subjected to phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation, and resuspended in TE (storage at - 20 °C). The RNA was 5' -[^{32}P]-radiolabeled by treatment with T4 polynucleotide kinase [40 μL; 80 pmol dephosphorylated RNA, 70 mM Tris (pH 7.6), 10 mM $MgCl_2$, 5 mM DTT, 2 μL T4 polynucleotide kinase (NEB, 10,000 units/mL), 150 μCi [γ -^{32}P]-ATP; 37 °C; 30 min]. Radiolabeled RNAs were purified by denaturing (8%) gel electrophoresis, excised from the gel, and recovered by overnight passive elution at 4 °C. The purified 5'-[^{32}P]-labeled RNAs were precipitated with ethanol, resuspended in 5 mM Tris (pH 7.5), and stored at -20 °C. The 5'-[^{32}P]-labeled RNA

was resuspended in 5 mM Tris buffer such that 1 μL of resuspended RNA measured approximately 10^6 dpm of [^{32}P].

2.4.3. RNA modification for molecular interference.

RNA (25-50 pmol) was denatured by heating to 90 °C for 2 min [32 μL; 30 pmol unlabeled RNA, 1.5 dpm 5'-[^{32}P]-radiolabeled RNA, 100 mM HEPES (pH 8.0)]. The denatured RNA was added to NMIA solution (1.2 μL, 0.4 M in DMSO) and allowed to react at 95 °C for 5 min. The modification process was repeated three times. After the third modification, the sample was placed on ice. A no-modification control reaction was performed identically using 1.2 μL DMSO. Any water evaporated during the modification was replaced to bring the volume to 36 μL (TPP samples were brought to 32 μL). For experiments in which band populations were quantified and visualized, 100,000 dpm of 5'-[^{32}P]-radiolabeled RNA was used per condition. Gels were visualized by phosphorimaging.

2.4.4. RNA folding and structural partitioning.

After the denatured RNA was modified, it was treated with 4 μL 10x folding buffer (100 mM $MgCl_2$, 1 M NaCl), and incubated at 37 °C for 30 min. Folding of the TPP riboswitch RNA was similar except that the RNA was incubated in folding buffer at 37 °C for 10 min, after which the TPP ligand (4 μL, 50 mM) was added and the sample was incubated at 37 °C for 20 min. TPP ligand concentration was saturating (K_d = 50-200 nM[27]). The 40 μL of folded RNA sample was immediately added to equal volume of an 80% glycerol solution containing bromophenol blue and xylene cyanol and resolved on a non-denaturing polyacrylamide gel [8% polyacrylamide, 19:1 acrylamide:bisacrylamide, 0.5× TB (45 mM Tris, 45 mM boric acid), 50 mM NaCl, 5 mM Mg_2Cl; 0.4 mm × 28.5 cm × 23 cm gel; 20 W, 8 h]. The gel was run in a cold room at 4 °C to ensure that the gel temperature did

not increase above 37 °C. The anode and cathode buffer wells were periodically refreshed to maintain ion homostasis. Bands were visualized by exposing the gel to film (Kodak BioMax) for 1 hr. The film was used as a template to identify and guide excision of the unfolded and folded band from the gel. The samples were recovered by passive elution overnight at 4 °C and were purified by ethanol precipitation and resuspended in 10 μL water.

2.4.5. Reverse transcription and adduct detection.

The general procedure was outlined previously.[13] DNA primers were 5'-end labeled with VIC or NED fluorophores (Applied Biosystems). RNA extracted from native gel separation (10 μL) was added to a fluorescently labeled DNA primer (5' -VIC-labeled GAA CCG GAC CGA AGC CCG; 3 μL, 0.3 μM) and allowed to anneal at 65 °C for 6 min and then cooled on ice. Reverse transcription buffer [6 μL; 167 mM Tris (pH 8.3), 250 mM KCl, 10 mM $MgCl_2$, 1.67 mM each dNTP] and Superscript III (1 μL, 200 units) were added, and samples incubated at 45 °C for 2 min, 52 °C for 20 min, and 65 °C for 5 min. The reactions were quenched with 4 μL 50 mM EDTA. The cDNAs were recovered by ethanol precipitation, washed twice with 70% ethanol, dried, and resuspended in 10 μL deionized formamide. Dideoxy sequencing ladders were produced using unlabeled, unmodified RNA by annealing a 5' -NED-labeled fluorescently labeled DNA primer (3 μL, 0.3 μM), and by adding 1 μL 2',3' -dideoxycytosine (10 mM) triphosphate before addition of Superscript III. cDNA fragments were separated by capillary electrophoresis using an Applied Biosystems 3130 DNA sequencing instrument. Capillary electrophoresis traces were analyzed using QuShape.[28]

2.4.6. Calculation of the HMX score.

HMX experiments measure the functional partitioning of the folded, $S^F(i)$, and

unfolded, $S^U(i)$, RNA ensembles in the presence of an 2'-*O*-ester adduct. Reactivity profiles from the folded and unfolded RNA populations were used to calculate an HMX score, $\text{Score}(i) \sim S^U(i)/(S^F(i) + \alpha S^U(i))$, where the coefficient α reflects the relative populations of RNAs with modifications at each nucleotide *i* (given that the absolute number of modifications in the ensemble cannot be determined). We estimated α by assuming that the total amount of adduct introduced at each nucleotide position is roughly the same, reflecting that modification was performed under denaturing conditions. We calculated α by minimizing the ratio between the average total modifications over the corresponding standard deviation, $\langle S^U(i) + \alpha S^F(i)\rangle / \delta(S^U(i) + \alpha S^F(i))$, where the average and standard deviation are taken over all nucleotide positions:

$$\alpha = \frac{[\langle S^U S^U\rangle - \langle S^U\rangle^2]\langle S^F\rangle - [\langle S^F S^U\rangle - \langle S^U\rangle\langle S^F\rangle]\langle S^U\rangle}{[\langle S^F S^F\rangle - \langle S^F\rangle^2]\langle S^U\rangle - [\langle S^F S^U\rangle - \langle S^U\rangle\langle S^F\rangle]\langle S^F\rangle}.$$

Physical separation of native and non-native ensembles was not completely quantitative. The native ensemble was generally well defined, but there was some native-like RNA in the non-native ensemble. We took this into account by subtracting the native-like reactivity profile from the non-native one:

$$\text{Score}(i) \sim (S^U(i) - \beta S^U(i))/(S^F(i) + \alpha S^U(i)).$$

The contribution of the native state was most pronounced for nucleotides with low adduct reactivity in the unfolded ensemble. Therefore, we estimated the coefficient β by identifying regions with relatively low adduct reactivity in the unfolded ensemble that also had the highest correlation coefficients between the unfolded and folded ensembles (Fig. 2.4A). The coefficient β was defined as the slope of the linear regression between unfolded and folded

2'-*O*-ester adduct intensity in these regions. Positions with HMX scores less than 0.3, between 0.3-0.6, and greater than 0.6 were defined as low, medium, and high, respectively (Fig. 2.4B).

2.4.7. Modeling of adduct disruption of native RNA tertiary structure.

The 2'-*O*-ester adducts were modeled as spheres (Fig. 2.6C). The RNA model was extracted from a pdb file,[29] and hydrogen atoms were added using the Molprobity web service.[30] Volume integrals were calculated using a Monte Carlo integration algorithm. The center of the adduct sphere was defined as a vector in the direction of the ribose C2-O2' bond of length *L* from the ribose O2' position. A grated potential was created to weight clashes nearer the center of an adduct by randomly sampling 100,000 random points from a normal distribution with σ defined as the radius of the adduct. Points falling within the van der Waals radii of atoms in the PDB, excluding the nucleotide origin or directly adjacent nucleotides, were scored as hits. High values of the molecular overlap (dimensionless) indicate regions where adducts have high probabilities of disrupting native RNA tertiary structure.

2.4.8. HMX-directed structure refinement by DMD.

DMD simulation and analyses consisted of three steps. First, the RNA was folded from the linear sequence, constrained by the accepted canonical base pairing pattern. Second, we performed replica exchange DMD simulations with the additional tertiary structure constraints derived from HMX scores. Finally, we selected the 100 structures with the lowest energies and highest correlations between the structure and the experimentally-derived HMX scores. We incorporated HMX information in terms of solvent accessibility, using an approached previously developed to model hydroxyl radical probing.[23] The relative solvent

accessibility was interpreted as the number of through-space contacts between sugar pseudo-atoms. Positions with high HMX scores were allowed a larger number of tertiary contacts than those with low scores.[23] We assigned two biasing potentials based on the number of tertiary contacts as calculated from the HMX score. The first term is an attractive potential that collapses the RNA to achieve a compact structure, the second term is a repulsive potential assigned to each nucleotide position if it exceeds a specified threshold of tertiary contacts. The minimum and maximum thresholds for number of tertiary contacts were 0.5 and 11, respectively[23]. Nucleotides with a number of contacts above the maximum threshold were given an attractive potential and those with a value below the minimum threshold were given a repulsive potential.

2.4.9. Replica exchange DMD simulations and consensus structure modeling.

We performed replica exchange DMD simulations for each RNA system using twelve replicas with temperatures of 0.200, 0.215, 0.230, 0.246, 0.262, 0.277, 0.293, 0.311, 0.330, 0.350, 0.375, and 0.400 kcal/(mol·k_B). We set each exchange event to occur at 1000 DMD time steps according to a Metropolis based Monte Carlo algorithm. For each replica, we performed a 5×10^5 DMD time step simulation. We then generated snapshots of the structures from each simulation, one snapshot at every 100 DMD time units. From these snapshots, we selected 1000 snapshots with the lowest energies, then calculated the correlations between the number of tertiary contacts and the relative solvent accessibilities derived from the HMX scores for each snapshot and selected the 100 structures with lowest (negative) correlations. We then performed the same selection procedure in reverse order, first selecting the 1000 structures with the lowest correlations based on the experimentally-determined HMX scores and then selecting the 100 structures with lowest energies. These

200 structures were then ranked by energies and structure reactivity correlation coefficients and 100 structures were chosen to represent the final structural ensemble.[23] These final ensembles were clustered by hierarchical clustering. For each RNA, the clustering cutoff was either 4 Å or three quarters of the average RMSD as the function of RNA length length.[25]

2.5. REFERENCES

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

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

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

(4) Dethoff, E. A., Chugh, J., Mustoe, A. M., and Al-Hashimi, H. M. (2012) Functional complexity and regulation through RNA dynamics. *Nature 482*, 322–330.

(5) Conway, L., and Wickens, M. (1989) Modification interference analysis of reactions using RNA substrates, in *Methods Enzymol.*, pp 369–379. Academic Press. Inc.

(6) Clarke, P. A. (1999) RNA footprinting and modification interference analysis, in *Methods Mol. Biol.* (Haynes, S., Ed.), pp 73–91. Humana Press.

(7) Merryman, C., and Noller, H. F. Footprinting and modification-interference analysis of binding sites on RNA, in *RNA:protein interactions, a practical approach*, pp 237–253. Oxford University Press.

(8) Ryder, S. P., and Strobel, S. A. (1999) Nucleotide Analog interference mapping. *Methods 18*, 38–50.

(9) Strobel, S. A. (1999) A chemogenetic approach to RNA function/structure analysis. *Curr. Opin. Struct. Biol. 9*, 346–352.

(10) Tullius, T. D., and Greenbaum, J. A. (2005) Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr. Opin. Chem. Biol. 9*, 127–134.

(11) Pastor, N., Weinstein, H., Jamison, E., and Brenowitz, M. (2000) A detailed interpretation of OH radical footprints in a TBP-DNA complex reveals the role of dynamics in the mechanism of sequence-specific binding. *J. Mol. Biol. 304*, 55–68.

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

(14) McGinnis, J. L., Dunkle, J. A., Cate, J. H. D., and Weeks, K. M. (2012) The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc. 134*, 6617–6624.

(15) Westhof, E., Dumas, P., and Moras, D. (1988) Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalaine transfer RNA. *Acta. Cryst. A44*, 112–123.

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

(17) Dann, C. E., III, Wakeman, C. A., Sieling, C. L., Baker, S. C., Irnov, I., and Winkler, W. C. (2007) Structure and Mechanism of a Metal-Sensing Regulatory RNA. *Cell 130*, 878–892.

(18) Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science 2733*, 1678–1685.

(19) Bhattacharyya, A., Murchie, A. I. H., and Lilley, D. M. J. (2002) RNA bulges and the helical periodicity of double-stranded RNA. *Nature 343*, 484–487.

(20) Lesnik, E. A., and Freier, S. M. (1998) What affects the effect of 2“-alkoxy modifications? 1. Stabilization effect of 2-”methoxy substitutions in uniformly modified DNA oligonucleotides. *Biochemistry 37*, 6991–6997.

(21) Gherghe, C. M., Leonard, C. W., Ding, F., Dokholyan, N. V., and Weeks, K. M. (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.

(22) Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA 14*, 1164–1173.

(23) Ding, F., Lavender, C. A., Weeks, K. M., and Dokholyan, N. V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat. Methods 9*, 603–608.

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

(25) Hajdin, C. E., Ding, F., Dokholyan, N. V., and Weeks, K. M. (2010) On the significance of an RNA tertiary structure prediction. *RNA 16*, 1340–1349.

(26) Lesnik, E. A., Guinosso, C. J., Kawasaki, A. M., Sasmor, H., Zounes, M., Cummins, L. L., Ecker, D. J., Cook, P. D., and Freier, S. M. (1993) Oligodeoxynucleotides containing 2'-O-modified adenosine: synthesis and effects on stability of DNA:RNA duplexes. *Biochemistry 32*, 7832–7838.

(27) Kulshina, N., Edwards, T. E., and Ferre-D'Amare, A. R. (2009) Thermodynamic

analysis of ligand binding and ligand binding-induced tertiary structure formation by the thiamine pyrophosphate riboswitch. *RNA 16*, 186–196.

(28) Karabiber, F., McGinnis, J. L., Favorov, O. V., and Weeks, K. M. (2012) QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA 19*, 63–73.

(29) Hamelryck, T., and Manderick, B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics 19*, 2308–2310.

(30) Davis, I. W., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acid Res. 32*, W615–W619.

CHAPTER 3. HMX REVEALS TERTIARY INTERACTIOS WITHIN MULTIPLE STABLE SUBSTRUCTURES OF THE *TETRAHYMENA* GROUP I INTRON

3.1. INTRODUCTION

Like proteins, large catalytically-active RNA molecules are commonly composed of multiple independent substructures.[1] Identification and characterization of these substructures are key to understanding the roles of these large RNA.[2,3] Substructures for catalytic RNA such as the RNase P and group I intron RNA have been identified through chemical probing,[4] and comparative sequence analysis[5] respectively. These domains have been shown to fold independently of the full length RNA and form native structures resembling those found in the active RNA.[6,7]

High resolution crystal structures have been solved for both the full length *Tetrahymena* group I intron as well as the P4-P6 subdomain.[8,9] The structure of the group I intron consists of two sets of coaxially stacked helices, the P4-P6 and P3-P9 domains, that form the active site of the RNA and the P1-P2 helix which contains the RNA spice site. Folding of the catalytically active RNA relies on formation of the P4-P6 domain to direct the folding of the P3-P9 domain and allow for the P1 helix to dock into the active site.[10]
HMX experiments provide a unique view of tertiary structure by detecting regions of high nucleotide density in stably folded RNA.[11] HMX uses electrophilic SHAPE reagents which form bulky adducts at the 2'-hydroxyl position of a denatured RNA sample. In a HMX experiment, a pool of modified RNA is partitioned by their ability to form tertiary structures under native folding conditions using native gel electrophoresis. Adducts that prevent folding will be over-represented in the unfolded RNA population. By comparing unfolded to folded

populations of RNA, HMX experiments identify nucleotides involved in tertiary interactions which map to high atomic density regions in the RNA.

This study presents a single probing technique that can accurately identify individual domains with in *Tetrahymena* group I intron. We use the HMX technique to probe the well-characterized group I intron in order to identify stable structural domains and to gain further insights into the tertiary interactions involved in the complete intron.

3.2. RESULTS

The group I intron RNA was modified under heat denaturing conditions using N-methylisatoic anhydride (NMIA).[12,13] Next, the RNA was then folded under native folding conditions and folded and unfolded states were separated using native gel electrophoresis. Folded RNA are more compact and thus migrate more rapidly through an acrylamide gel. Modified group I intron RNA separated into four distinct states (Figure 3.1A). The lowest, most compact, band represents the fully folded RNA. Each slower moving band represent RNA with decreasing levels of tertiary structure. Unmodified RNA also separated into four distinct bands; however, in the unmodified RNA the top three bands were a significantly smaller fraction of the sample than for the modified RNA (Figure 3.1B). The increase in the slow moving band populations between the modified and unmodified samples indicates that NMIA adduct disrupts cause disruption of RNA tertiary structure formation.

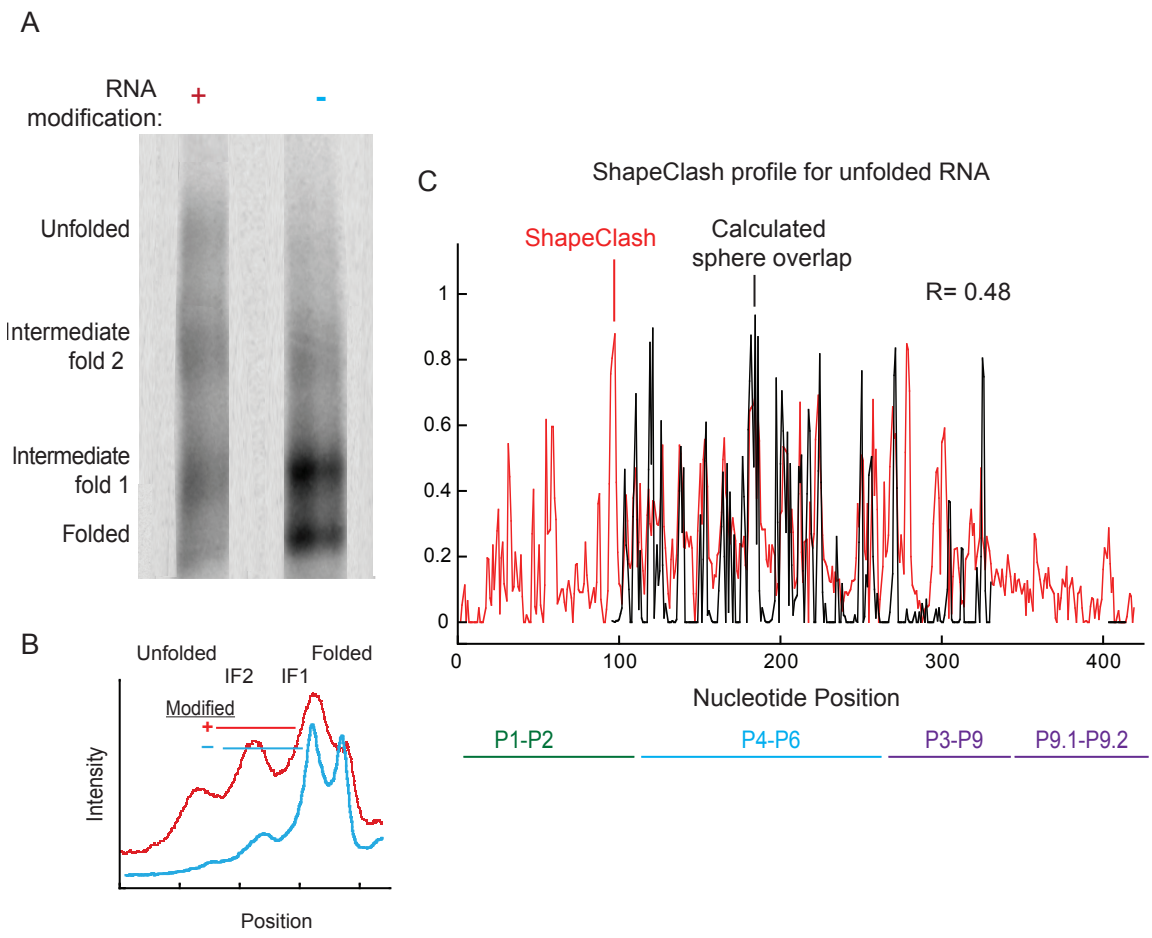


Figure 3.1 Identification of multiple domains with in the group I intron by ShapClash. (**A**) Folded and unfolded along with intermediate fold populations for modified and unmodified RNA were separated on a 0.5x TB polyacrylamide gel containing 50 mM NaCl and 5 mM $MgCl_2$. For clarity, gel images were straightened and scaled to provide similar representations for each RNA. (**B**) Band intensities were measured for each lane. The intensity of the unfolded and intermediate bands was larger in the modified sample compared to the unmodified sample for each RNA. (**C**) Correlation between HMX data and calculated sphere overlap related to RNA structure. Cross correlation normalized HMX score profiles (red) show strong correlation when compared to calculated sphere overlap (black) for the group I intron. Sections of the RNA that are not present in the crystal structure are shown as gas in the calculated sphere overlap trace.

Standard HMX analysis compares the unfolded and folded bands to create HMX scores for each position in the RNA. The higher the HMX score the more likely the nucleotide is to disrupt RNA tertiary structure. Calculation of a HMX intensity plot between the top, most unstructured RNA to the fully folded band should represent all positions in high atomic density regions of the RNA structure (Fig. 3.1C, red trace). The quality of HMX score, as it relates to the accepted crystal structure, can be evaluated by comparison to a sphere overlap calculation. This calculation uses the accepted crystal structure to estimate the degree to which an adduct can disrupt RNA structure (Fig. 3.1C, black trace). While the entire sequence was used to crystallize the group I intron the P1-P2 helix along with sections of the P9 helix could not be well resolved and so are absent in the final crystal structure.[8] The sphere overlap calculation shows a modest correlation with the HMX score. However, agreement between predicted and experimental data is much higher in the P4-P6 domain (r = 0.63). Additional intensities in the HMX profile could reflect interactions between regions of the RNA that cannot be crystallized.

HMX profiles can also be created for each intermediate band for each band can by subtracting the fully folded profile from the intermediate band profile. Since band migration in native gels is related to the degree to which an RNA is unstructured, the fastest migrating intermediate band should be the most native like structure. The HMX profile of this band indicates that only a few positions have a high HMX score (Fig. 3.2A). These positions are centralized with in the P1-P2 region and part of the P3-P9 domain. The positions highlighted in this profile are indicated on the secondary structure representation of the group I intron crystal structure (Fig. 3.2D, green circles). These nucleotides are centralized around the intron active site. The HMX profiles shows that the intermediate band represents a structure

in which the P4-P6 and P3-P9 domain is fully formed without the P1-P2 domain docked in the active site. The nucleotides identified in this profile show the key tertiary interactions that must be formed to allow for docking of the P1-P2 helix with in the active site of the RNA.

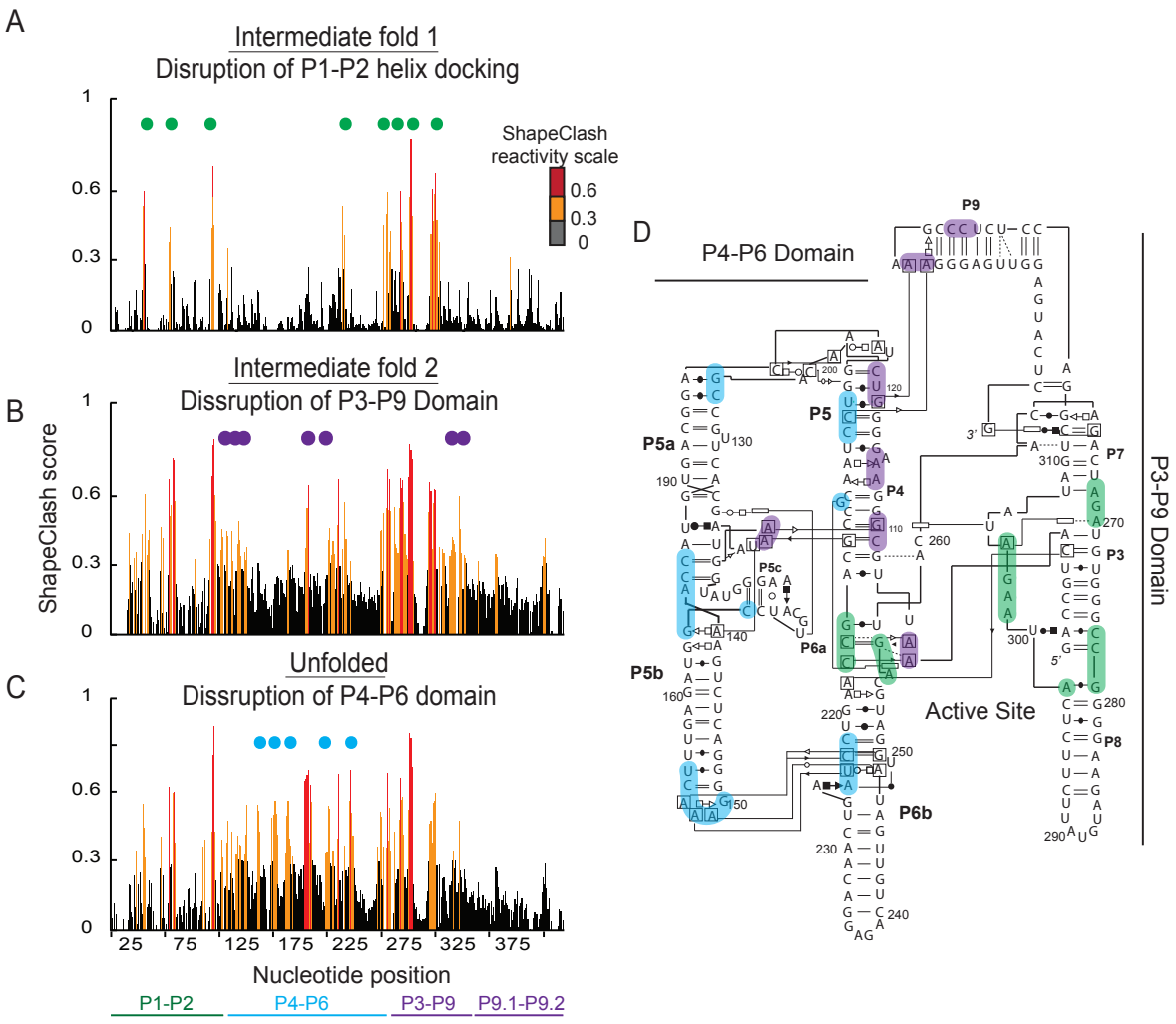


Figure 3.2 Analysis of intermediate structures in Group I intron. HMX score profiles were created for the (**A**) first and (**B**) second intermediate bands as well as for the (**C**) unfolded band. Positions with low reactivity below 0.3 are colored in black; positions of medium reactivity (between 0.3 and 0.6) are colored in orange; and positions of high reactivity (greater than 0.6) are colored in red. Reactive positions in the first intermediate profile are marked by green circles and. Positions that become reactive in the second intermediate and unfolded profile are marked in purple and blue respectively. (**D**) Highlighted positions on the HMX profiles are superimposed on the secondary structure representation of the crystal structure.[15]

Analysis of the second intermediate band reveals a profile with many more reactive nucleotides (Fig 3.2B). While nucleotides found in the first intermediate band remain reactive, previously uncreative nucleotides in the P9 helix and select positions with in the P4-P6 domain also become reactive. Positions that become reactive with in the second intermediate band are indicated on the secondary structure (Fig. 3.2D, purple circles). These nucleotides map to the junction between the P4-P6 and P3-P9 domains. The second intermediate band represents a structure in which only the P4-P6 domain remains intact and the reactive positions indicate nucleotides involved in the docking of the P3-P9 domain around the P4-P6 domain.

Only in the fully unfolded band the P4-P6 domain becomes reactive (Fig. 3.2A). These added reactivities indicate that the P4-P6 has become disrupted along with the P3-P9 and P1-P2 domains. The fully unfolded band represents all interactions that can be disrupted by the NMIA adduct. By identifying uniquely reactive positions in each band we can indentify the specific interactions in which each nucleotide is involved.

3.3. DISCUSSION

HMX analysis of the group I intron indicates that the RNA can form multiple, stable, partially formed structures. These structures show that the group I intron can adapt a conformation in which the P4-P6 and P3-P9 domains are fully folded with out the P1-P2 helix docked into the active site. The RNA can also form a structure in which only the P4-P6 structure is formed. These results support previous studies that describe the group I intron folding pathway in which the P4-P6 folds into its native tertiary structure which then directs the organization of the P2-P9 domain and allow for the docking of the P1-P2 helix into the active site.[10]

This study shows the versatility of the HMX experiment. HMX can provide tertiary information for large RNA that are not easily amenable to high-resolution techniques. HMX provides tertiary information for nucleotides in the group I intron that could not be resolved in the crystal structure. The lack of crystallographic data in the P3-P9 and P1-P2 regions leads to an incomplete representation of the RNA structure. The crystal structure is not able to provide information for the P1-P2 region docking with in the RNA active site, resulting in an incomplete understanding of splice site recognition and cleavage. HMX however, identified the specific nucleotides involved in P1-P2 helix docking with in the active site. HMX scores also provides an advantage over other biochemical techniques. The group I intron has been studied extensively by hydroxyl radical footprinting.[6] Only through the systematic deletion of specific regions in the RNA could the independently folding P4-P6 domain be identified. Through native gel separation of group I intron, multiple substructures in the group I intron could be identified. Furthermore HMX provides tertiary information for the entire group I intron as well as identifies the specific interactions in which each nucleotide is involved.

This single experiment was able to identify multiple stable substructures with in a single pool of RNA. Furthermore we were able to obtain singe nucleotide resolution tertiary information for each of these structures. The HMX experiment and the analysis outlined above can be useful to gain insight into the tertiary structure and folding pathway of many other large RNA.

3.4. METHODS

3.4.1. RNA constructs

The DNA template for the *Tetrahymena* group I intron RNA was imbedded within 5' and 3' structure cassette flanking sequences[13] and was generated by PCR. The RNA was transcribed *in vitro* [1 mL; 40 mM Tris (pH 8.0), 10 mM $MgCl_2$, 10 mM dithiothreitol, 2 mM spermidine, 0.01% (v/v) Triton X-100, 4% (w/v) poly(ethylene) glycol 8000, 2 mM each NTP, 50 μL PCR-generated template, 0.1 mg/mL T7 RNA polymerase; 37 °C; 4 h] and purified by denaturing polyacrylamide gel electrophoresis [8% polyacrylamide, 7 M urea, 29:1 acrylamide:bisacrylamide, 0.4 mm × 28.5 cm × 23 cm gel; 32 W, 1.5 h]. The RNA was excised from the gel, recovered by passive elution overnight at 4 °C, and precipitated with ethanol. The purified RNA was then resuspended in 50 μL TE and stored at -20 °C.

3.4.2. 5'-[^{32}P] RNA radiolabeling

Purified RNA was first dephosphorylated [300 μL; 50 mM Tris (pH 8.5), 0.1 mM EDTA, 10 μM RNA, 300 units SUPERase-In (Ambion), 200 units alkaline phosphatase (Roche); 50 °C; 1 h]. The RNA was then purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation and then resuspended in TE (storage at - 20 °C). The RNA was 5' -[^{32}P]-radiolabeled by treatment with T4 polynucleotide kinase [40 μL; 80 pmol dephosphorylated RNA, 70 mM Tris (pH 7.6), 10 mM $MgCl_2$, 5 mM DTT, 2 μL T4 polynucleotide kinase (NEB, 10,000 units/mL), 150 μCi [γ -^{32}P]-ATP; 37 °C; 30 min]. Radiolabeled RNA was purified by denaturing (8%) gel electrophoresis, excised from the gel, and recovered by overnight passive elution at 4 °C. The purified 5'-[^{32}P]-labeled RNAs were precipitated with ethanol, resuspended in 5 mM Tris (pH 7.5), and stored at -20 °C. The 5'-[^{32}P]-labeled RNA was resuspended in enough 5 mM Tris buffer so that the amount of ^{32}P

in 1 μL of resuspended RNA measured approximately 1 million cpm.

3.4.3. Denatured RNA modification

RNA was denatured by heating to 90 °C for 2 min [32 μL; 30 pmol unlabeled RNA, 1.5 cpm 5'-[^{32}P]-radiolabeled RNA, 100 mM HEPES (pH 8.0)]. The denatured RNA was added to NMIA (1.2 μL, 0.4 M in DMSO) and allowed to react at 95 °C for 5 min. The modification process was repeated three times. After the third modification the sample was placed on ice. A no-modification control reaction was performed identically using 1.2 μL DMSO. Any water evaporated during the modification was replaced to bring the volume to 36 μL. For experiments in which band populations were quantified and visualized, 100,000 cpm of 5'-[^{32}P]-radiolabeled RNA was used per condition. Gels were visualized using a phosphorimager.

3.4.4. RNA folding and native gel separation

After the denatured RNA was modified it was treated with 4 μL 10x folding buffer (100 mM $MgCl_2$, 1 M NaCl), and incubated at 37 °C for 30 min. The folded RNA sample was immediately added to an 80% glycerol solution containing bromophenol blue and xylene cyanol and loaded on to a native polyacrylamide gel [8% polyacrylamide, 19:1 acrylamide:bisacrylamide, 0.5x TB (45 mM Tris, 45 mM boric acid), 50 mM NaCl, 5 mM Mg_2Cl; 0.4 mm × 28.5 cm × 23 cm gel; 20 W, 8 h]. The gel was run in a cold room kept at 4 °C to ensure that the gel temperature did not rise above 37 °C. To minimize the effect of a salt front on separation, the anode and cathode buffer wells were periodically refreshed. The bands were visualized by exposing the gel for 1 hr to Kodak BioMax maximum sensitivity film. The film was used as a template to identify and guide excision of the unfolded and folded band from the gel. The samples were recovered by passive elution overnight at 4 °C

and were purified by ethanol precipitation and resuspended in 10 μL water.

3.4.5. Reverse transcription and adduct detection

The general procedure was outlined previously.[13] DNA primers were 5'-end labeled with VIC or NED fluorophores (Applied Biosystems). RNA extracted from native gel separation (10 μL) was added to a fluorescently labeled DNA primer (5' -VIC-labeled GAA CCG GAC CGA AGC CCG; 3 μL, 0.3 μM) and allowed to anneal at 65 °C for 6 min and then cooled on ice. Reverse transcription buffer [6 μL; 167 mM Tris (pH 8.3), 250 mM KCl, 10 mM $MgCl_2$, 1.67 mM each dNTP] and Superscript III (1 μL, 200 units) were added, and samples were incubated at 45 °C for 2 min, 52 °C for 20 min, and 65 °C for 5 min. The reactions were quenched with 4 μL 50 mM EDTA. The cDNAs were recovered by ethanol precipitation, washed twice with 70% ethanol, dried in a SpeedVac for 5 min, and resuspended in 10 μL deionized formamide. Dideoxy sequencing ladders were produced using unlabeled, unmodified RNA by annealing a 5' -NED-labeled fluorescently labeled DNA primer (3 μL, 0.3 μM), and by adding 1 μL 2',3' -dideoxycytosine (10 mM) triphosphate before addition of Superscript III. cDNA fragments were separated by capillary electrophoresis using an Applied Biosystems 3130 DNA sequencing instrument. Raw capillary electrophoresis traces were analyzed using QuSHAPE.[14]

3.5. REFERENCES

(1) Batey, R. T., Rambo, R. P. & Doudna, J. A. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed* 38, 2326–2343 (1999).

(2) Reiter, N. J., Chan, C. W. & Mondragón, A. Emerging structural themes in large RNA molecules. *Curr. Opin. Chem. Biol.* 21, 319–326 (2011).

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

(4) Loria, A. & Pan, T. Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA* 2, 551–563 (1996).

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

(6) Murphy, F. L. & Cech, T. R. An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. *Biochemistry* 32, 5291–5300 (1993).

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

(8) Golden, B. L. A preorganized active site in the crystal structure of the Tetrahymena ribozyme. *Science* 282, 259–264 (1998).

(9) Cate, J. H. et al. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 2733, 1678–1685 (1996).

(10) Doherty, E. A. & Doudna, J. A. The P4-P6 domain directs higher order folding of the Tetrahymena ribozyme core. *Biochemistry* 36, 1–11 (1997).

(11) Homan, P. et al. RNA tertiary structure analysis and refinement by 2'-hydroxyl molecular interference**.** *Submitted (2014)*.

(12) Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* 127, 4223–4231 (2005).

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

(14) Karabiber, F., McGinnis, J. L., Favorov, O. V. & Weeks, K. M. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* 19, 63–73 (2012).

(15) Lescoute, A. & Westhof, E. The interaction networks of structured RNAs. *Nucleic Acids Research* 34, 6587–6604 (2006).

CHAPTER 4. TERTIARY STRUCTURE REVEALED BY SINGLE-MOLECULE CORRELATED CHEMICAL PROBING OF RNA*

4.1. INTRODUCTION.

RNA plays a central role in gene expression and regulation. These functions are mediated by tiered levels of information, from simple primary sequence to higher-order structures that govern interactions with ligands, proteins, and other RNAs[1,2]. Many RNAs can also form more than one stable structure, and these distinct conformations often have different biological activities[3,4]. Currently, the rate of describing new RNA sequences vastly exceeds abilities to examine their structures. Here we characterize tertiary interactions and multiple conformations in single RNAs by melding chemical probing and massively parallel sequencing. As massively parallel sequencing reports the sequences of single templates, each read is fundamentally a single-molecule observation[8]. We first modified RNA with a reagent that is sensitive to the underlying RNA structure and then detected multiple adducts in individual RNA strands (Fig. 4.1). Chemical adducts were detected as sequence mutations based on their ability to induce misreading of the template nucleotide by a reverse transcriptase enzyme[5-7]. Single-molecule probing data were used to detect correlated (through-space) RNA modifications to identify tertiary interactions (Fig. 4.1A) and to examine multiple conformations in single in solution ensembles (Fig. 4.1B).

*This chapter has been submitted for publication in PNAS.

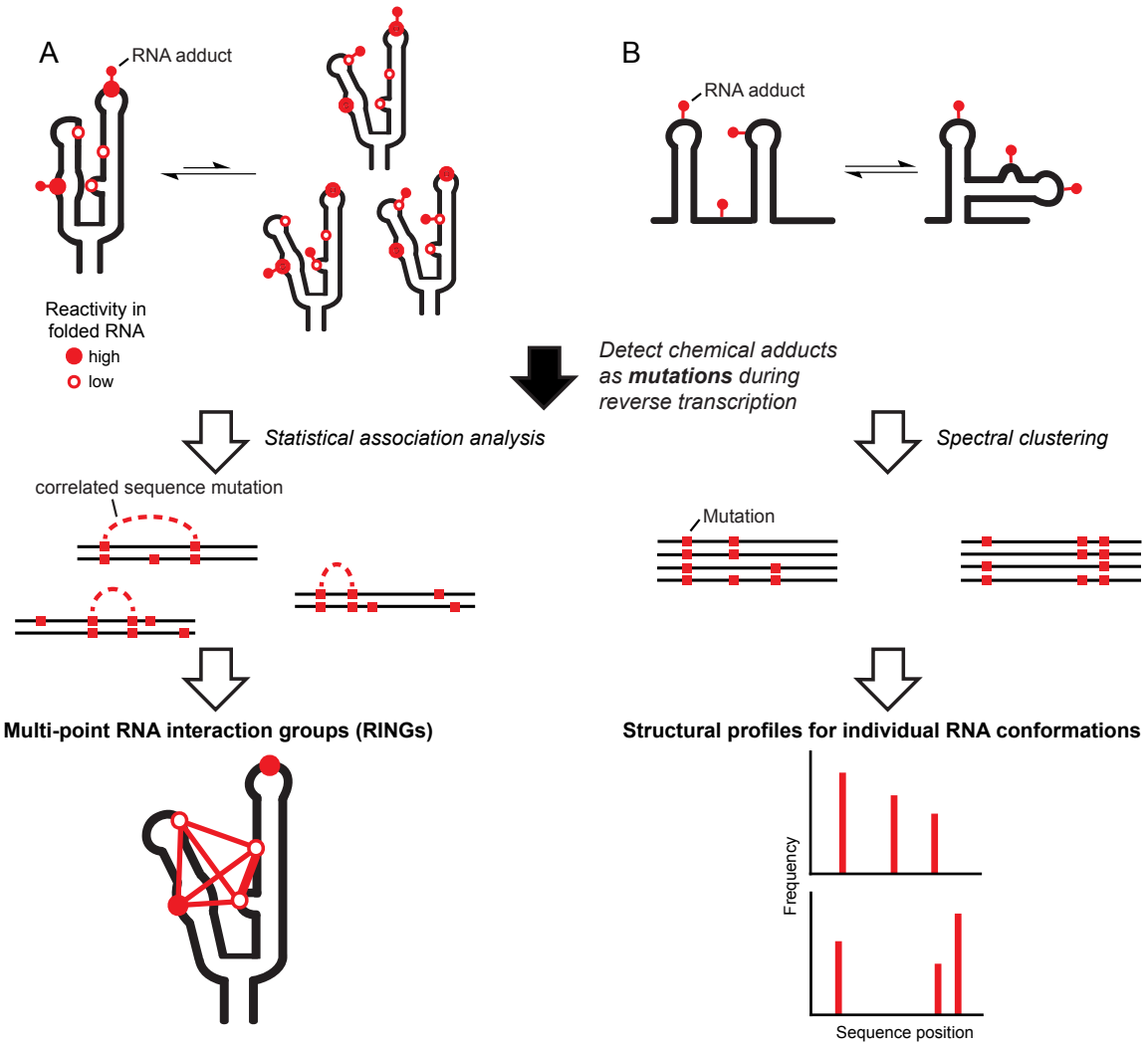
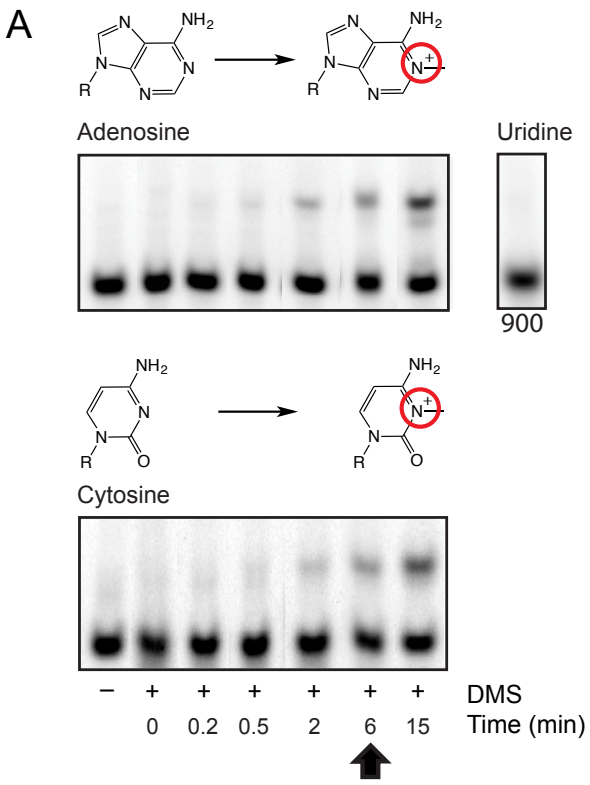


Figure 4.1 Single-molecule RNA structure analysis by massively parallel sequencing. (**A**) RNA molecules experience local structural variations and 'breathing' in which regions of an RNA structure become reactive to a chemical probe in a correlated way. Nucleotides that interact (open red circles) show correlated reactivity. Statistical association analysis is used to detect and quantify the strengths of these interdependencies, ultimately revealing multi-point RNA interaction groups or RINGs. (**B**) In solution, RNAs often adopt multiple conformations. Spectral clustering analysis based on similarity of nucleotide reactivity patterns was used to separate data on individual RNA stands into different conformations.

4.2. RESULTS

4.2.1. Multi-site DMS reactivity with RNA.

We used dimethyl sulfate (DMS) to probe the structures of three RNAs: the *E. coli* thiamine pyrophosphate (TPP) riboswitch (79 nucleotides), which binds the TPP ligand when functioning in gene regulation[9]; the *Tetrahymena* group I intron P546 domain (160 nts)[10]; and the *Bacillus stearothermophilus* RNase P catalytic domain (265 nts)[11]. DMS forms adducts at the N1 position of adenosine and N3 position of cytosine (Fig. 4.2A). We optimized conditions to yield multiple modifications in an RNA strand without disrupting native-like RNA folding (Fig. 4.3). RNA samples were treated with 170 mM DMS in 10 mM Mg^{2+} and 300 mM cacodylate buffer at pH 7 for 6 min. The reactions were quenched by addition of excess 2-mercaptoethanol. Cytosine and adenosine nucleotides were methylated with equal efficiencies, and ~12% of nucleotides were modified under these conditions (Fig. 4.2B).



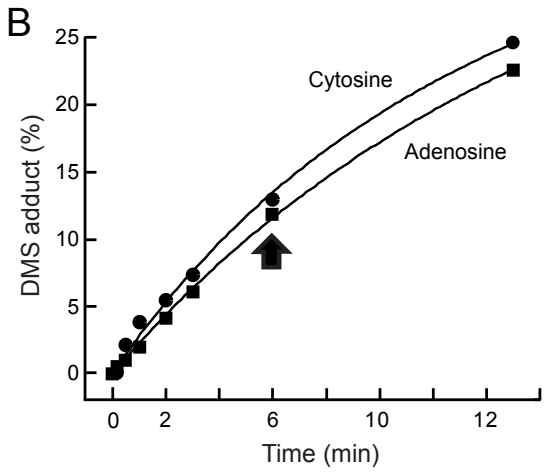


Figure 4.2 Efficient DMS adduct formation at the base-pairing faces of adenosine and cytosine. (**A**) Reaction of radioactively labeled nucleotides with 170 mM DMS in 300 mM cacodylate (pH 7) monitored by gel electrophoresis. (**B**) Time-course of DMS reaction with adenosine and cytosine. Unconstrained nucleotides react to form methyl adducts with ~12% efficiency (arrow).

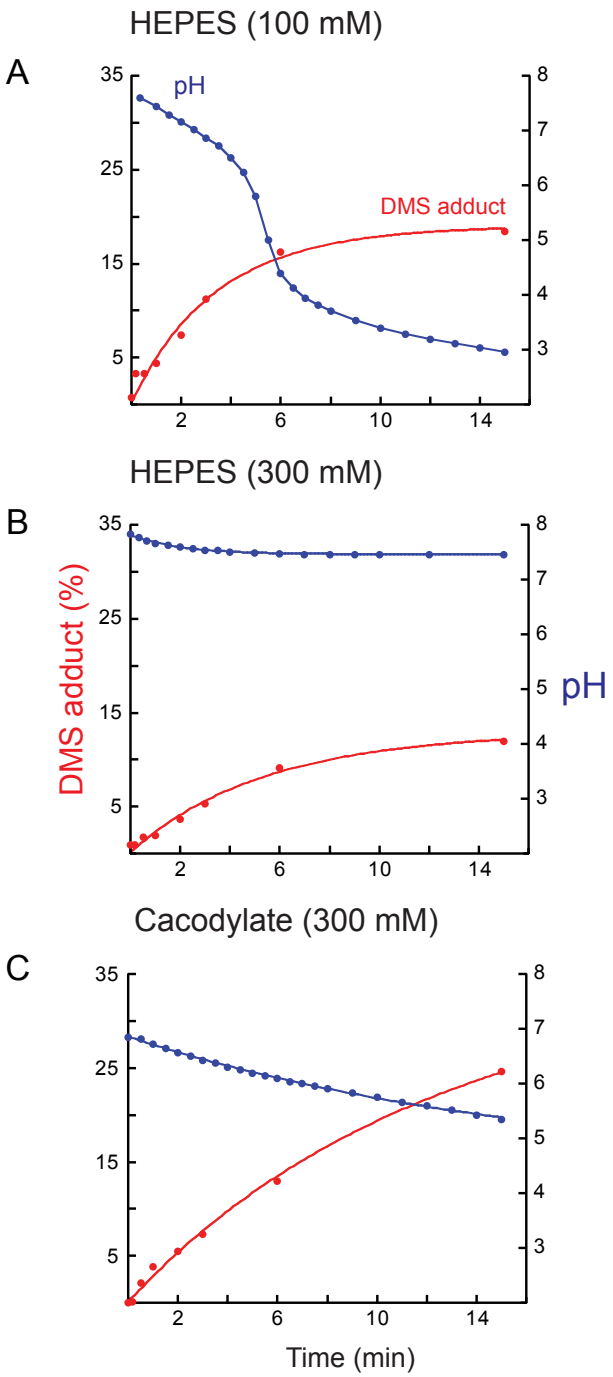


Figure 4.3 Optimization of DMS adduct formation. The pH of the DMS reaction (blue lines) and adduct formation with adenosine (red) were monitored as a function of time. The DMS concentration was 170 mM. (**A**) In 100 mM HEPES (pH 8.0), pH dropped over time, quenching the DMS reaction. These conditions closely resemble those widely employed in conventional DMS experiments. (**B**) Reactions performed in 300 mM HEPES (pH 8.0) limited the pH drop; however, the organic buffer reacted directly with DMS, quenching the reaction with adenosine. (**C**) The pH in reactions performed in 300 mM cacodylate was well-controlled, and the buffer did not react with DMS to quench the reaction.

We detected sites of DMS methylation as adduct-induced mutations in the cDNA generated during reverse transcription (Fig. 4.1). For the TPP riboswitch, the P546 domain, and the RNase P domain, averages of 2, 5, and 7 adducts were detected in each sequencing read, respectively (Fig. 4.4A). Approximately 15% of the single-stranded A and C nucleotides in each RNA were modified by DMS, comparable to the level of modification in free nucleotides.

We visualized the overall reactivity patterns for each RNA in two-dimensional mutation frequency profiles (Fig. 4.4B). The observed high mutation frequencies over background allowed DMS-induced mutations to be analyzed without the requirement for background correction. Comparison to the high-resolution structures for these RNAs[9-11] shows nucleotides modified at high levels are those that are not involved in base pairing or tertiary interactions (Fig. 4.4C).

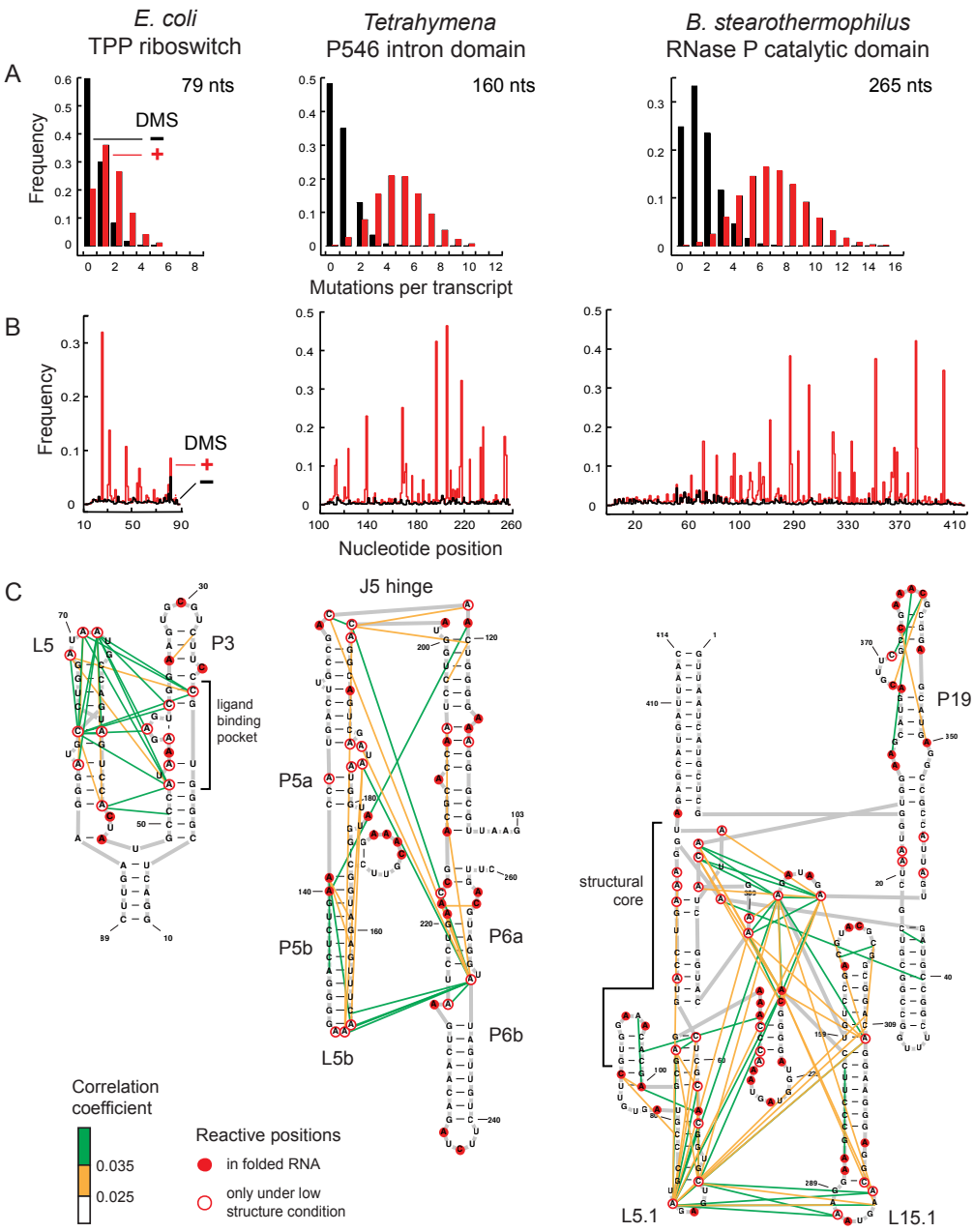


Figure 4.4 RING analysis of RNA structure. (**A**) Number of mutations per transcript detected by reverse transcription with (red) and without (black) DMS modification. (**B**) DMS modification induced mutation frequencies as a function of nucleotide position. Data from DMS-treated samples are shown in red and no-reagent controls are shown in black. (**C**) RINGs for the TPP riboswitch, P546 domain, and RNase P RNAs showing strong (green) and moderate (yellow) correlations. Correlations occur between positions that are reactive in the native structure (filled red circles) or become reactive during 'breathing' motions (open circles), reflecting the structural breathing component of reactivity interdependencies. Correlation coefficients of 0.025 and 0.035 correspond to median increases in correlated mutations of 2.5- and 2.8-fold, respectively (Fig. 4.12C). Secondary structures are drawn to approximate relative helical orientations in three-dimensionalspace based on known structures[9-11].

4.2.2. Through-space RNA interactions detected by statistical association analysis.

Nucleotides involved in through-space interactions will show correlated chemical reactivities, reflective of a 'breathing' mechanism in which an RNA nucleotide becomes transiently accessible for modification. This breathing mechanism suggests that correlated probing will be highly selective for transient, dynamic interactions rather than static structural differences (Fig. 4.1A). These interdependencies were readily quantified because multiple chemical modification events were detected in sequencing reads of single RNA strands. We call these nucleotides that are modified in a correlated fashion RNA interaction groups or RINGs. We used a two-part strategy to identify reactive nucleotide pairs with statistically significant correlations and to quantify the strengths of these correlations. The interdependencies for DMS reactivities for any two positions in a single RNA strand were evaluated using a chi-square test. The strength of the interaction between each pair of correlated nucleotides was then quantified using Pearson's phi metric.

RINGs for the TPP riboswitch, P546 intron domain, and RNase P RNAs (Fig. 4.4C) include nucleotides known to interact based on high-resolution structures[9-11]. For example, correlated positions in the TPP riboswitch corresponded to nucleotides involved in a docking interaction between the L5 loop and P3 helix and in formation of the ligand-binding pocket. In the P546 domain, correlated modifications were observed at nucleotides in the L5b loop and P6a helix docking interaction, within the J5 hinge region, and through the length of the P5a and P5b helices. In the RNase P RNA, RINGs report tertiary interactions between the L5 and L15.1 loops and in the structural core. We also observed a second set of interactions in the RNase P P19 element.

4.2.3. Through-space RNA interactions detected by statistical association analysis.

We evaluated differences in the tertiary interactions, as reported by RINGs, under different solution conditions or in mutant RNAs. In the presence of Mg^{2+}, the P546 domain forms a U-shaped structure in which a tetraloop-receptor interaction forms between L5b and P6a and the J5 region acts as a hinge[12,13]. These interactions are correctly reported by multiple correlated chemical modifications in these structural elements (Fig. 4.5A). Disruption of this tertiary structure by folding the RNA in the absence of Mg^{2+} eliminated the majority of observed interactions (Fig. 4.5B).

The tertiary structure of the P546 domain can also be perturbed by mutations in the P6a helix and in the J5 hinge. Mutation of the C223-G250 base pair in the P6a helix to A-U disrupts the L5- P6a interaction[12]. RING analysis of this mutant showed that the correlation between L5 and P6b was lost and that other parts of the RNA also underwent significant reconfiguration. Interactions involving the hinge appeared to be strengthened and more and stronger correlations were observed within the helical domains of P5a and P5b (Fig. 4.5C). Mutations which base pair nucleotides in the J5 hinge likely yield a relatively linear conformation for the P546 domain[13]. RING analysis of the J5 mutant showed the expected loss in correlated nucleotides within the J5 region and consequently the loss of the L5b-P6a interaction (Fig. 4.5D). The correlations among nucleotides in the P5b helix were strengthened in this mutant relative to those in observed in the wild-type RNA, but no changes were observed in correlations among nucleotides in the P5a helix. This analysis of P546 domain confirmed that RINGs accurately reflect structural interactions in an RNA molecule at nucleotide resolution.

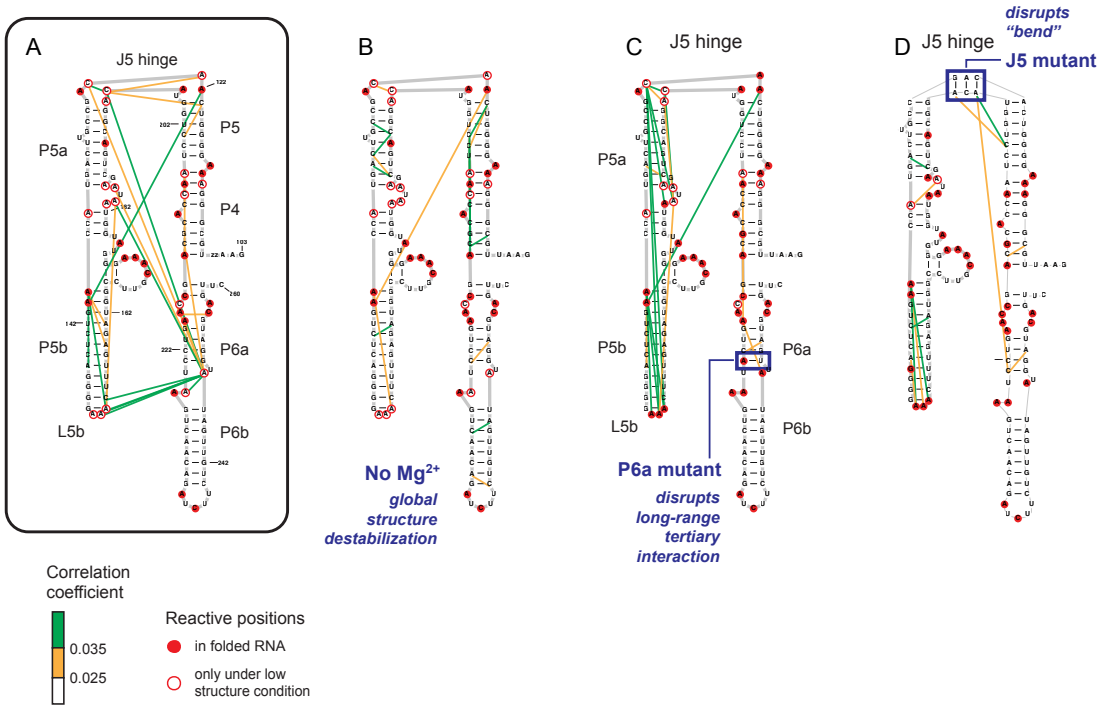


Figure 4.5 RINGs report the tertiary structure of the P546 domain and mutation-induced structural changes. Strong and medium internucleotide correlations are shown with green and yellow lines, respectively. (**A**) RINGs in the P546 domain folded in the presence of Mg^{2+}. (**B**) RINGs in the P546 domain folded in the absence of Mg^{2+}. (**C**) RINGs in the P6a mutant. (**D**) RINGs in the J5 hinge mutant. For clarity, panel A is identical to Figure 4.4C.

4.2.4. Multiple RNA conformations detected by spectral clustering.

Each RNA strand is sequenced independently in a massively parallel sequencing experiment. RNA strands of different conformations will tend to exhibit distinct groups of co-reactive nucleotides (Fig. 4.1B). Such groups can be detected by spectral clustering and will be reflective of distinct, relatively stable individual structures in solution. Spectral clustering produces an objective estimate of the number of clusters and therefore the number of distinct conformations adopted by a particular RNA in solution. Spectral clustering analysis of the modification data obtained on the TPP riboswitch and RNase P RNAs indicated that each RNA formed multiple distinct conformations under the conditions used in our probing experiments (Table 4.1).

Table 4.1 Summary of spectral clustering analysis for multiple RNA conformations in single ensembles. Clustering analysis is summarized for the TPP riboswitch, P546 domain, P546 mutants, and the RNase P RNA as a function of different levels of structure. The eigengap value measures the structural difference between clusters; samples with eigengaps greater than 0.03 are taken to have two (or more) distinct clusters. The population of each cluster is given in the last column with the most highly structured cluster listed first. An asterisk indicates that smallest cluster population for the TPP riboswitch, at saturating ligand concentration, was too small to accurately generate DMS reactivity profiles. For analysis, the sample was therefore clustered into two conformations with populations of 81% and 19%.

	Condition	Eigengap	Clusters	Population (%)
TPP	Saturating ligand	0.027 / 0.035	2	75 : 20 : 5*
	No ligand	0.033	2	17 : 83
	No ligand + no Mg^{2+}	0.024	1	–
	200 nM Ligand	0.056 / .044	3	32 : 31: 37
P546 domain Native	Mg^{2+}	0.013	1	–
	no Mg^{2+}	0.005	1	–
P6a mut	Mg^{2+}	0.014	1	–
	no Mg^{2+}	0.003	1	–
J5 mut	Mg^{2+}	0.002	1	–
	no Mg^{2+}	0.009	1	–
RNase P	Mg^{2+}	0.047	2	76 : 24
	no Mg^{2+}	0.023	1	–

RINGs identified for the TPP riboswitch with saturating ligand revealed interactions in the L5-P3 docked structure and in the ligand-binding pocket (Fig. 4.6A). There were significantly fewer internucleotide tertiary interactions in the absence of TPP ligand than in its presence; however, specific interactions in J2-4 were still observed (Fig. 4.6B). Spectral clustering revealed that both the saturating ligand and no-ligand RNAs are composite states with constituent major and minor conformations (Fig. 4.6C, D). The minor cluster in the saturating ligand RNA is characterized by increased DMS reactivities at precisely the positions that become reactive when no ligand is bound (Fig. 4.6A, C; open circles). Therefore, even under saturating ligand conditions, the TPP riboswitch RNA samples both ligand-bound and unstructured conformations. In the absence of ligand, the major cluster has a DMS reactivity pattern similar to the less structured state in the presence of ligand. In contrast, the minor cluster detected in the absence of ligand has reduced DMS reactivities in the thiamine-binding pocket, suggestive of a conformation that is more highly structured than that of the major cluster (Fig. 4.6B, D). We infer from this analysis that the riboswitch thiamine-binding pocket samples a "hidden" pre-folded structure similar to that formed upon ligand binding.

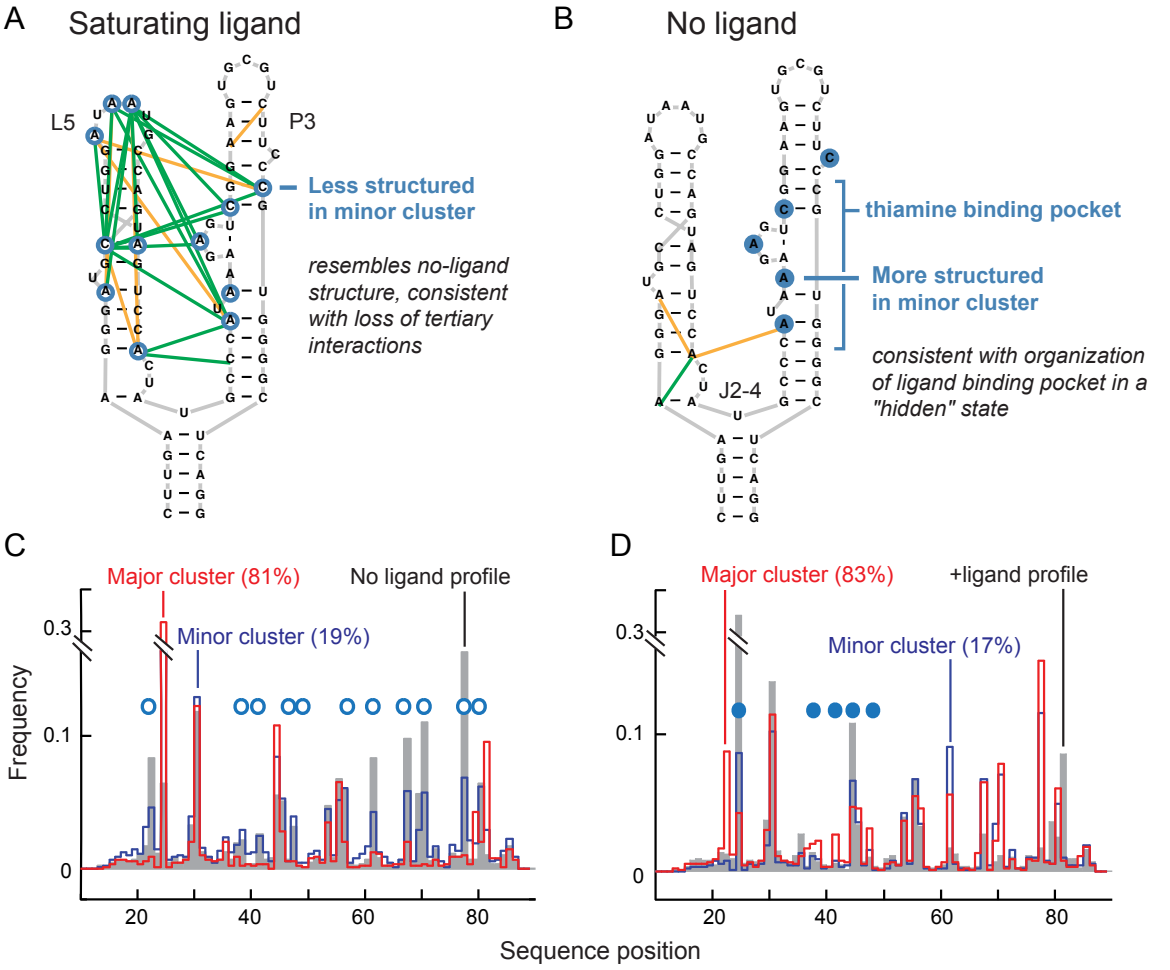


Figure 4.6 RINGs and clustering analysis of the TPP riboswitch in the presence and absence of TPP ligand. RING analysis in the presence of (**A**) saturating ligand and (**B**) absence of ligand. Strong and moderate internucleotide associations are shown with green and yellow lines, respectively. Nucleotides that are less or more structured in the minor, less populated, cluster are emphasized with open and closed spheres, respectively. Spectral clustering analysis in the (**C**) presence of saturating ligand and (**D**) absence of ligand. There are two clusters in each state. In the presence of saturating ligand, the major cluster (red) corresponds to the fully folded riboswitch. In the absence of ligand, the major cluster (red) reflects an unstructured state with few internucleotide interactions. The minor cluster (blue) in the saturating ligand sample is more unstructured than the major cluster and is similar to the no-ligand structure (gray). The minor cluster (blue) in the no-ligand sample is more highly structured than the major cluster specifically in the region of the thiamine binding pocket (blue closed circles).

We next probed the TPP riboswitch RNA in the presence of sub-saturating ligand concentrations (200 nM TTP; $K_d \sim$ 50-200 nM[14]). Clustering analysis of chemical reactivity data produced three well-defined clusters in the ratio of 1:1:1.2 (Table 4.1) corresponding to (*i*) the fully folded, ligand-bound state, (*ii*) the state in which the ligand-binding pocket is structured but the rest of the RNA shows weak internucleotide interactions, and (*iii*) the unstructured state with only a few interacting nucleotides (Fig. 4.7). Each of these clusters corresponds to states identified in either saturating ligand or no-ligand RNA. Spectral clustering analysis thus identified multiple distinct conformations from a single in-solution RNA ensemble, including a previously uncharacterized state in which the ligand-binding pocket is pre-folded. This partially folded state is likely important for recognition of the TPP ligand.

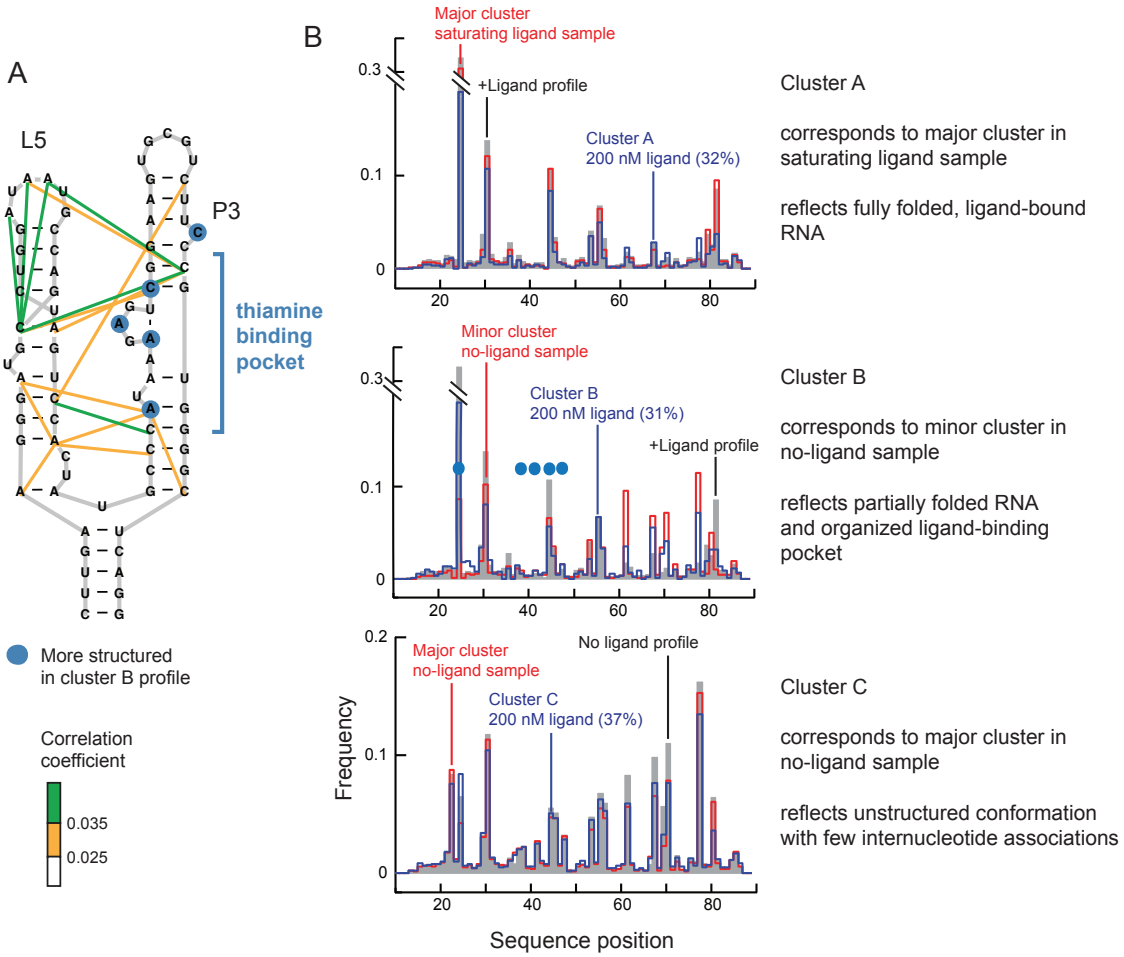


Figure 4.7 Spectral clustering analysis of the TPP riboswitch at sub-saturating ligand concentration of 200 nM ligand. (**A**) RING analysis of internucleotide association interactions. Interactions are fewer in number and weaker than those for the RNA under saturating ligand conditions (compare with Fig. 4.6A). (**B**) Three clusters were identified with population fractions of 32, 31, and 37% (blue). Each of these clusters corresponds to a state identified in either saturating ligand concentration or in the absence of ligand (red) with nucleotides corresponding to the ligand bound or no ligand structures (gray).

Finally, we examined the structure of the RNase P RNA as a function of Mg^{2+}. The networks of interactions were strikingly different in the presence and absence of Mg^{2+} (Fig. 4.8). The strong networks of interactions between L5 and L15.1 and in the structural core disappeared in the absence of Mg^{2+}+ and were replaced by interactions between P5.1 and P2 and within P7 (Fig. 4.8A and B). Spectral clustering identified two clusters in the plus- Mg^{2+} state (Table 4.1). The minor cluster in the plus- Mg^{2+} sample is distinct from both the fully folded RNA and from the no- Mg^{2+} structure (Fig. 4.8A and C). Reactive nucleotides in the minor cluster comprise the L5- L15.1 and structural core interactions, indicating that these interactions are weakened in this state (Fig. 4.8A). Critically, single-molecule spectral clustering analysis shows that RNAs natively adopt multiple unique states, even under conditions generally assumed to promote a single structure.

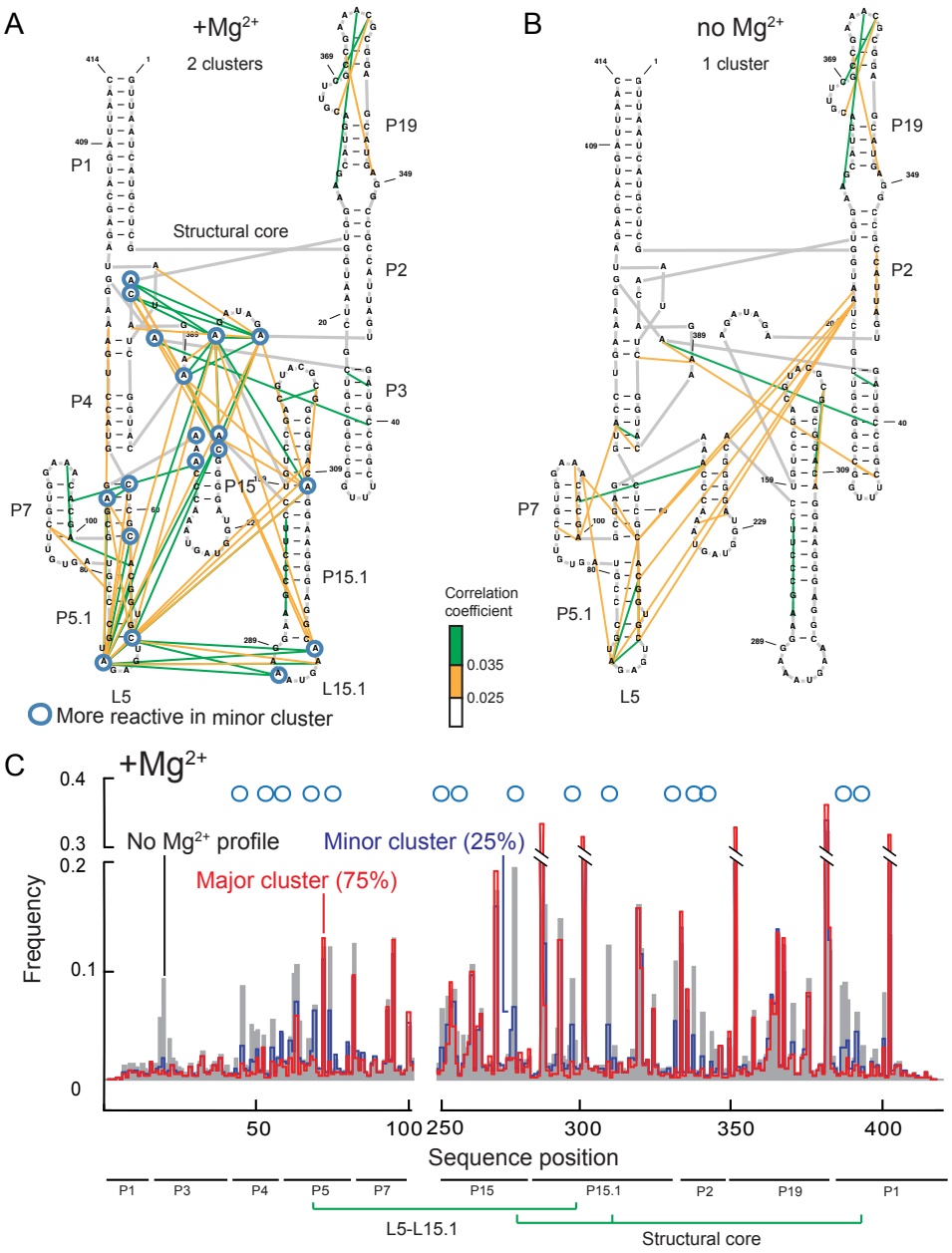


Figure 4.8 Clustering analysis of the RNase P domain RNA in the presence and absence of Mg^{2+}. (A) RING analysis of the RNA structure in the presence of Mg^{2+}. (**B**) RING analysis in the absence of Mg^{2+}. (**C**) Separation of the plus- Mg^{2+} data into two clusters. The minor cluster (blue) is characterized by a subset of nucleotides (blue circles) that are more reactive (and thus less structured) than those in the major cluster structure. Positions more reactive in the minor cluster mediate the L5-L15.1 loop-loop tertiary interaction and form the structural core. In most regions, the no- Mg^{2+} state has a RING pattern that is structurally distinct from both of the plus- Mg^{2+} states. In contrast, the P19 element shows the same RING pattern as was observed in the presence of Mg^{2+} suggesting that this region folds independently and is not stabilized by Mg^{2+}.

4.3. DISCUSSION

4.3.1. Principles of RNA folding.

RNA structure formation can be usefully approximated by assuming that stable helices, formed locally and stabilized by Watson-Crick pairing, are subsequently organized into a three-dimensional structure by longer-range tertiary interactions. RING analysis of the three RNAs studied here is consistent with this structural hierarchy. For example, we observed RINGs that reflect non-canonical base pairs and loop-helix and loop-loop tertiary interactions that have been widely noted in prior structural studies[15] (Fig. 4.9A, in magenta and red). RING analysis also identified interactions whose prevalence was previously not fully appreciated. Approximately one-third of all observed nucleotide interdependencies involve single stranded or loop nucleotides at the opposite ends of individual helices. These through-helix interactions mean that structural communication in RNA can extend over long distances. In some cases, through-helix structural coupling extends through multiple stacked helices (Fig. 4.9A, in yellow and orange).

In addition, RING analysis indicates that tertiary interactions are not independent but instead are strongly dependent on other structural elements. We observed coupled interactions between well-defined individual tertiary structure motifs in both the TPP riboswitch and the P546 domain RNA (Fig. 4.9A, in blue). Disruption of any one tertiary interaction, by exclusion of ligand or by mutation, resulted in loss of the tertiary motif itself and also disrupted other interactions. Our data also indicate the importance of close helical packing. These interactions are especially obvious in the TPP riboswitch and in the structural core of the RNase P RNA (Fig. 4.9A, in green).

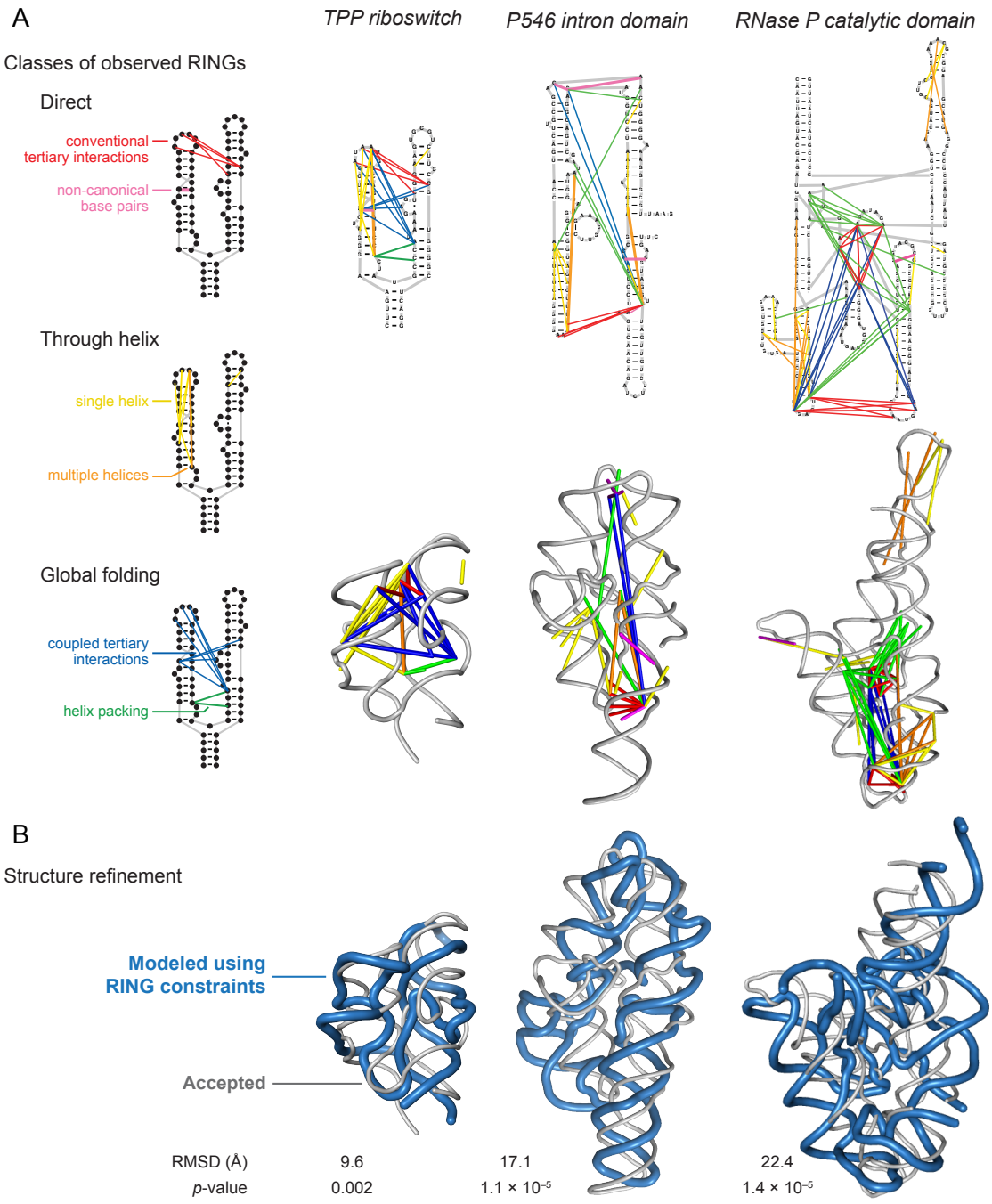


Figure 4.9 Through-space RNA structural relationships revealed by RINGs. (**A**) Direct, through-hhelix, and global internucleotide interactions are illustrated on both secondary structures (*top*) and three-dimensional models[9-11] (*bottom*). (**B**) Three-dimensional models determined for each RNA using RING interdependencies as constraints. The *p*-values report the significance of each model; the secondary structure was input during refinement[19].

Our analyses of the TPP riboswitch, the P546 domain, and the RNase P domain indicate that mutations (Fig. 4.5) or absence of ligand (Figs. 4.6 and 4.7) or divalent ions (Figs. 4.5B and 4.8) do not simply "subtract" an interaction from the structure but cause large-scale reorganization of RNA folding. None of the unfolded or less-folded states characterized were simply less structured versions of a fully folded state. Instead, we find that less structured states are stabilized by unique sets of interdependent interactions that, in general, have not been detected in prior ensemble or single-molecule studies.

4.3.2. Three-dimensional RNA structure refinement.

Because RING analysis identifies dense arrays of nucleotide interdependencies reflective of RNA tertiary structure, we explored whether these interactions could be used as restraints to model three-dimensional RNA folds. A small number of constraints, reflective of through-space RNA structure, are often sufficient to yield high-quality structure models[16,17]. We used a two-step interaction potential (Fig. 4.10A-C) to introduce free-energy bonuses when constituent nucleotides come into proximity during discrete molecular dynamics simulation[16-18]. Introduction of RING constraints caused each RNA to sample collapsed states during the simulation (Fig. 4.10D). Following filtering by radius of gyration, representative structures were selected by hierarchical clustering. For each RNA characterized, we obtained high quality and statistically significant[19] models (Fig. 4.9B) that correctly recapitulated RNA architecture defined by high-resolution structures[9-11]. For the RNase P RNA, we observed overlapping RINGs spanning two-thirds of the molecule and a second non-overlapping set in P19 (Figs. 4.4 C and 4.9A, right-hand panels); this suggests that the P3-P2-P19 element is not structurally linked to the rest of the RNA architecture. The accuracy of the three-dimensional model for the RNase P RNA is especially high in the

structural core (excluding the P3-P2-P19 element) with a 14.4 Å RMSD (p-value < 10^{-6}) compared to the crystal structure. These comparisons indicate that RING-network interactions allow both *de novo* identification of structural elements and highly accurate refinement of folded domains of large RNAs (Fig. 4.9B and 4.11).

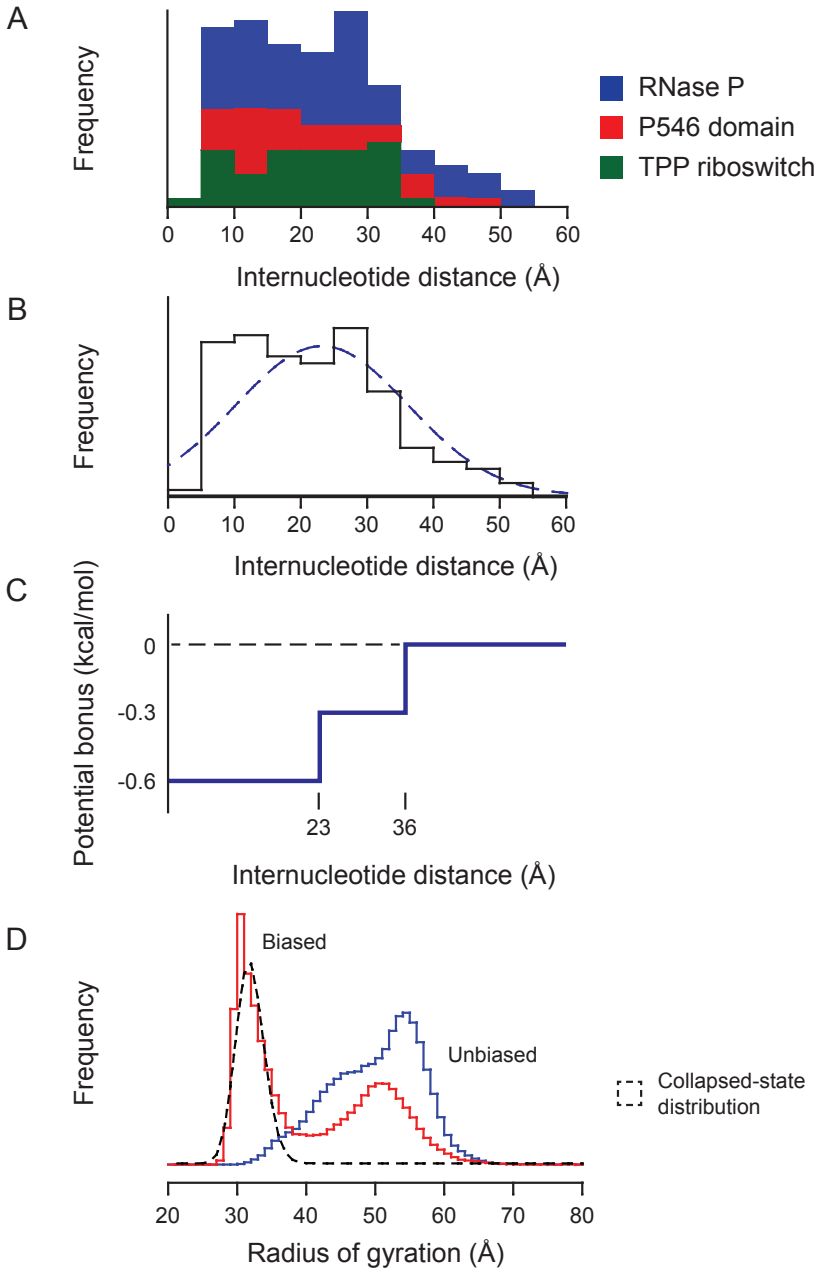


Figure 4.10 Long-range constraints for using RING interdependencies to refine three-dimensional RNA structure models. (**A**) Distribution of distances corresponding to nucleotide associations with correlation coefficients greater than 0.025. (**B**) Histogram summed over all observations. Smooth curve corresponds to the normal distribution based on the average and standard deviation. (**C**) Interaction potential for RING-based distance constraints. (**D**) Radius of gyration based filtering of structure models. Representative histograms of radii of gyration for models of the P546 domain RNA generated during unbiased simulations (blue) or simulations biased by RING data (red). The fit log-normal distribution for the bias-dependent collapsed state is shown with a dashed line.

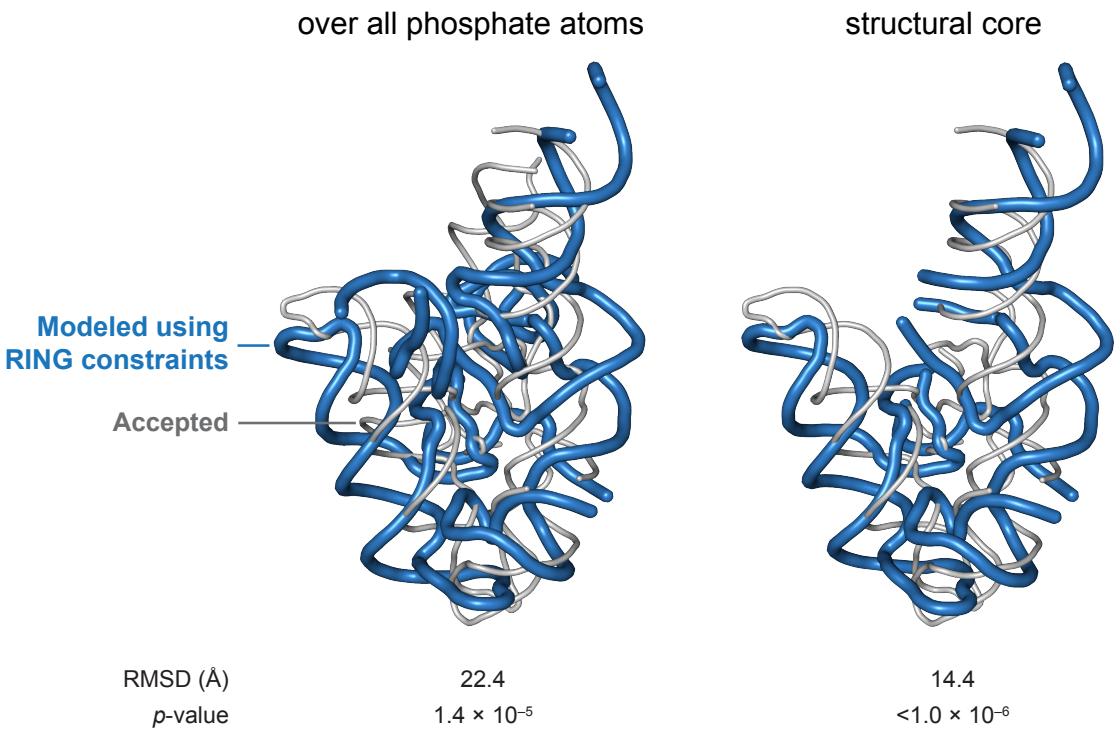


Figure 4.11 Comparison of the RING-directed three-dimensional model for the RNase P RNA with that based on the crystallographically visualized structure[11]. The core accepted structure (right) excludes helices P3-P2-P19 which folds independently (see Fig. 4.4). The p-values report the significance[19] of each model relative to the accepted structure.

4.3.3. Perspective.

Single-molecule structure analysis by sequencing represents a remarkably simple and generic approach for analysis of the global architectures of functionally important RNAs or DNAs. The basic insight of recording multiple events in the same RNA or DNA strand is completely general and should inspire development of numerous new classes of experiments and biological discoveries. The approaches developed here can be readily extended to other RNA modification agents, to SHAPE experiments that interrogate all four nucleotides simultaneously[20], and to RNA-protein crosslinking[21], providing that modifications are detected with high frequency. Single molecule experiments that explore protein-RNA, RNA-RNA and DNAmediated interactions simultaneously are clearly feasible. Higher-order structure is tightly linked to biological function. Identifying RINGs in large RNAs and transcriptomes will enable widespread biological functional motif discovery.

4.4. METHODS

4.4.1. Characterization of reaction between dimethyl sulfate and RNA nucleobases.

Adduct formation between DMS (Sigma-Aldrich) and [γ-^{32}P]-labeled ATP, CTP, and UTP was performed by adding 10% (vol/vol) DMS (1 μL; 1.7 M, in absolute ethanol) to [γ-^{32}P]-NTP in 1x reaction buffer [9 μL, 10 mM $MgCl_2$, 300 mM sodium cacodylate (pH 7.0)] at 37 °C. Reactions were quenched with an equal volume of neat 2-mercaptoethanol after 10, 30, 60, 120,180, 360, and 900 s. For pre-quench control reactions, 1.3 M DMS solution [1:2:5 (vol/vol) DMS:ethanol:H_2O] was first added to equal volume neat 2-mercaptoethanol. This mixture (2.8 μL, 625 mM DMS) was then immediately added to [γ-^{32}P]-NTP in 7.2 μL of 1.4x reaction buffer [7.2 μL, 14 mM $MgCl_2$, 417 mM sodium cacodylate (pH 7.0)], and the reaction was incubated at 37 °C for 15 min. Quenched reactions were resolved by gel

electrophoresis (30% polyacrylamide; 29:1 acrylamide:bisacrylamide; 0.4 mm x 28.5 cm x 23 cm gel; 30 W, 45 min) and quantified by phosphorimaging. Data were consistent with a mechanism in which DMS forms adducts at the N1 position of adenosine and N3 position of cytosine and does not react with uridine. The change in pH during DMS adduct formation was followed in reactions without NTP at 37 °C using an Accument 25 pH meter. Direct measurements of DMS adduct formation with cytosine and adenosine suggest roughly equal reactivities with these two nucleotides (Fig. 4.2). This observation differs from the widespread view that adenosine reacts more rapidly than does cytosine with DMS. We attribute this misconception to the relatively inefficient ability of N3-methyl cytosine to cause pausing by reverse transcriptase enzymes.

4.4.2. RNA constructs.

DNA templates for the *E. coli* TPP riboswitch, *Tetrahymena* group I intron P546 domain, and *B. stearothermophilus* RNase P catalytic domain RNAs, each imbedded within 5' and 3' structure cassette flanking sequences, were generated by PCR[22]. RNAs were transcribed *in vitro* [1 mL; 40 mM Tris (pH 8.0), 10 mM $MgCl_2$, 10 mM dithiothreitol, 2 mM spermidine, 0.01% (v/v) Triton X-100, 4% (w/v) poly(ethylene) glycol 8000, 2 mM each NTP, 50 μL PCR-generated template, 0.1 mg/mL T7 RNA polymerase; 37 °C; 4 h] and purified by denaturing polyacrylamide gel electrophoresis [8% polyacrylamide, 7 M urea, 29:1 acrylamide:bisacrylamide, 0.4 mm x 28.5 cm x 23 cm gel; 32 W, 1.5 h]. RNAs were excised from the gel, recovered by passive elution overnight at 4 °C, and precipitated with ethanol. The purified RNAs were resuspended in 50 μL TE and stored at -20 °C.

4.4.3. RNA folding and DMS modification.

RNA structure probing experiments were performed in 10 mM $MgCl_2$ and 300 mM cacodylate at pH 7.0. RNAs [5 pmol in 5 μL 5 mM Tris (pH 7.5), 0.5 mM EDTA (1/2Å~ TE)] were denatured at 95 °C for 2 min, cooled on ice, treated with 4 μL 2.5x~ folding buffer [750 mM cacodylate (pH 7.0), 25 mM $MgCl_2$], and incubated at 37 °C for 30 min. After folding, the P546 domain and the RNase P catalytic domain RNAs were treated with DMS (1 μL, 1.7 M in absolute ethanol) and allowed to react at 37 °C for 6 min. No-reagent control reactions were performed with 1μL absolute ethanol. Reactions were quenched by the addition of an equal volume of neat 2-mercaptoethanol, and immediately placed on ice. No-Mg^{2+} experiments were performed identically except that the 2.5Å~ folding buffer omitted Mg^{2+}. The TPP riboswitch RNA was incubated in folding buffer at 37 °C for 10 min, after which the TPP ligand was added at the desired concentration, and samples were incubated at 37 °C for 20 min. (Note: DMS is a known carcinogen and neat 2ME has a very strong odor. Manipulations involving DMS and 2ME should be performed in a chemical fume hood. Solutions containing DMS should be neutralized with 5N NaOH. Solutions containing DMS or 2ME should be disposed of as chemical waste.)

4.4.4. Reverse transcription and adduct detection.

The reverse transcription and adduct detection method was adapted from the SHAPE-MaP protocol[5]. After treatment with DMS, RNAs were purified using G-50 spin columns (GE Healthcare). Reverse transcription reactions were performed using SuperScript II reverse transcriptase (Invitrogen) for 3 hrs at 42 °C [0.5 mM premixed dNTPs, 50 mM Tris HCl (pH 8.0), 75 mM KCl, 6 mM $MnCl_2$, and 10 mM DTT]. Reactions were desalted with G-50 spin columns (GE Healthcare). Under these conditions (long incubation time, in the

presence of Mn^{2+}), the reverse transcriptase reads through methyl adducts at the N1 and N3 positions of adenosine and cytosine, respectively, yielding a mutation at the site of the adduct. Double-stranded DNA libraries with adaptors and indices compatible with Illumina-based sequencing were generated by PCR. Resulting libraries were pooled and sequenced with an Illumina MiSeq instrument (500 cycle kit) so that the first sequencing read in paired end mode covered the RNA sequence of interest. Resulting FASTQ data files were aligned to reference sequences and per-nucleotide mutation rates were calculated using an inhouse pipeline[5]. Phred scores used to count mutations were required to be ≥ 20.

4.4.5. Measurement of inter-nucleotide interactions by statistical association analysis.

To detect nucleotide reactivity interdependencies, all possible pairs of nucleotides were subjected to Yates' corrected version of Pearson's chi-squared test of independence versus association[23]. Yates' corrected chi-squared statistic was computed as:

$$\chi^2_{Yates} = \frac{N\left(|ad - bc| - 0.5N\right)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (1)$$

where $N = (a + b + c + d)$ is the total number of strands in the dataset, and *a, b, c*, and *d* are defined by the following 2 Å~ 2 contingency table of the observed numbers of four possible cooccurrences:

Nucleotide *i* / Nucleotide *j*	Not mutated 0	Mutated 1
Not mutated 0	*a*	*b*
Mutated 1	*c*	*d*

A pair of nucleotides was taken to have a statistically significant association if $\chi_{Yates}^2 > 20$ ($p < 0.00001$). With this high acceptance threshold for an individual nucleotide pair, we expect to make no more than one false positive determination for RNAs at least up to 500 nucleotides long. For pairs of nucleotides that passed the chi-squared significance test, the sign and strength of the statistical association was determined by computing Pearson's correlation coefficient, ρ. In the case of two binary variables, ρ is equal to Pearson's measure of association, the *phi coefficient*. The correlation coefficient and the chi-squared statistic are related:

$$\rho^2 = \chi^2 / N \tag{2}$$

Heat maps of nucleotides with statistically significant associations are illustrated in Fig. 4.12. Although correlation coefficients were typically less than 0.05, coefficients were highly significant. Based on χ^2 statistics, the probability that identified correlated nucleotides were independent was less than 0.00001.

The following guidelines were also imposed for nucleotide association analysis and clustering. The average number of modifications detected per read was required to be approximately 15% of the estimated number of single-stranded nucleotides, yielding an average equal to or greater than two mutations per read. Nucleotides with a mutation rate greater than 0.05 in the no-modification control were excluded from the chi-squared calculation. Correlated nucleotide pairs with a standard deviation of their correlation coefficient (estimated by bootstrapping) greater than 20% were not used as RING constraints. Bootstrapping iterations were sufficiently large so that the absolute difference between bootstrapped and calculated correlation coefficients was less than 1% of the calculated correlation coefficient.

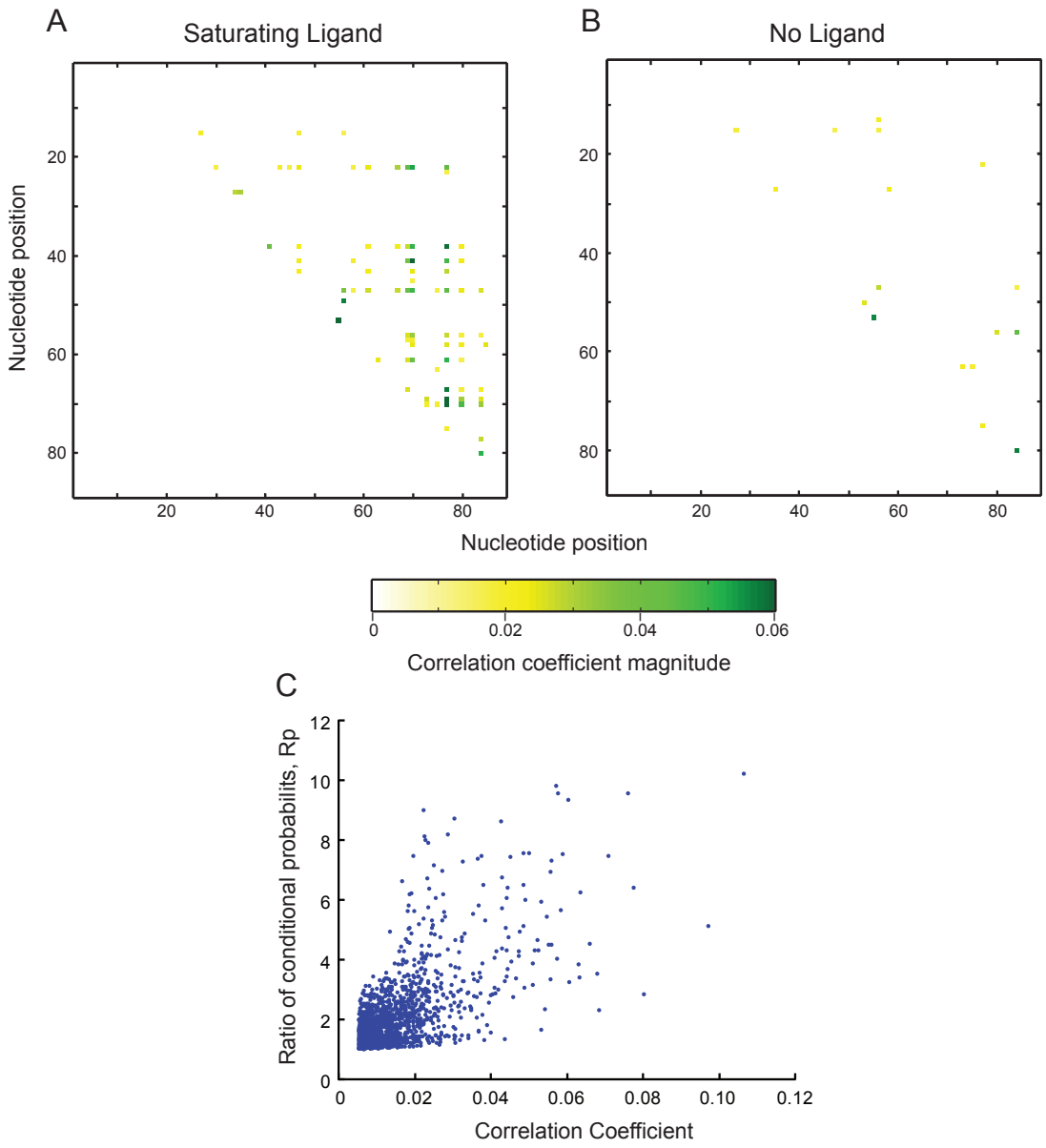


Figure 4.12 Visualization of internucleotide associations in the presence and absence of ligand for the TPP riboswitch RNA. (**A, B**) Heat map showing positive statistically significant ($\chi^2 > 20$) nucleotide associations for the TPP riboswitch in (**A**) the presence and the (**B**) absence of the ligand. (**C**) Illustration of relationship between correlation coefficient and the effect that mutation of one nucleotide has on the probability of mutation of the other nucleotide in the pair. Such an effect is measured as a ratio of conditional probabilities [$R_p = P_{(A = 1 \mid B = 1)} / P_{(A = 1 \mid B = 0)}$], and it is plotted for statistically significant ($\chi^2 > 20$) positive associations as a function of correlation coefficient. In the RING analysis, we focused on nucleotide pairs with correlation coefficients greater than 0.025, corresponding to a median ratio of conditional mutation probabilities, R_p, greater than 2.50.

4.4.6. Spectral clustering of multiple conformations in a single RNA ensemble.

Nucleotides making up an RNA define the dimensions of an abstract high-dimensional space, in which any single read of an RNA strand is represented by a point whose coordinates are defined by which of the nucleotides, if any, reacted with the chemical reagent (each coordinate is set to either 1 or 0, depending on whether the nucleotide that the coordinate represents was reactive or not). In this space, strands with similar structural conformations will tend to cluster, reflecting differences in frequency of modification profiles for each conformation. We used spectral clustering[24-26], which is particularly effective in finding arbitrarily shaped clusters, to define RNA structural clusters, without making any assumptions about the form of the data clusters.

To detect the presence of multiple structural conformations in an RNA pool using spectral clustering, we summarized the locations in the primary sequence (*N* nucleotides in length) of the nucleotides that were modified by the chemical reagent in *M* RNA strands, in a "hit" matrix, $H_{M \times N}$. This dataset was treated as a simple, complete, undirected, weighted graph, in which each nucleotide is represented by a vertex with all of the *N* vertices linked by edges. Each edge was assigned a weight corresponding to the similarity of the reactivity patterns of the two nucleotides, measured as a number of reads in the dataset in which both nucleotides were modified:

$$S = H^T H \tag{3}$$

This similarity matrix *S* was then used to construct normalized graph Laplacian matrix:

$$L_{NCut} = D^{-1/2} \cdot (D - S) \cdot D^{-1/2} \tag{4}$$

where D is a diagonal matrix, in which $D_{ii} = \sum_j S_{ij}$. The eigenvectors of matrix L_{NCut} were used to perform a normalized cut partitioning of the dataset into clusters by cutting the edges between vertices in a way that minimized the sums of the weights of the cut edges and maximized the sum of weights of the preserved edges[24-26]. In our application to RNA mutation data, the eigenvalues and eigenvectors of the normalized graph Laplacian matrix L_{NCut} were used to 1) determine how many structural conformations are present in the studied RNA pool, 2) estimate relative fractions of different conformations in the sample, and 3) reconstruct mutation frequency profiles for the individual conformations. These procedures are described in detail in the following paragraphs. Spectral clustering was applied to adenosine and cytosine nucleotides that were modified with frequencies greater than 0.01. Strands with no modifications carry no information and were excluded from spectral clustering.

The eigenvalues were sorted in ascending order, from the smallest eigenvalue, λ_1, to the largest, λ_N. The first eigenvalue, λ_1, is always zero. Eigenvalues express the effectiveness of each normalized cut partitioning of the vertices. The more effective a particular cut is in cutting edges between dissimilar vertices (such as those belonging to different clusters) while preserving edges between similar vertices (such as those belonging to the same cluster), the smaller its eigenvalue. Therefore, if a dataset has K distinct clusters, the first K eigenvalues will be distinctly smaller than the K+1 eigenvalue and the rest of the eigenvalues. Thus, to estimate the number of clusters, K, in a dataset, we chose K such that all eigenvalues λ_2, λ_3 … λ_K were relatively small, and λ_{K+1} was relatively large. To make jumps between consecutive eigenvalues more apparent, "eigengaps" (defined as a difference $\Delta\lambda_i = \lambda_{i+1} - \lambda_i$; with the first eigengap, $\Delta\lambda_1$, set to zero) rather than eigenvalues were evaluated (Fig. 4.13)[26]. In general, if

a dataset has K clusters, the eigengap plot will have an outstanding eigengap in the K position ($\Delta\lambda_K$) and also likely to the left of it, but not to the right of it (Fig. 4.14).

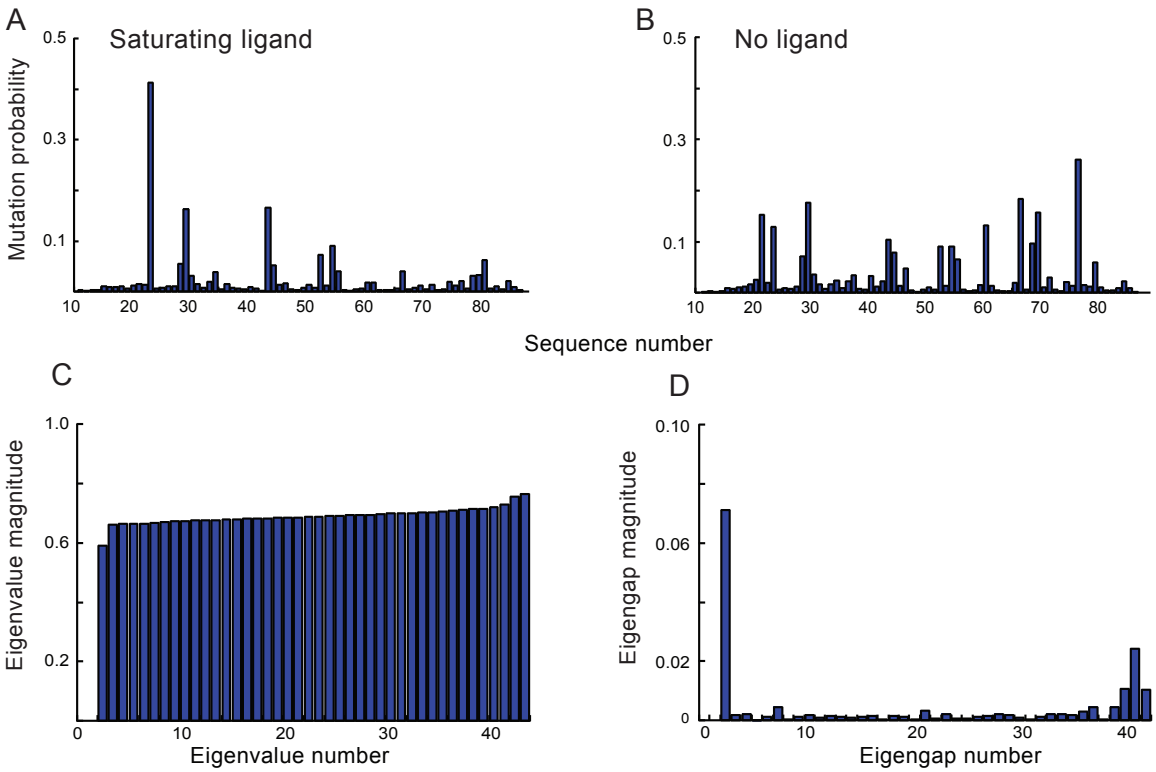


Figure 4.13 Spectral clustering of RING-MaP data obtained from analysis of the TPP riboswitch RNA. (**A, B**) A synthetic data sample of a known conformational composition was created by combining two sets of sequencing reads obtained after modification of the TPP riboswitch RNA: one set (**A**) of 50,000 reads was obtained in the presence of saturating TPP ligand, and the other set (**B**), also of 50,000 reads, was obtained in the absence of TPP ligand. Note that these two sets have distinctly different modification frequency profiles, reflecting their different RNA conformations. Any conformational variations that happened to be present within the either set were deliberately destroyed by randomly shuffling the recorded instances of nucleotide modifications among the reads. Such shuffling, performed independently for every nucleotide in each set, preserved the total number of modifications to a given nucleotide in a given set, but made co-occurrences of modifications among nucleotides statistically random. Thus, this data sample has only two conformations, present at 50:50 ratio. This synthetic data sample is representative of data collected in single experiments, since nucleotide modifications occur on each strand independently of other strands in the RNA pool. (**C**) Eigenvalues of the normalized graph Laplacian matrix, L_{NCut}, of the similarity matrix, S, constructed for the 43 reactive adenosine and cytosine nucleotides of the synthetic TPP RNA data sample. The first eigenvalue, λ_1, is always zero, whereas the magnitudes of successive eigenvalues reflect inversely the effectiveness of each successive normalized graph cut. If the data points form two distinct clusters, the first graph cut will be most effective, cutting the links between the points lying in different clusters, thus cleanly separating the clusters. Indeed, in the plot the largest difference between successive

eigenvalues is between λ_2 and λ_3, indicating that the first cut of the data graph (at λ_2) was exceptionally effective. (**D**) Eigengap plot of the differences in magnitude between successive eigenvalues, which make cluster-indicating jumps between successive eigenvalues more apparent. There is just one outstanding eigengap in this plot, $\Delta\lambda_2$, between eigenvalues 2 and 3, indicating that the nucleotides form two prominent clusters reflecting two conformations.

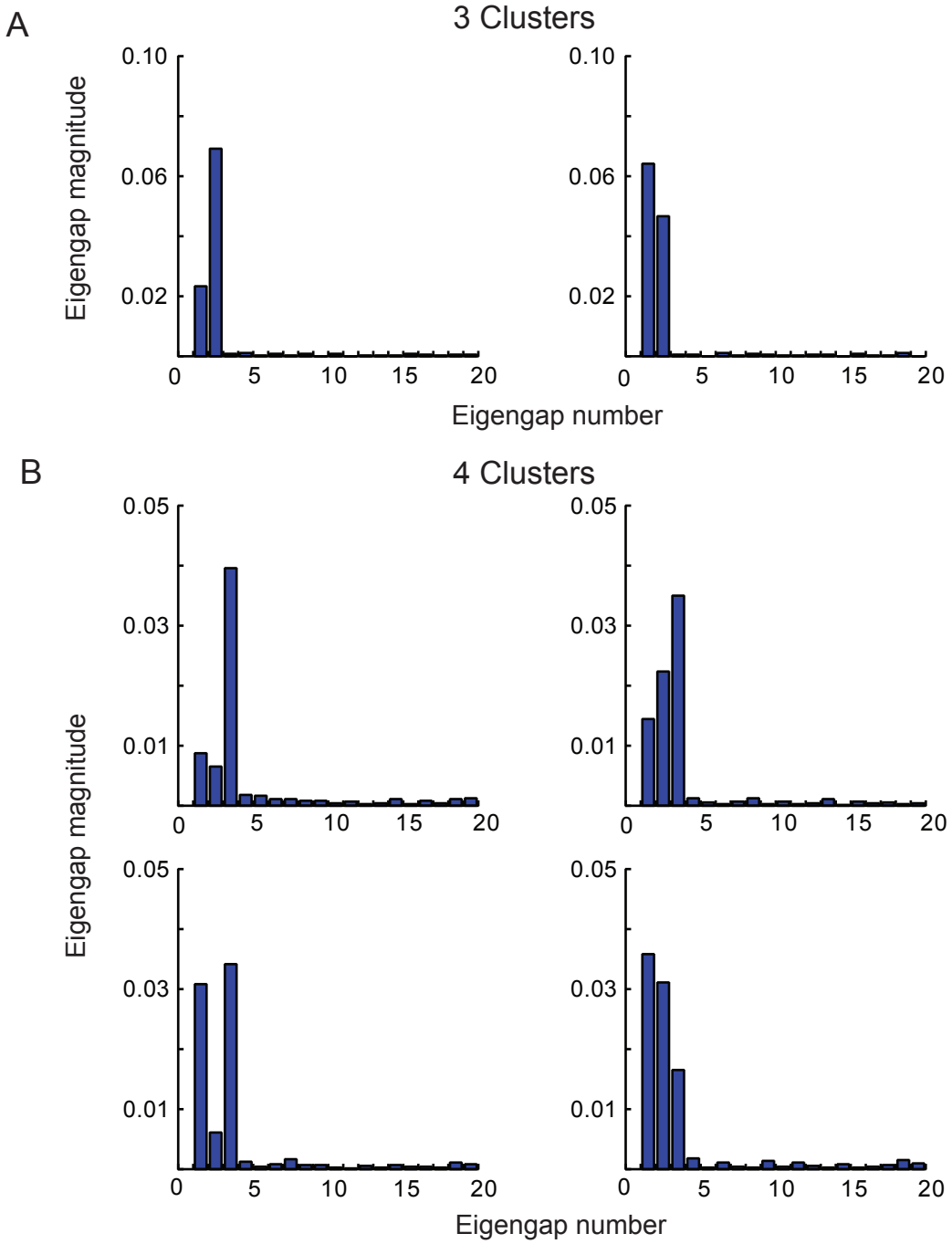


Figure 4.14 Examples of eigengap plots containing either three (top row) or four (two bottom rows) clusters in varying proportions. These plots were generated using artificial datasets, derived from a hypothetical RNA containing 50 adenosine and cytosine nucleotides with varied frequencies of modification comparable to those observed in real RNAs. For each cluster of reads, nucleotides were assigned their modification probabilities at random.

4.4.7. Estimating relative fractions of different conformations in RNA sample.

If a dataset has *K* clusters, eigenvectors $\vec{x}_2 \ldots \vec{x}_K$ can be used to assign individual RNA strands to clusters. Specifically, if a dataset of *M* strands is recognized to have *K* clusters, the strand scores are

computed as:

$$Y_{M\times(K-1)} = H \cdot \left[\vec{x}_2, \vec{x}_3, \cdots \vec{x}_K\right] \tag{5}$$

where $\vec{x}_2, \vec{x}_3 \ldots \vec{x}_K$ are the second to the *K*-th eigenvectors. The scores have *K*-1 dimensions, and the strands are partitioned into *K* clusters by performing *K*-means clustering of *M* data points in the *K*-1 dimensional score space (Fig. 4.15).

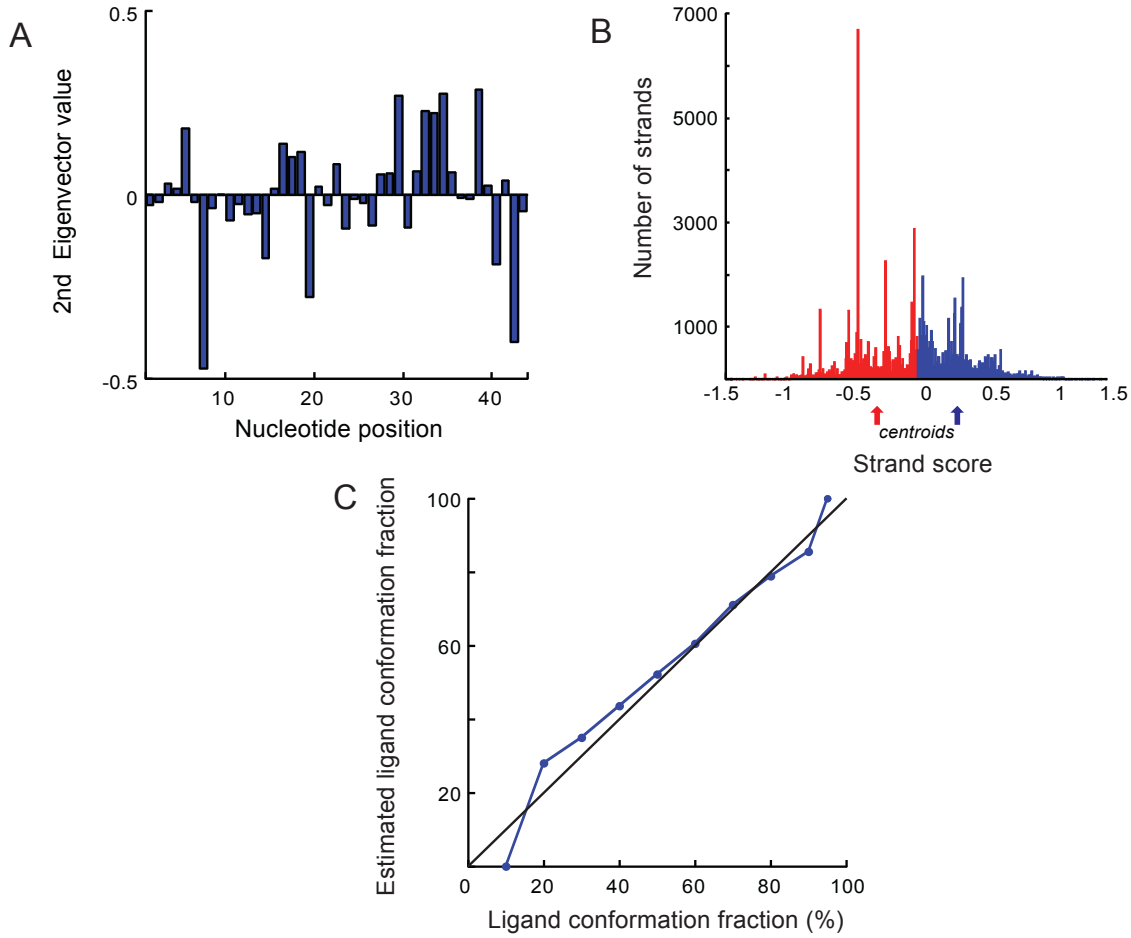


Figure 4.15 Determination of the conformational identity of individual RNA strands. (**A**) The second eigenvector, $\vec{x}_2$, of the synthetic data set created for the TPP riboswitch as described in Fig. 4.13, was chosen because the second eigengap indicates that these data are split into two clusters (see Fig. 4.13D). This eigenvector has 43 values, corresponding to 43 adenosine and cytosine nucleotides with high modification rates. For each nucleotide, the eigenvector magnitude specifies how strongly that nucleotide is associated with either of the two detected clusters. (**B**) Distribution of the scores for all strands computed from the second eigenvector. The strands that belong to one cluster are on the left side of the score distribution, whereas the strands that belong to the second cluster are on the right side of the distribution. The classification boundary was determined using *K*-means clustering; red and blue arrows indicate centroids of each cluster. The ratio of strands in the two clusters is 52:48, which is in good agreement with the true ratio of 50:50 for synthetic dataset. (**C**) Estimated relative fractions of the ligand-bound and unstructured conformations plotted as a function of their true relative fractions. The plot was generated by performing spectral and *K*-means clustering on eight different synthetic datasets with different proportions of strands belonging to the two TPP RNA conformations.

4.4.8. Reconstructing modification frequency profiles of individual conformations in an RNA sample.

Once strands are assigned to clusters reflective of distinct conformations, the modification frequency profiles can be computed specifically for each conformation. The accuracy of such profile reconstruction was improved by computing the modification frequencies of each nucleotide separately using the following procedure. To compute the modification frequencies of nucleotide *i*, this nucleotide was first removed from the hit matrix *H* and then spectral clustering and strand partitioning by *K*-means clustering was performed on this reduced matrix (without the contribution of this nucleotide). Next, the modification frequency of nucleotide *i* was computed separately for each partitioned group of strands, yielding estimates for different conformations (Fig. 4.16).

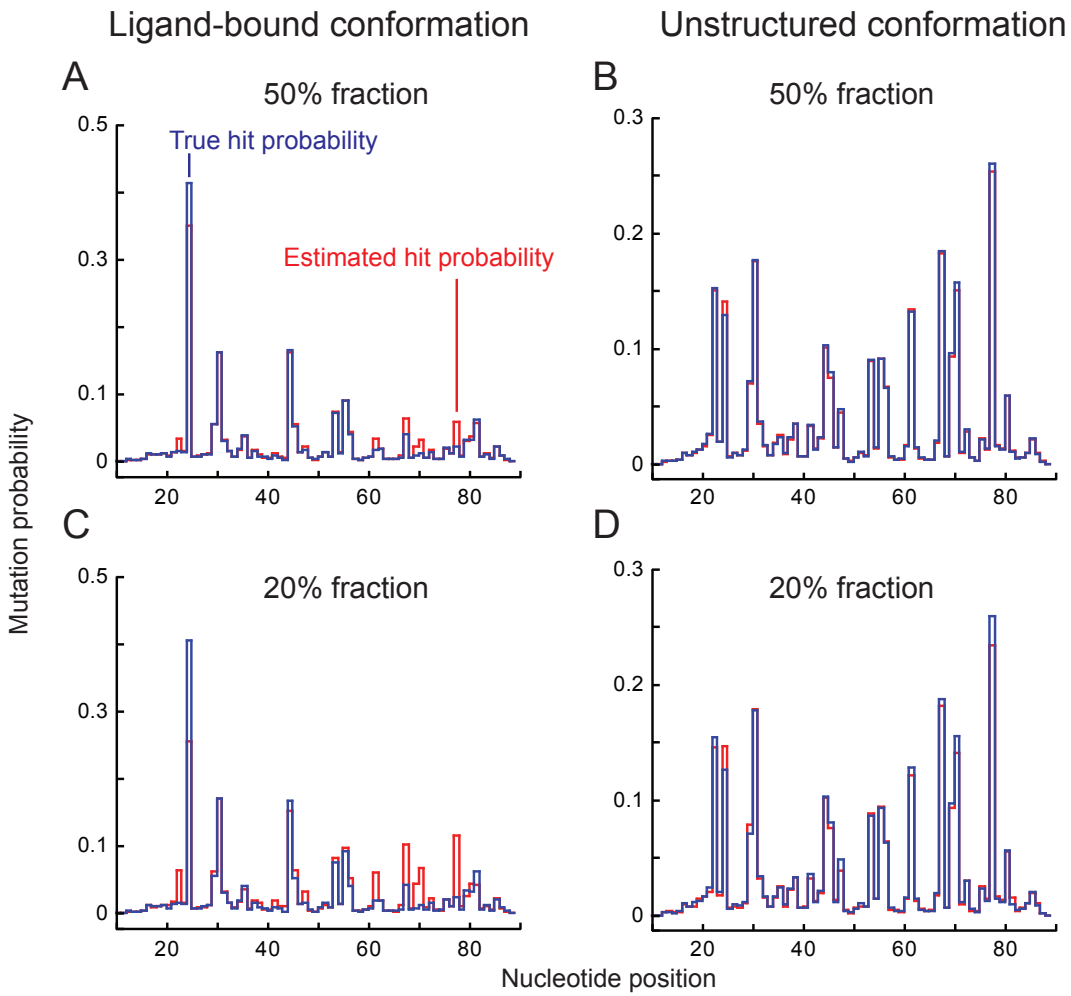


Figure 4.16 Reconstruction of mutation probability profiles of individual conformations for data samples containing two conformations each. The two conformations were represented by sequencing reads obtained after modification of the TPP riboswitch RNA in the presence of saturating TPP ligand (ligand-bound conformation) or in the absence of TPP ligand (unstructured conformation), respectively. Any conformational variations within the either set were intentionally destroyed, by randomly shuffling the recorded instances of nucleotide modifications among the reads (see Fig. 4.13, legend). To make data samples, the ligand-bound and unconstrained sets of reads were combined at 50:50, 20:80 or 80:20 ratios. (**A**) The true mutation probability profile (blue) and the profile estimated from spectral and *K*-means clustering (red) for the ligand-bound conformation making up 50% of the data sample. (**B**) The true and estimated mutation probability profiles for the unstructured conformation making up 50% of the data sample. (**C**) The true and estimated mutation probability profiles for the ligandbound conformation making up 20% of the data sample. (**D**) The true and estimated mutation probability profiles for the unconstrained conformation making up 20% of the data sample. The 80% fractions are not shown because the true and estimated profiles are nearly identical (as expected).

4.4.9. Three-dimensional RNA structure modeling.

Three-dimensional RNA fold reconstruction was performed using a restrained molecular dynamics approach in which free energy bonuses were incorporated based on pair-wise nucleotide correlations[16, 17]. Each nucleotide was modeled as three pseudo-atoms corresponding to the phosphate, sugar, and base groups. Pair-wise interactions including base pairing, base stacking, packing interactions, and electrostatic repulsion are approximated using square-well potentials[18]. Accepted base-pairing arrangements[9-11] were used to constrain modeling.

To incorporate information from RING analysis, free energy potentials were applied during the discrete molecular dynamics (DMD) simulations between nucleotide pairs found to interact. A free energy bonus was included for interacting nucleotide pairs if the absolute correlation coefficient was above 0.025. Free energy potentials were not included if two nucleotides were implicated as being in contact by proximity in primary sequence or by participation in a common secondary structure element. For primary sequence neighbors, a free energy potential was not included between nucleotides within 11 positions in sequence. For secondary structure neighbors, potentials were not added for nucleotides in the same structural element. Structural elements were defined as nucleotides at positions n_i and n_j in a RING pair and nucleotides at positions m_i and m_j in any given base pair, if $|n_i - m_j| + |n_j - m_j|$ is less than or equal to 11 nucleotides. The 11-nt threshold was selected based on the number of base pairs in a single turn of an A-form RNA helix.

Free energy potentials were imposed between RING-correlated nucleotide pairs based on through-space distances between constituent nucleotides during simulations. For distances between correlated nucleotides within 36 Å and 23 Å, the applied bonuses were -0.3 and -0.6

kcal/mol, respectively (Fig. 4.10). The maximum -0.6 kcal/mol bonus is equivalent to stabilization afforded by a single RNA stacking interaction in the molecular dynamics force field.

Molecular dynamics simulations were performed using the DMD engine with replica exchange[27]. Eight replicas were run in parallel for 1,000,000 time units each with replica temperature factor values of 0.1000, 0.1375, 0.1750, 0.2125, 0.2500, 0.2875, 0.3250, and 0.3625. From each replica, models at every 100 time units were taken forward. This list of models was then filtered based on radius of gyration. Radii of gyration for these models were compared against a control simulation where no RING-based potentials were incorporated. For both RING-dependent and control models, radius of gyration histograms were constructed. The control histogram was scaled to minimize its difference from the experimental histogram, and the frequency of the control histogram was then subtracted from the experimental. For this difference histogram, a log-normal distribution was found by least-squares fitting that describes the distribution of radii of gyration for constraint-dependent collapsed structures (Fig. 4.10D). To be considered further, RING-dependent models had to be within a geometric standard deviation of the geometric mean described by this fit distribution.

Following filtering by radius of gyration, the 250 models with the lowest energies were then analyzed by hierarchical clustering[17]. Clustering was performed considering RMSD values between analyzed models. Clustering was constrained such that maximum RMSD between any two constituent members of a cluster was less than the sum of the average and standard deviation of the RMSD distribution predicted for the analyzed structure. The medoid of the most populated cluster was taken as the predicted structure.

4.5. GUIDELINES FOR OF RING-MAP DATA ANALYSIS

4.5.1. Qualifying data used in RING-MaP calculations

RING-Map analysis can be done using a single RNA sample, without the requirement of any background correction. It is important to determine the quality of the samples used in order to correctly interpret the resulting RING and clustering data. The quality of the samples used can be evaluated based on the criteria of read depth, the number of mutations per sequence, and the level of modifications in the sample. In addition a few control experiments are suggested to aid in the evaluation of RING-MaP data. The recommended control experiments include a no-modification and a no-Mg^{2+} sample.

4.5.1.1. Number of mutations per sequence.

RING and clustering analysis of RNA structure is possible because multiple structurally-dependent adducts can be detected in a single sequencing read. As a result, the number of mutations per read affects the quality of the calculations. When evaluating the number of mutations detected in a single read it is important to consider the size of the RNA. Smaller RNAs have fewer single-stranded nucleotides available for modification and so the number of modifications per sequence should scale to the size of the RNA. Figure 4.4A shows the distribution of the number of adducts detected for each RNA studied and can serve as guide for other RNAs. For an RNA with an accepted secondary structure, the average number of mutations detected per read should be approximately 15% of single-stranded nucleotides. For the no-modification samples, the average number of mutations per read should not be more than one mutation per read. A higher average mutation rate per read in the control sample indicates mutation rates at certain positions are caused by factors other

than adduct formation. Examining the mutation frequency of each sample can identify these positions.

4.5.1.2. Mutation frequency of samples.

An important aspect of the analysis described in this paper is that RING and clustering calculations can be carried out on a single sample with out the need for a background subtraction. These calculations can only be done if the mutation frequency at single stranded A's and C's is sufficiently high. To determine if the mutation frequency of a modified sample is sufficiently high, a no-modification control can be used. First, the mutation frequency of the modified sample should be much greater than unmodified sample. Single-stranded nucleotides in the modified sample should have a mutation rate approximately 5x or greater than those in the unmodified samples. Figure 4.4B can serve as an example to evaluate the mutation frequency.

Second, the no-modification control can also be used to identify high mutation positions that are not caused by modification. Positions that have high mutation rates in the no-modification control can result in a large number of correlations that do not reflect structural interactions or produce clusters that do not reflect native structures. If the mutation frequency at a nucleotide in the no modification control is greater than 0.05 it should be removed from chi square and clustering calculations.

Finally, comparison of modified and no-modified sequencing data can identify an increase in the mutation rate toward the 5' end of the RNA. A gradual increase in mutation rate from the 3' to the 5' end of the RNA is a result of decreased sequencing accuracy by the MiSeq instrument. The decrease in sequencing accuracy has a greater effect on larger RNA and can be seen in the RNase P RNA (Fig. 4.17A). Data with high mutation rates at the 5'

end can greatly affect the resulting RINGs, producing a large number of correlations between positions at the 5' end of the RNA that do not correspond to structural interactions (Fig. 4.17B). High mutation rates due to inaccurate sequencing can be corrected for by adjusting the phred score threshold used for counting mutations. The standard phred score outlined for SHAPE-MaP analysis is ten.[6] Increasing the phred score threshold to twenty will compensate for the increased mutation frequency due to instrument error. The increased phred score does not affect high quality base calls, so the average mutation rate at the 5' end of the RNA will decrease and allow for accurate calculation of RINGs (Fig. 4.17 C,D). One consideration for this analysis is that increasing the phred score above twenty can eliminate mutations due to adduct formation and miss key structural correlations. In this study all samples were analyzed using a phred score of twenty, which did not affect correlations for the shorter TPP riboswitch and P546 domain.

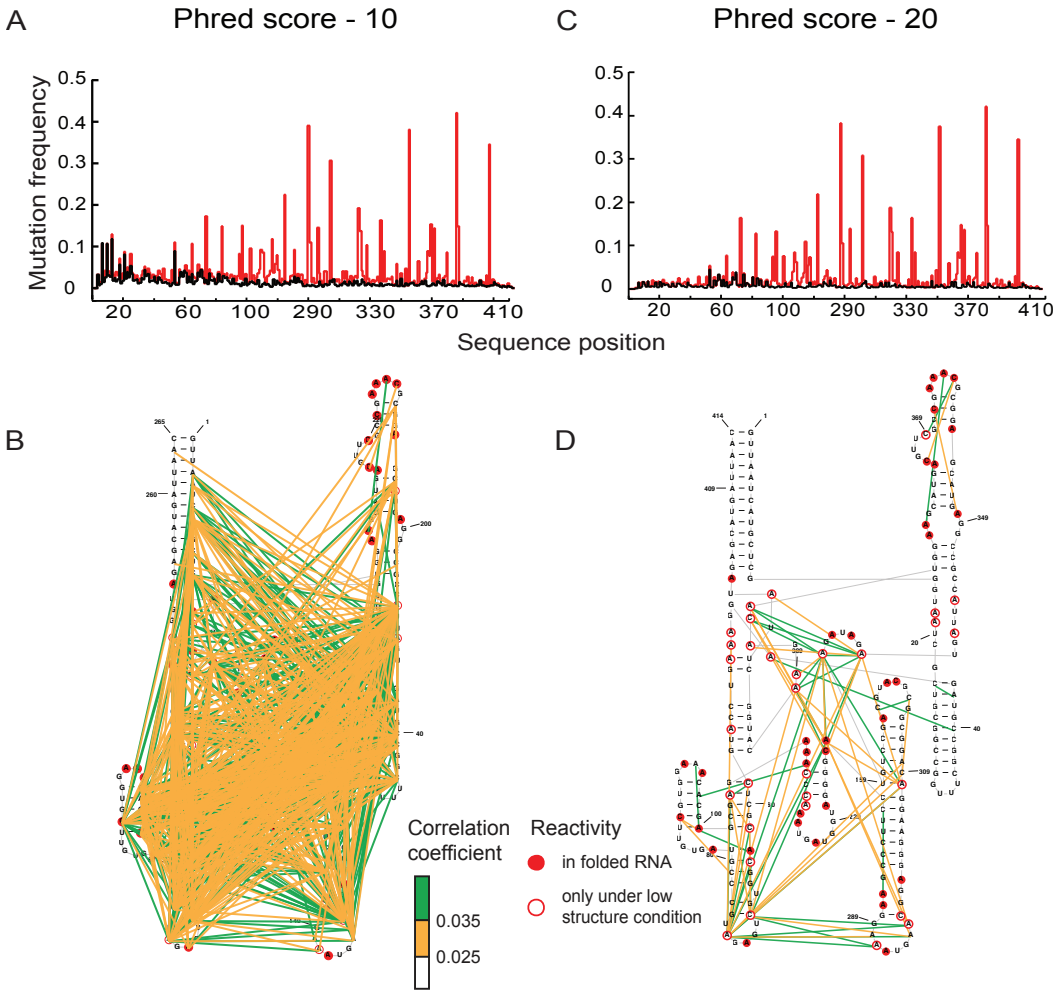


Figure 4.17 Increasing phred score for counting mutations eliminates base call errors from sequencer. (**A**) Reactivity plot of DMS modified RNA (Red) and no modification control (Black) with the phred score set at 10 produces increased average mutation rate at 5' end. (**B**) RING of DMS modified RNA processed with phred score of 20. High mutation rate due to insturment error causes spurious correlations unrealted to RNA structure. (**C**) Increasing phred score to 20 for counding mutations produces reactivity profile with decreased average mutation rate at 5' end of trace. (**D**) RING network of same DMS modified sample with increased phred score eliminates spurious correlations. RINGs now accuratly represent structural interactions in RNA.

4.5.1.3. Read depth

Both RING and spectral clustering calculations require a large number of full-length sequences that map to the RNA of interest to produce high quality results. Both calculations are sensitive to the presence of random mutation events in the cDNA sequence. These calculations cannot distinguish between random mutations or those caused by nucleotide adducts. Having a high number of reads for each sample ensures that mutations considered in these calculations are a result of adducts and not random background mutations. These calculations consider every A and C position in an RNA sequence and so the number of reads required to produce high quality data is proportional to the length of the RNA of interest. When determining the number of reads needed for a sample it is important that only full-length sequences are used. All positions in an RNA must be equally represented in an RNA sample to ensure that resulting correlations or clusters represent the entire RNA structure. The number of full-length sequences is greatly affected by the quality of the sequencing data. Sequencing data for used in this work was generated using a MiSeq instrument. Generally, the quality of sequencing reads decline as the cDNA is sequenced on MiSeq platforms. In order to ensure that the maximum number of full-length sequences are generated an appropriate Illumina MiSeq kit should be used in paired end more so that the first sequencing read from the 3' end covers the entire RNA sequence of interest. The remaining cycles should be used in the second sequencing read from the 5' end in order to raise the sequence quality at the 5' end of the RNA. Table 4.2 outlines the number of reads mapped to an RNA sequence from a MiSeq instrument and the number of those reads which are full-length sequences for each RNA. The comparison of these values can serve as a guideline to determine the quality of sequencing runs.

Furthermore the full length reads mapped to the RNA must also contain two or more mutations for RING or clustering calculations to accurately represent native RNA structure. The number of reads with two or more mutations is outlined in Table 4.2 for each sample. The number of full length reads given for each RNA of increasing size should serve as guide to determine the number of reads required for future experiments.

Table 4.2. Summary of processed massively parallel sequencing data for RNA nucleotide association analysis and spectral clustering. Clustering analysis requires reads that have two or more mutation events. The fourth column lists the number of nucleotides with mutation frequencies greater than 0.01 used in spectral clustering analysis.

	Condition	# of reads mapped to RNA	# of full length reads	# of Reads with ≥ 2 mutations	# of nucleotides used for clustering
TPP	Saturating ligand	87603	80465	51036	17
	No ligand	92593	83732	62731	21
	No ligand + no Mg^{2+}	92667	69805	62203	28
	200 nM Ligand	95283	87592	56732	19
P546 domain Native	Mg^{2+}	413205	182897	181399	44
	no Mg^{2+}	486526	198993	198128	57
P6a mut	Mg^{2+}	427618	234066	229775	50
	no Mg^{2+}	434924	212381	210451	55
J5 mut	Mg^{2+}	621159	216419	214468	45
	no Mg^{2+}	461769	186863	183724	46
RNase P	Mg^{2+}	808239	727877	724240	71
	no Mg^{2+}	641156	561986	561017	96

4.5.2. RING analysis

RINGs measure a wide range of nucleotide interactions and so it can often be difficult to identify important interactions, specifically when the data is noisy. A few methods can be

employed to help identify the most useful interactions within an RNA. Within the test set of RNAs it was observed that most correlations occur between positions that are reactive in the native structure or those become reactive in the no Mg^{2+} condition (Fig. 4.4C, red circles). By using a no Mg^{2+} control a complicated RING analysis can be quickly evaluated by only considering correlations between positions that are reactive in the native structure or become reactive in the no Mg^{2+} sample. It is also possible to calculate the standard deviation of each correlation measurement through a boot strapping measurement. This allows us to determine of the quality of each correlation measurement. By considering only nucleotides with a standard deviation below a certain threshold, inaccurate correlations can be eliminated. Only correlations with a standard deviation less then 20% of the correlation coefficient should be considered. In this work, bootstrapping iterations were sufficiently large so that the absolute difference between bootstrapped and calculated correlation coefficients was less than 1% of the calculated correlation coefficient. If RINGs include a large fraction of correlations that are not between reactive positions or have large standard deviations, one or more of the criteria used to qualify RING-Map data should be addressed.

4.5.3. Spectral clustering Analysis

Spectral clustering analysis allows for the identification of multiple native structures in a single pool of RNA. The ability to accurately identify and separate distinct conformations by spectral clustering analysis can be determined by three criteria; the eigengap values, fraction of the minor cluster, and the differences between the cluster profiles.

4.5.3.1. Eigengap value analysis

Determining the number of clusters that can be definitively separated is an important first step in evaluating clustering data. The quantity and quality of cluster separation can be determined by evaluating the eigengap plot. If a sample has multiple, well-defined clusters the sample will have an eigengap of 0.03 or greater. The number of possible clusters can be determined by the position of the eigengap. A sample with two clusters will have an eigengap greater than 0.03 in the second position, three clusters will have an eigengap greater than 0.03 in the third position, and so on. Figure 4.14 outlines the expected eigengap profiles for samples with multiple clusters.

4.5.3.2. Fraction of the minor cluster

The clustering algorithm partitions all sequences into a derived cluster and so the fraction of each cluster in an RNA sample can be determined. However, reactivity profiles for small fractions cannot be reliably determined. In this study, fractions of less then 10% were not considered well-resolved clusters. If a sample has a low population cluster it should not be considered when producing the reactivity profiles for different clusters.

4.5.3.3. Difference in cluster profiles

It is also important to compare the resulting profiles from the separated clusters, especially when eigengap values are close to 0.03. Clusters, which represent distinct structures, can be identified when reactivity at certain positions are preferentially represented in a single cluster. It is important to note that a sample with poorly defined clusters will still produce reactivity profiles for each cluster. While poorly defined cluster profiles may show small differences in reactivity, no position will be exclusively represented in either cluster. As a result, the separate profiles describe the same RNA conformation.

4.5.3.4. Analyzing the TPP riboswitch.

The analysis of the TPP riboswitch with decreasing levels of structure serves as a good example of how clustering data should be analyzed. The eigengap values indicate that the saturating ligand sample has potentially thee clusters, the no-ligand sample has two, and the no Mg^{2+} sample has one cluster (Fig. 4.18A). The smallest cluster of the saturating ligand sample is only 5%, so only two reactivities profiles can be generated for this sample (Table 4.1). For the no-ligand and no Mg^{2+} samples, the eigengap values are just above and below the eigengap threshold, respectively. As a result, the differences between the resulting profiles should be compared to determine if each cluster represents a distinct conformation (Fig. 4.18B). The no ligand sample has a few nucleotides, indicated by blue circles, which are exclusively represented in a single cluster. This verifies that each profile represents a distinct conformation. The no Mg^{2+} sample produces profiles in which position both increase and decrease in reactivity between cluster profiles but no positions are exclusively represented in one cluster. As a result, the no ligand sample contains only one conformation.

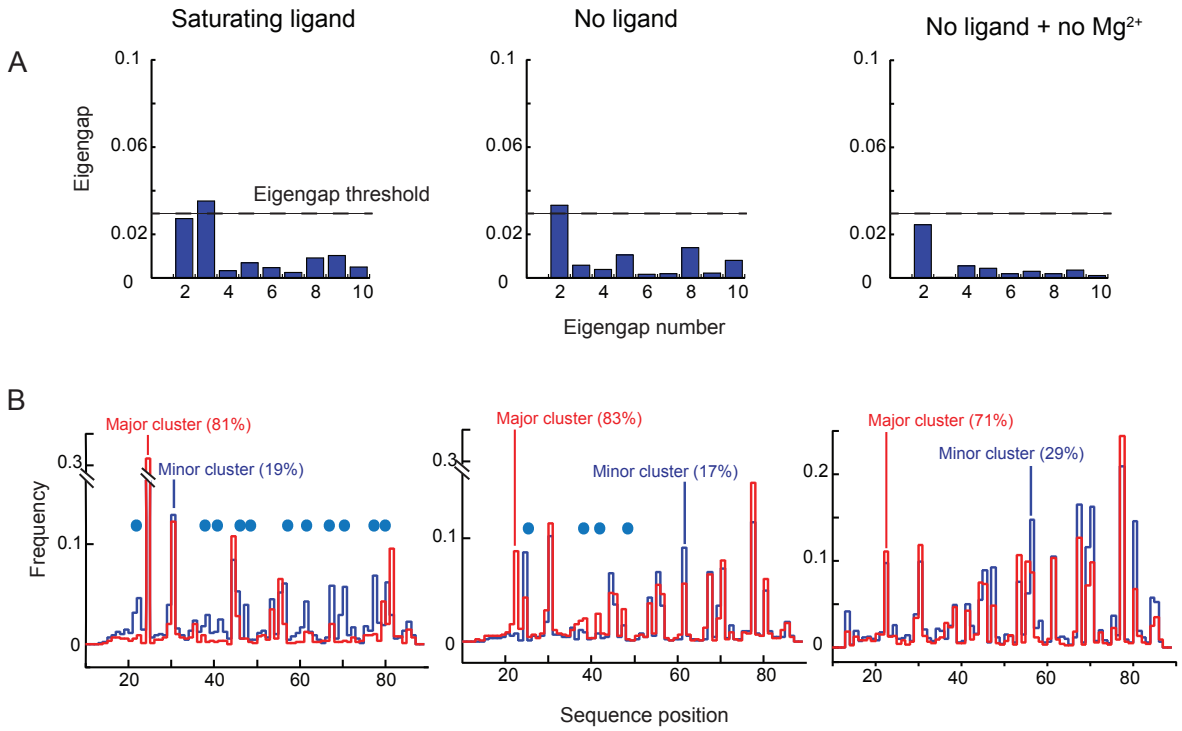


Figure 4.18 Determining number of clusters in TPP samples using eigengap values and cluster profiles. (**A**) Eigengaps of saturating ligand, no ligand and no Mg^{2+} indicate that samples have 2,1 and 0 clusters respectivly. How well samples are seprated can be determined by how above or below the eigengap threshold egengap values are. Samples with multiple positons above the egiengap threshold determine number of clusters. (**B**) reactivity profiles for culsters derived from TPP samples. The major cluster (Red) is compaired to minor cluser (Blue). The more difinitivly separated clusters have positions in which are preferentially represented in a single cluster (blue circles). The poorley separated, no Mg^{2+} sample produces two clusters with few major differences between reactive positions.

4.5.3.5. Additional criteria for evaluating data for spectral clustering.

In addition to the general criteria set for all RING-MaP data, additional criteria must be considered to determine if data is suitable for clustering analysis. A mutation frequency threshold was used to set the number of positions used for clustering analysis. Only positions with a sufficiently high mutation rate were considered. In this work, only positions with a mutation frequency greater than 0.02 were used. The numbers of positions used in spectral clustering calculations for each RNA are outlined in the final column of Table 4.2. The number of positions used should reflect the number of single-stranded A and C nucleotides in an RNA. If the number of nucleotides analyzed does not fit this criteria for spectral clustering, it may reflect a larger problem with the sample quality and one or more of the criteria used to determine the quality of RING-MaP data should be addressed.

RING-MaP provides data that can be analyzed to identity structurally correlated nucleotides and the presence of multiple structures within a pool of RNA. These analyses provide a unique view of RNA structure that has not been possible in a single experiment before. The criteria outlined here provide a simple set of guidelines to assist data analysis of new RNAs and an efficient way to evaluate data used for RING-MaP calculations.

4.6. ACKNOWLEDGEMENTS

This work was supported by grants from the NIH to K.M.W.: AI068462 (for development of MaP technologies) and GM064803 (for three-dimensional structure refinement).

4.7. REFERENCES

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

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

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

(4) E. A. Dethoff, J. Chugh, A. M. Mustoe, H. M. Al-Hashimi, Functional complexity and regulation through RNA dynamics, Nature 482, 322–330 (2012).

(5) N. A. Siegfried, et al. & K. M. Weeks, submitted (2014).

(6) P. Ryvkin et al., HAMR: high-throughput annotation of modified ribonucleotides, RNA 12, 1684–1692 (2013).

(7) J. Spitzer et al., in Methods Enzymol, (Elsevier Inc., 2014), vol. 539, pp. 113–161.

(8) J. Shendure, H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. 26, 1135–1145 (2008).

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

(10) J. H. Cate et al., Crystal structure of a group I ribozyme domain: principles of RNA packing, Science 2733, 1678–1685 (1996).

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

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

(13) A. A. Szewczak, T. R. Cech, An RNA internal loop acts as a hinge to facilitate ribozymefolding and catalysis, RNA 3, 838–849 (1997).

(14) N. Kulshina, T. E. Edwards, A. R. Ferre-D'Amare, Thermodynamic analysis of ligand binding and ligand binding-induced tertiary structure formation by the thiamine pyrophosphate riboswitch, RNA 16, 186–196 (2009).

(15) S. E. Butcher, A. M. Pyle, The Molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks, Acc. Chem. Res. 44, 1302–1311 (2011).

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

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

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

(19) C. E. Hajdin, F. Ding, N. V. Dokholyan, K. M. Weeks, On the significance of an RNA tertiary structure prediction, RNA 16, 1340–1349 (2010).

(20.) E. J. Merino, K. A. Wilkinson, J. L. Coughlan, K. M. Weeks, RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE), J. Am. Chem. Soc. 127, 4223–4231 (2005).

(21) J. König, K. Zarnack, N. M. Luscombe, J. Ule, Protein–RNA interactions: new genomic technologies and perspectives, Nat. Rev. Genet. 13, 77–83 (2012).

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

(23) F. Yates, Contingency Tables Involving Small Numbers and the $\chi 2$ Test, Supp. J. Roy. Stat. Soc. 1, 217–235 (1934).

(24) J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22, 888–905 (2000).

(25) A. Y. Ng, M. I. Jordan, Y. Weiss, in Advances in Neural Information Processing Systems, T. Dietterich, S. Becker, Z. Ghahramani, Eds. (MIT Press, 2001), pp. 849–856.

(26) U. von Luxburg, A tutorial on spectral clustering, Stat. Comp. 17, 395–416 (2007).

(27) F. Ding, C. A. Lavender, K. M. Weeks, N. V. Dokholyan, Three-dimensional RNA structure refinement by hydroxyl radical probing, Nat. Methods 9, 603–608 (2012).