HIGH-THROUGHPUT EXPERIMENT DRIVEN MODELING OF RNA INTERACTIONS AND STRUCTURES

Greggory Mathew Rice

A dissertation submitted to 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
2015

Approved by:

Gary Pielak

Alain Laederach

Nikolay Dokholyan

David Lawrence

Kevin Weeks

ABSTRACT

GREGGORY MATHEW RICE: High-throughput Experiment Driven Modeling of RNA Interactions and Structures
(Under the direction of Kevin M. Weeks)

The higher order structure of an RNA is often essential to its biological function, modulating its interactions with ligands, protein partners, and other RNAs. Modeling RNA secondary structure, assessing the accuracy of RNA structural models, and discovering new functional motifs are challenging problems that are confounded by the length and complexity of the studied RNA. Improvements in structure modeling accuracy can be made by incorporating SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) and DMS (dimethyl sulfate) chemical probing data, however these models remain imperfect. In this work I apply principals of molecular modeling in interpret chemical probing experiments and create analytical and experimental tools that enable large-scale experiment-driven modeling of RNA interactions and structures. First, I use electronic structure modeling to propose a mechanism to explain the preferential reactivity of the SHAPE reagent 1M6 at nucleotides exhibiting an available open stack in folded RNAs. I also use molecular modeling at the nucleotide level to develop a model that accurately predicts the disruptive effects of SHAPE adducts on RNA tertiary structure. Second, I create a new energy potential for RNA structure prediction using information from a three-reagent SHAPE experiment that increases the accuracy of modeling accuracy from 85% to above 90% for some of the most difficult-to-predict RNA structures. Third, working collaboratively with others in the lab, I validate a new approach for melding SHAPE chemical probing with deep sequencing in a new technique termed SHAPE mutational profiling (SHAPE-MaP). The ability to quickly generate structural data for RNAs of unprecedented size using SHAPE-MaP presents a new challenge: accurately modeling large RNA secondary structures. To solve this problem I develop software, called *SuperFold*, which uses a windowed modeling algorithm

to enable rapid secondary structure prediction and discovery of the most stable structural motifs in long RNAs. Fourth, I extend the RING-MaP experiment and analysis to enable use with random primers and applied it to the bacterial ribosome. With the improved RING-MaP experiment I am able to detect structural interaction networks within the small ribosomal subunit. Additionally, I am able to perturb those structural networks by adding the antibiotic spectinomycin. Coupled together, the work presented here will provide valuable tools that democratize RNA structure analysis and help others in the RNA community understand the role of RNA structure at new and exciting scales.

"…If one advances confidently in the direction of his dreams, and endeavors to live the life which he has imagined, he will meet with a success unexpected in common hours."

–Henry David Thoreau

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Kevin. Thank you for giving me the freedom to try new ideas (that often didn't work) and letting me learn by doing. I will miss going into your office to bounce a new idea off you. I have become a better writer and scientist because of you.

I would also like to thank the wonderful teachers and mentors that I've had at UNC, specifically Barry, Gary, Alain, Nikolay, and many others. You have given me the foundation and provided the support for me to succeed in graduate school and beyond.

Next, I would like to thank the members of the Weeks group past and present. I couldn't have asked for a better group of scientists (and friends) to work with. Your collective wisdom is a force to be reckoned with.

To team SHAPE-MaP, thank you for reminding of me the value of close collaboration. Together we accomplished something that alone would not have been as great. At times it was challenging, but in the end, we made it.

To the lifelong friends that I have made since coming to UNC, you guys rock. Matt, Fatima, Kathleen, Tim, and Hannah you have become a second family for me. Graduate school has been easier because of you all. Wherever I am in the world, you will always have a place to stay.

Last, I would like to thank my family: my parents, sisters, grandparents, aunts, and uncles. Thank you for always being there. I feel that I can accomplish anything, no matter where I am, with your support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

$^{2'}$OH	ribose 2'-hydroxyl
1M6	1-methyl-6-nitroisatoic anhydride
1M7	1-methyl-7-nitroisatoic anhydride
A	adenosine
Asp	aspartic acid
C	cytidine
CA	capsid
cDNA	complementary DNA
CE	capillary electrophoresis
CMCT	*N*-cyclohexyl-*N*'-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate
CO_2	carbon dioxide
cryo-EM	cyro-electron microscopy
dA	deoxyadenosine
DMD	discrete molecular dynamics
DMS	dimethyl sulfate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
ENV	Envelope
G	guanosine
g	gram
g	gravity
HCV	hepatitis C virus
HEPES	N-2-hydroxyethylpiperazine-N′-ethanesulfonic acid
HIV-1	human immunodeficiency virus type 1

HMX	2'-hydroxyl molecular interference
IRES	internal ribosome entry site
k_B	Boltzmann constant
KCl	potassium chloride
kcal	kilocalorie
ln	natural log
LNA	locked nucleic acid
M	molar
MaP	mutational profiling
MDa	mega Dalton
$MgCl_2$	magnesium chloride
min	minute
mL	milliliter
mM	millimolar
mol	mole
NaCl	sodium chloride
ng	nanogram
NGS	next-generation sequencing
nM	nanomolar
NMIA	*N*-methylisatoic anhydride
NMR	nuclear magnetic resonance spectroscopy
nt	nucleotide
OD_{600}	optical density at 600 nanometers
PARS	parallel analysis of RNA structure
PCR	polymerase chain reaction
Phe	phenylalanine

PK	pseudoknot
PPT	polypurine tract
ppv	positive predictive value
Rif	rifampicin
RING-MaP	RNA interaction groups identified by mutation profiling
RNA	ribonucleic acid ribonuclease
rpm	revolutions per minute
RRE	Rev response element
rRNA	ribosomal RNA
RT	reverse transcriptase
sens	sensitivity
SDS	sodium dodecyl sulfate
SHAPE	selective 2'-hydroxyl acylation analyzed by primer extension
SHAPE-MaP	selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling
SHAPE-seq	selective 2'-hydroxyl acylation analyzed by primer extension sequencing
Spc	spectinomycin
STMV	satellite tobacco mosaic virus
TPP	thiamine pyrophosphate
T	temperature
TAR	trans-activation response element
TE	10 mM Tris (pH 8.0), 1 mM EDTA
TPP	thiamine pyrophosphate
tRNA	transfer RNA
U	uridine
UTR	untranslated region
Å	Ångstrom

°C	degree Celsius
ΔG	Gibbs free energy change
μg	microgram
μL	microliter
μM	micromolar

CHAPTER 1: THE IMPORTANCE OF RNA STRUCTURE

Establishment of the structure-function relationship in ribonucleic acid

During the early years of molecular biology in the 1950s, ribonucleic acid (RNA) was considered a passive intermediate carrying information coding for the sequence of proteins stored in deoxyribonucleic acid (DNA) to the ribosome (Crick 1970). In the intervening decades since the postulation of this now-famous central dogma, RNA has been found to be a far more versatile molecule than first thought. RNA is able to affect gene expression through a number of complex systems including alternative exon splicing (Amaral et al. 2008), microRNA (Ha and Kim 2014) and RNA interference (Fire et al. 1998), and non-coding RNAs (Storz et al. 2011; Geisler and Coller 2013).

The ability of an RNA to fold back on itself and form higher order structures likely plays an essential role in many of these biological systems (Sharp 2009). RNA often folds three-dimensionally into a single functionally relevant structure. RNA structure can be described on three different hierarchical levels, with each level encoding essential and increasingly complex information. The first, and simplest level is the primary sequence which consists of the linear linkage of the four RNA nucleotides: adenine (A), cytosine (C), guanine (G), and uracil (U) (Leontis et al. 2006). Many biological systems operate solely on the primary sequence of an RNA, such as RNA interference. However, the next level of RNA structure –secondary structure, can modulate the efficiency of these systems.

At the secondary structure level, the RNA polymer bends back on itself and the nucleic acid bases are able to form hydrogen-bonding interactions known as base pairs. The canonical base pairs found in RNA are, in decreasing order of strength, G-C, A-U, and G-U base pairs. Continuous stretches of base pairs form helices and are connected by non-paired stretches of nucleotides that are

typically described as loops, bulges, and single stranded regions (Fig. 1.1) (Tinoco and Bustamante 1999; Leontis et al. 2006). Together, the helices of an RNA arrange to form tertiary structure. RNA tertiary structure is stabilized by non-canonical base pairing, base stacking, and hydrogen bonding (Fig. 1.1b). The tertiary structure of an RNA can also play important functional roles in gene regulation, splicing, and even protein synthesis. In prokaryotes (and some eukaryotes), riboswitches are small RNA elements that change their structure based on recognition of small molecule ligands (Serganov et al. 2006; Dethoff et al. 2012). In the case of riboswitches, the proper formation of tertiary structure is essential for ligand binding and gene regulation.

High resolution methods for determining RNA structures

After the discovery of functional RNAs, high-resolution structure determination methods that were first applied to proteins were applied to RNA. One such technique is X-ray crystallography, which, when suitable crystals are obtained, can often determine the atomic-level tertiary structure of a molecule based on the diffraction pattern produced by X-rays traveling through the crystallized molecule of interest. X-ray crystallography was first applied to RNA in order to determine the structure of 76-nucleotide tRNA in 1974 (Kim et al. 1974; Klug et al. 1974). Decades later, crystallography has revealed molecular details of the ~3000-nucleotide large ribosomal subunit (Ban et al. 2000) and the small and large ribosomal subunit complex (Korostelev et al. 2006). Nuclear magnetic resonance spectroscopy (NMR), which works on solutions of molecules, was applied to studying RNA structure later than crystallography. NMR revealed not only structural features, but the dynamic motion of RNA as well (something that crystallography is unable to do) (Cheong et al. 1990; Al-Hashimi and Walter 2008).

Despite the ability of X-ray crystallography and NMR to reveal the molecular details of RNA structure in exquisite detail these techniques have many limitations. X-ray crystallography has not been as successful with RNA as it has been with proteins owing to the fact that most RNAs are difficult to crystallize. Because the NMR signal exhibits a high degree of similarity between the

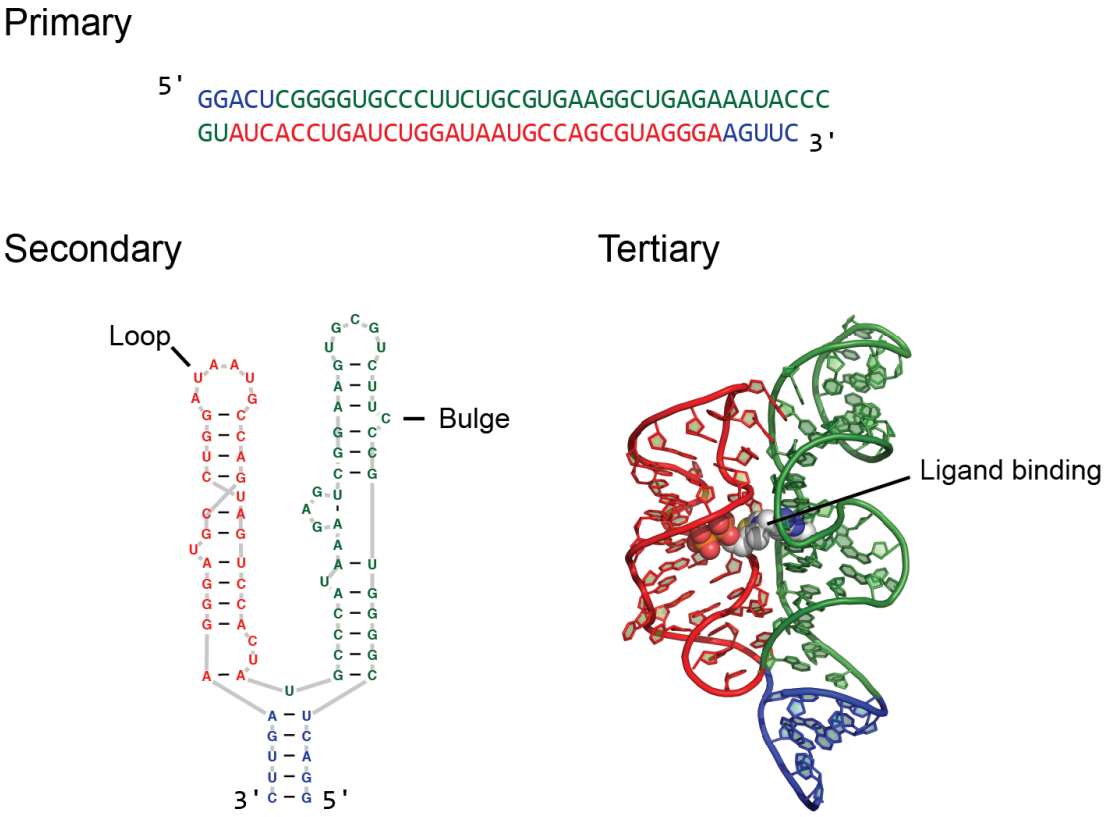

Figure 1.1: Explaination of the levels of RNA structure. The primary structure of RNA consists of four different kinds of nucleotides linked together as a long chain. Nucleotides come together to form base pairs in the secondary structure, shown as black lines in the secondary structure. Base pairing interactions ranked from strongest to weakest are GC > AU > GU. Stretches of base pairs form helices that are broken up by loops and bulges. Helicies arrange together to form the tertiay structure. This RNA is the TPP riboswitch that is able to recognize the ligand Thiamine pyrophosphate (TPP) (PDB code 2GDI). Colors are consistent throughout the panels to highlight how nucleotides come together to form higher order structure.

different nucleobases, NMR is limited to small RNAs –typically less than 100 nucleotides in length– without heroic efforts of partially enriching segments of RNA molecules with NMR active atomic isotopes (Lu et al. 2011). Because of these drawbacks, both NMR and X-ray crystallography are only useful to a small fraction of all RNAs that are of functional interest.

RNA foot-printing methods for determining structure

Determining the secondary structure of an RNA is often key to understanding its function. However flexibility, structural heterogeneity, and time constraints make high-resolution methods intractable for many RNAs. Experimental evidence for base pairing can be obtained using ribonuclease enzymes (RNases) or chemical agents that are sensitive to base pairing and structural flexibility. Some RNases can cleave RNA in a structure specific manner: RNase V1 cleaves at base paired nucleotides whereas RNase S1 cleaves at single stranded nucleotides (Ehresmann et al. 1987). Despite their structure selectivity, RNase enzymes are large relative to small molecules and often have a sequence bias in the sites they will cleave. One small molecule reagent that reacts with RNA in at unpaired nucleotides is dimethyl sulfate (DMS), which reacts most detectably at the pairing face of A and C nucleobases to form a methyl adduct (Fig. 1.2a). However, determining locations of DMS reactivity is difficult at G nucleotides and DMS does not react broadly at U nucleotides.

Within the last decade SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) has become popular because of its ability to react at all four nucleotides and its convenient hydrolysis quench reaction with water (Merino et al. 2005; Wilkinson et al. 2006; Weeks and Mauger 2011). SHAPE uses electrophiles (typically variations of an isatoic anhydride scaffold) that react at the 2'-hydroxyl position of the ribose sugar (Fig. 1.2b). The reactivity of the hydroxyl position is modulated based on nucleotide flexibility, with flexible nucleotides being the most reactive (Gherghe et al. 2008; McGinnis et al. 2012). Since most RNA nucleotides have a 2'-hydroxyl, SHAPE chemistry provides a generic method to obtain flexibility information at all four nucleotides. Additionally, SHAPE reagents

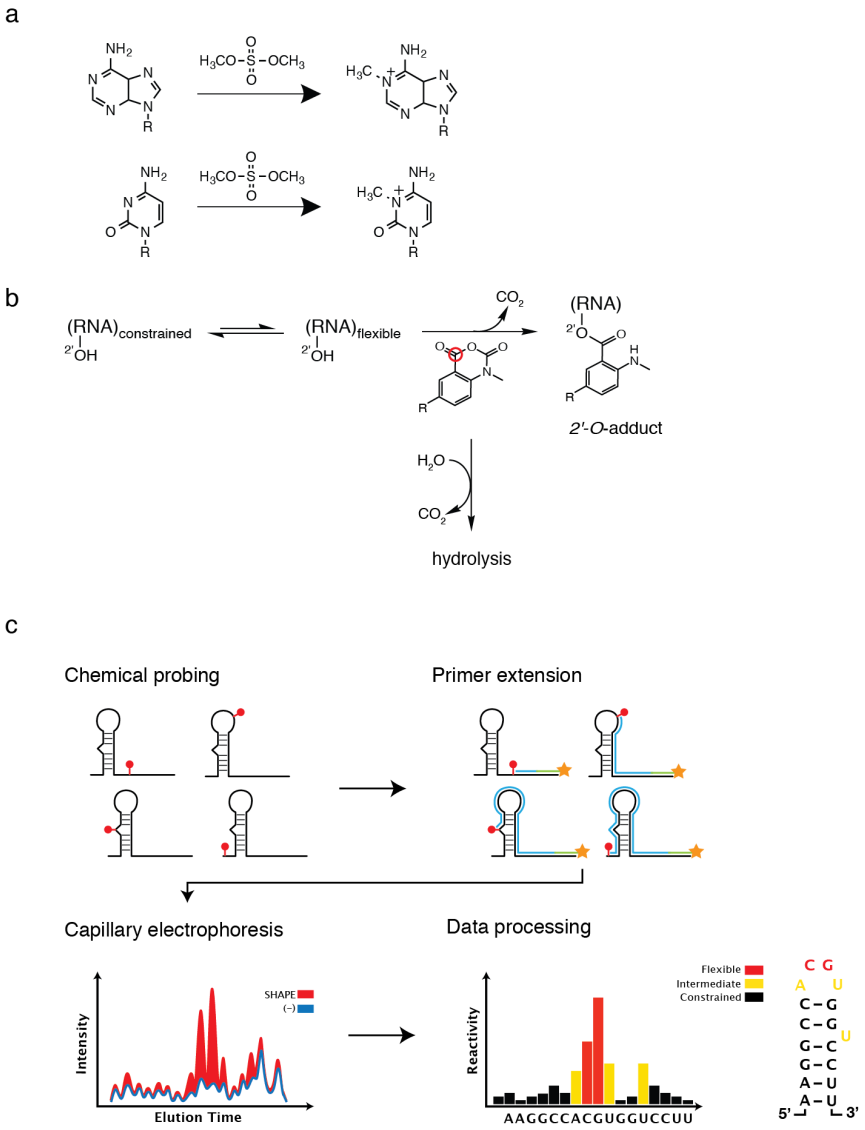

Figure 1.2: Chemical probing methods and detection by capillary electrophoresis. (A) Dimethyl sulfate (DMS) reacts on the base pairing face of single stranded A and C nucleotides to form covalent adducts. (B) RNA nucleotides can sample both constrained and flexible conformations. Flexible conformations of RNA (such as those in loops, bulges and single stranded regions) react more readily at the ribose 2'-hydroxyl with isatoic anhydrides to form bulky *2'-O*-adducts. The reactive position of the SHAPE reagent is circled in red. Different variations on the isatoic anhydride scaffold are possible by varying functional groups around the aromatic ring (R). A competeing hydrolysis with water inactivates the SHAPE reagent, limiting the amount of modification that can occur. (C) Detection of RNA adducts by primer extension. A pool of RNAs is modified with a structure selective reagent (red lolipops). Next a fluorescently labeled primer is annealed at the 3' end of the RNA. Reverse transcriptase extends the primer, creating a cDNA, until it encounters an adduct or break in the RNA and dissociates. Fluorescently labeled cDNAs are resolved using capillary electrophoresis and the peak intensity, corresponding to the number of adducts, is subtracted from a background control. After data processing, the reactivity of each nucleotide corresponds to its structural context.

and DMS are able to pass through cell membranes and can be used to probe the structure of RNA in living cells (Spitale et al. 2013; Tyrrell et al. 2013; Ding et al. 2014; McGinnis and Weeks 2014). Both nuclease and chemical probing experiments are quantified using reverse transcription primer extension (Low and Weeks 2010; Karabiber et al. 2013). Using this detection method, a fluroescently labeled primer is extended from the 3 prime end of an RNA by a reverse transcriptase enzyme (Fig. 1.2c). When reverse transcriptase encounters an adduct, or a break in the RNA in the case of nucleases, it is unable to proceed and dissociates. The cDNA products are resolved using capillary electrophoresis and quantified based on their intensity. In order to account for inherent reverse transcription pauses, a no reagent control is used to subtract the background signal away from the reagent signal.

Secondary and structure prediction methods

Since the RNA "alphabet" is so small, consisting of only four "letters" (as compared to the canonical 20 amino acid alphabet of proteins) bioinformaticions have been spent considerable time creating computer algorithms to predict the secondary structure of RNA. The energetics of RNA base pairing and stacking, determined from experiments, have been incorporated to create a nearest neighbor free energy potential (Mathews et al. 1999; Mathews and Turner 2006; Reuter and Mathews 2010). These algorithms are able to obtain accuracy between 50-70% depending on the sequence. One typical assumption is that there are zero non-nested base pairs. These non-nested base pairs, known as pseudoknots (Staple and Butcher 2005), are known to occur in nature and are often in functionally important RNAs. For example, internal ribosomal entry site (IRES) element of hepatitis C virus (HCV) contains a pseudoknot, and enables the virus to hijack cellular replication machinery; when the pseudoknot is disrupted, the virus is unable to function (Nicholson and White 2014). Pseudoknots are also critical to the function of some small riboswitches (Serganov et al. 2008). Despite the ubiquity of prediction algorithms, the structure of RNA is dependent on more than just the sequence alone. Factors such as tertiary structure interactions, kinetic folding pathways, and

ligand binding are difficult to extract from sequence alone and can significantly decrease the accuracy of RNA structure modeling. In order to compensate for these interactions, some RNA modeling programs have successfully incorporated SHAPE and DMS chemical probing data (Deigan et al. 2009; Hajdin et al. 2013) (Cordero et al. 2012) to significantly increase the accuracy of generated models.

Rise of deep sequencing and coupling to structure probing methods

One of the consequences of the human genome project was the rapid development of platforms that enable rapid sequencing of DNA bases (International Human Genome Sequencing Consortium 2004). Rather than sequencing a single stretch of DNA, deep sequencing platforms sequence millions of individual DNA fragments simultaneously, on a massively parallel scale. The current state of the art, and widely used platform is marketed by Illumina. This platform works using a sequencing by synthesis approach with reversible terminator chemistry (Bentley et al. 2008)(Fig. 1.3), and requires that all DNA fragments have the same sequences on the 5' and 3' ends of the molecules to act as molecular "handles" for PCR amplification and sequencing primer binding sites.

RNA can be indirectly sequenced on massively parallel sequencers by first copying the RNA into complementary DNA (cDNA) using reverse transcriptase enzymes. Because of this simple transformation to DNA, several RNA structure groups have adapted structure-probing approaches to deep sequencing scales (Table 1.1). Nearly all of these methods attempt to recover the location of the 5' end of the cDNA (where reverse transcriptase dissociated from encountering a chemical probing adduct or reached the end of an RNA cleaved by a nuclease). In order to recover the location of the end of the cDNA, a ligation step is required. These ligation steps are inefficient and biased in unpredictable ways (Weeks 2011). Finding a new method to determine the location of adducts formed by structure-sensitive reagents would be transformative in the field of RNA structure modeling

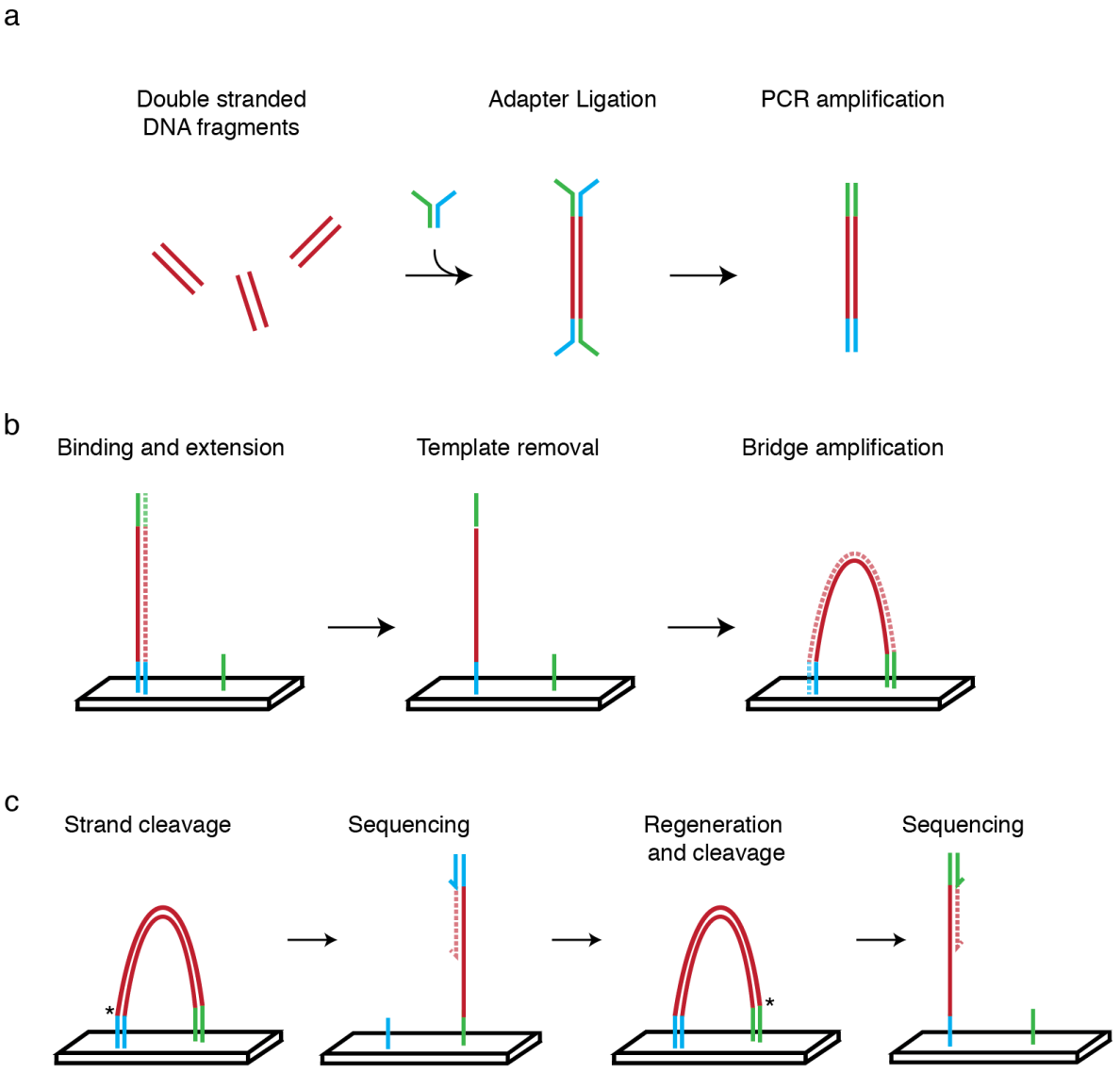

Figure 1.3: Explaination of paired end sequencing. (A) Sequencing library construction. Starting with a complex mixture of short double stranded DNA fragments (red lines), semi-complementary adapters (green and blue lines) are ligated on to the ends on the double stranded DNA. This ligation results in a library of DNA fragments with both a green and a blue sequence. PCR amplification is performed using sequences complementary to the known green and blue stretches. (B) Next, the the amplified DNA is denatured and flowed over a glass surface fuctionalized with sequences of DNA complementary to the adapters where it binds (green and blue sequeces). The template is extended (dashed line) and removed using a denaturation step. Since both green and red sequences are present, several cycles of PCR can be performed in a brige amplification to create clonal copies near in space on the glass support. (C) Prior to sequencing, a restriction enzyme cleaves one end of the template and one strand is removed –leaving all the sequences on the support in the same orientation. A sequencing primer is bound (blue) and sequencing is performed one cycle at a time in the direction of the plate. Following the first round of sequencing, bridges are re-amplifed and the opposite end is cleaved. A second sequencing primer is bound to (green) end and sequencing proceeds at the other end of the libraries. This figure is adapted from Bentley *et al* 2008.

Method name	Citation	Reagent	RNA folding	Organism	Ligation?
PARS	(Kertesz et al. 2010) (Wan et al. 2014)	RNase V1(dsRNA) RNase S1 (ssRNA)	*In vitro*	Yeast, Human	Yes
Frag-seq	(Underwood et al. 2010)	P1 nuclease (ssRNA)	*In vitro*	Mouse	Yes
DMS-seq	(Rouskin et al. 2014)	DMS	*In vitro*	Yeast	Yes
Structure-seq	(Ding et al. 2014)	DMS	*In vivo*	Arabidopsis	Yes
Mod-seq	(Talkish et al. 2014)	DMS	*In vivo*	Yeast	Yes
CIRS-seq	(Incarnato et al. 2014)	DMS, CMCT	*In vivo*	Mouse	Yes
SHAPE-MaP	(Siegfried et al. 2014)	1M7	*In vitro*	HIV	No

Table 1.1: A selection of deep sequencing structure probing methods. A large number of have created methods to couple secondary struture probing to next generation sequencing. With the exception of SHAPE-MaP, most methods require the use of a single stranded ligation at the cDNA or RNA level to detect structure informative stops. This ligation step often introduces bias into the final pool of sequences that is difficult to correct.

Research overview

The overall goals of this project are three-fold. First, to use computational modeling in order to help interpret chemical probing experiments and second, to create new approaches that enable the study of RNA structure at large scales. Finally, I want to use the understanding and new technical achievements gained in the first two goals to find and solve important, biologically relevant problems.

In Chapter 2, I highlight my endeavors to give a molecular interpretation of two different chemical probing phenomenon by applying principles of molecular modeling. The first part of the chapter is devoted to explaining the preferential SHAPE reactivity of 1-methyl-6-nitroisatoic (1M6) to positions in folded RNAs that have available base stacking interactions. Using density functional theory (DFT) calculations I showed that 1M6 interacts more favorable at these positions relative to other reagents. In the second part of the chapter I developed a molecular overlap model to predict how disruptive an adduct at the 2'-hydroxyl position of the ribose would be to the folding of the tertiary structure of an RNA. My model gave a correlation relative to a real experiment as high as R=0.7 in several RNAs.

RNA secondary structure prediction is a challenging problem that only increases in difficulty as the length of the RNA increases. In Chapter 3, I define a novel energy potential to constrain RNA modeling predictions using information gained from position specific differences in reactivity by different SHAPE reagents. This energy potential is able to increase the accuracy of already "good" predictions to above 90% ("excellent") across several RNAs with structures that were previously difficult to predict accurately. This increase in prediction accuracy is important since most RNAs of global interest, such as HIV-1, HCV, and Dengue virus, are thousands of nucleotides in length.

In Chapter 4, I work with two colleagues in order to validate a new method that couples deep sequencing with SHAPE chemistry (SHAPE-MaP), utilizing a reverse transcription mutational profiling approach to uncover the location of structure-specific adducts. Part of this work involved creating an automated motif discovery algorithm (*Superfold*) to uncover novel functionally important regions in HIV-1 genome. The automated motif discovery approach solves a problem created by the

ability to quickly generate large amounts of data. Using these tools we uncover and validate three previously unreported pseudoknots in HIV-1 –a remarkable feat for a virus that has been intensely studied for more than twenty years.

Finally, in Chapter 5, I extend the RING-MaP (Homan et al. 2014) experiment and analysis with random priming in order to uncover structural dynamics in the bacterial ribosome. Previously, the RING-MaP experiment was limited to only small RNAs due to inefficiencies in reverse transcription and challenges in the analysis. In order to address these challenges, I found new conditions to dramatically improve the efficiency of reverse transcription in the presence of adducts and created a new algorithm that enables the rapid analysis of long RNAs (>10,000 nts). This extension will become an excellent tool for validating new structures in long RNAs.

Perspective

In this project I apply principles from biochemistry, physical chemistry, and molecular biology in order to address the challenge of modeling RNA structures (some of which are thousands of nucleotides in length) at large scales. I use chemical and molecular modeling to give support for observed biochemical phenomena and create a new energy potential based on elements of tertiary structure detected from biochemical probing experiments in order to increase the accuracy of RNA secondary structure models. Additionally I validate a new technique (SHAPE-MaP) to couple RNA structure probing with deep sequencing, create a structure modeling package (*SuperFold*) to *de novo* discover well structured regions in long RNAs, and extend the RING-MaP approach (both experiment and analysis) to work with random priming and long RNAs in order to uncover structural dynamics within the small subunit of the ribosome.

My hope is that the approaches and techniques that I have developed will be used broadly in the RNA community to uncover and enrich our understanding of RNA interactions, especially those that are relevant human disease. I expect that these methods presented here will enhance our

understanding of RNA structure at both small and transcriptome-wide scales, and will enable us to learn more about the incredibly versatile molecule that is RNA.

REFERENCES

Al-Hashimi HM, Walter NG. 2008. RNA dynamics: it is about time. *Curr Opin Struct Biol* **18**: 321–329.

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. *Science* **319**: 1787–1789.

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**: 905–920.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

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

Cordero P, Kladwang W, VanLang CC, Das R. 2012. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* **51**: 7037–7039.

Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561–563.

Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106**: 97–102.

Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM. 2012. Functional complexity and regulation through RNA dynamics. *Nature* **482**: 322–330.

Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**: 696–700.

Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res* **15**: 9109–9128.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**: 806–811.

Geisler S, Coller J. 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol* **14**: 699–712.

Gherghe CM, Shajani Z, Wilkinson KA, Varani G, Weeks KM. 2008. Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S2) in RNA. *J Am Chem Soc* **130**: 12244–12245.

Ha M, Kim VN. 2014. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* **15**: 509–524.

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci* **110**: 5498–5503.

Homan PJ, Favorov OV, Lavender CA, Kursun O, Ge X, Busan S, Dokholyan NV, Weeks KM. 2014. Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci* **111**: 13858–13863.

Incarnato D, Neri F, Anselmi F, Oliviero S. 2014. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol* **15**: 491.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Karabiber F, McGinnis JL, Favorov OV, Weeks KM. 2013. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19**: 63–73.

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**: 103–107.

Kim SH, Suddath FL, Quigley GJ, McPherson A, Sussman JL, Wang AH, Seeman NC, Rich A. 1974. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185**: 435–440.

Klug A, Robertus JD, Ladner JE, Brown RS, Finch JT. 1974. Conversation of the Molecular Structure of Yeast Phenylalanine Transfer RNA in Two Crystal Forms. *Proc Natl Acad Sci* **71**: 3711–3715.

Korostelev A, Trakhanov S, Laurberg M, Noller HF. 2006. Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* **126**: 1065–1077.

Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16**: 279–287.

Low JT, Weeks KM. 2010. SHAPE-directed RNA secondary structure prediction. *Methods* **52**: 150–158.

Lu K, Heng X, Garyu L, Monti S, Garcia EL, Kharytonchyk S, Dorjsuren B, Kulandaivel G, Jones S, Hiremath A, et al. 2011. NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**: 242–245.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.

Mathews DH, Turner DH. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* **16**: 270–278.

McGinnis JL, Dunkle JA, Cate JHD, Weeks KM. 2012. The mechanisms of RNA SHAPE chemistry. *J Am Chem Soc* **134**: 6617–6624.

McGinnis JL, Weeks KM. 2014. Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry* **53**: 3237–3247.

Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* **127**: 4223–4231.

Nicholson BL, White KA. 2014. Functional long-range RNA-RNA interactions in positive-strand RNA viruses. *Nat Rev Micro* **12**: 493–504.

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf* **11**: 129.

Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**: 701–705.

Serganov A, Huang L, Patel DJ. 2008. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* **455**: 1263–1267.

Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. 2006. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**: 1167–1171.

Sharp PA. 2009. The centrality of RNA. *Cell* **136**: 577–580.

Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Meth* **11**: 959–965.

Spitale RC, Crisalli P, Flynn RA, Torre EA, Kool ET, Chang HY. 2013. RNA SHAPE analysis in living cells. *Nat Chem Biol* **9**: 18–20.

Staple DW, Butcher SE. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **3**: e213.

Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* **43**: 880–891.

Talkish J, May G, Lin Y, Woolford JL, McManus CJ. 2014. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20**: 713–720.

Tinoco I, Bustamante C. 1999. How RNA folds. *J Mol Biol* **293**: 271–281.

Tyrrell J, McGinnis JL, Weeks KM, Pielak GJ. 2013. The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry* **52**: 8777–8785.

Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, Haussler D. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Meth* **7**: 995–1001.

Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**: 706–709.

Weeks KM. 2011. RNA structure probing dash seq. *Proc Natl Acad Sci* **108**: 10933–10934.

Weeks KM, Mauger DM. 2011. Exploring RNA structural codes with SHAPE chemistry. *Acc Chem Res* **44**: 1280–1291.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1**: 1610–1616.

CHAPTER 2: CHEMICAL INTERACTIONS OF SHAPE ADDUCTS WITH RNA IN THREE DIMENSIONAL SPACE[1]

Introduction

RNA is a central information carrier in biology (Sharp 2009). Information directing the function of an RNA is encoded at several levels. The RNA primary sequence composed of the four nucleotide alphabet arranges into base pairs to form helices. These helices further pack and arrange in order to form specific higher-order three-dimensional structure structures (Leontis et al. 2006). Higher-order RNA structures are typically comprised of secondary structure elements held together by a few key tertiary interactions (Weeks 2010; Butcher and Pyle 2011) including long-range stacking, loop-loop and loop-helix contacts, and pseudoknots. Regions of an RNA that contain significant tertiary structures ultimately have numerous important functional roles.

Nucleotides that participate in either base pairing or stable higher-order tertiary structure interactions can be detected by protection from solution-phase chemical probing reagents, whereas single-stranded and relatively unstructured elements are reactive (Wilkinson et al. 2006). Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) has emerged as an especially informative approach for probing RNA structure and dynamics (Gherghe et al. 2008; Mortimer and Weeks 2009). SHAPE chemistry exploits the discovery that the reactivity of the ribose 2′-hydroxyl is highly sensitive to local nucleotide flexibility (Fig. 2.1A). Flexible nucleotides sample many conformations, a few of which preferentially react with hydroxyl-selective, electrophilic reagents to form *2′-O-*

[1] The text and figures from this chapter was adapted from modeling experiments undertaken in collaboration with other students for two different projects. My critical contribution to the differential SHAPE project was to confirm the hypothesis that 1M6 more favorably interacts at available base stacks for RNA nucleotides compared to NMIA using density functional theory (DFT). In the HMX experiment I worked to establish the expected disruption a SHAPE adduct could cause to a folded (packed) RNA based on principles of molecular overlap. Figures and text from this chapter originally appeared in Steen, K.-A., Rice, G. M., & Weeks, K. M. (2012). Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *Journal of the American Chemical Society*, *134*(32), 13160–13163. doi:10.1021/ja304027m and Homan, P. J., Tandon, A., Rice, G. M., Ding, F., Dokholyan, N. V., & Weeks, K. M. (2014). RNA Tertiary Structure Analysis by 2′-Hydroxyl Molecular Interference. *Biochemistry*, *53*(43), 6825–6833. doi:10.1021/bi501218g

adducts (Fig. 2.1A). However, it is not obvious based on the chemical reactivity of a nucleotide whether a given constraining interaction reflects a base pairing or tertiary interaction. Nucleotides involved in tertiary interactions often have unusual backbone or stacking geometries (Holbrook 2008; Butcher and Pyle 2011), adopt the *syn* conformation (Sokoloski et al. 2011), or undergo conformational changes on slow timescales (Gherghe et al. 2008; Mortimer and Weeks 2009).

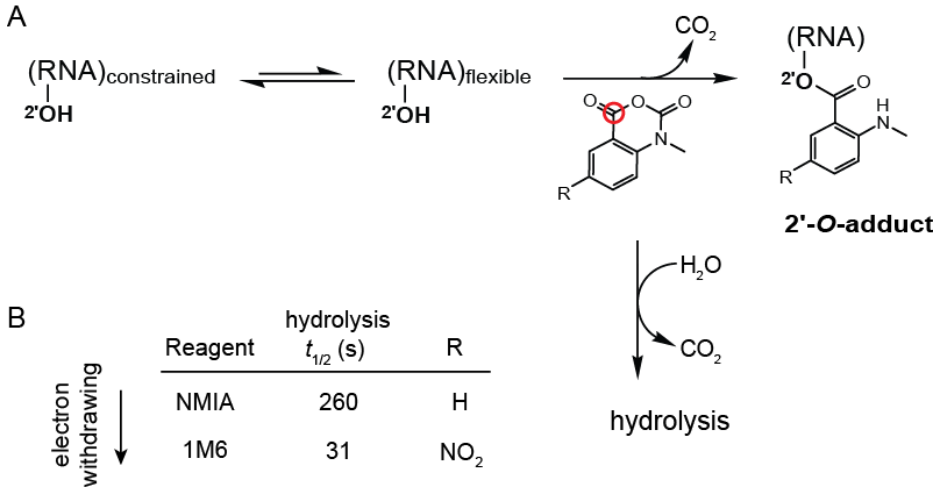

Figure 2.1: RNA SHAPE chemistry. (A) Mechanism in the context of the concurrent hydrolysis reaction. The red circle denotes the reactive center of the reagent. (B) SHAPE reagents and hydrolysis half-lives.

In a separate experiment, RNA secondary and tertiary interactions can be interrogated by modifying an RNA with chemical probes under denaturing conditions or by incorporating nucleotide substitutions that disrupt native structure. In modification interference, an RNA is treated to introduce chemical modifications, usually at the nucleobases, and then the RNA is subjected to a partitioning experiment to distinguish functional from non-functional molecules (Conway and Wickens 1989; Clarke 1999). For the nucleotide analog interference mapping (NAIM)(Ryder and Strobel 1999; Strobel 1999) strategy, 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 (Conway and Wickens 1989;

Clarke 1999; Ryder and Strobel 1999; Strobel 1999). These approaches generally require multiple distinct experiments to interrogate the tertiary environment of every nucleotide in an RNA.

Here we describe two strategies in which 2'-hydroxyl-selective reagents are used to interrogate higher order structures in RNA. The first strategy, which we term *differential SHAPE*, compares differences in the position specific SHAPE reactivity in order to determine features of higher-order RNA structure from chemical probing experiments. The second strategy, which we call *2'*-hydroxyl molecular interference (HMX), a hydroxyl-selective reagent is used to create a pool of RNAs with evenly distributed *2'-O*-ester adducts. Next, a structure-selective pressure, such as RNA folding, is placed on the pool of modified RNAs. A subset of *2'-O*-ester groups will interfere with molecular interactions and prevent native structure formation. 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.

Results

Fingerprinting RNA structure using multiple SHAPE reagents

We initially screened potential SHAPE reagents for the ability to "fingerprint" RNA tertiary structure motifs using the aptamer domain of the TPP riboswitch in the ligand-bound state. The TPP riboswitch has been extensively characterized by crystallography (Serganov et al. 2006; Haller et al. 2013) and SHAPE chemistry (Steen et al. 2011). This RNA contains many tertiary structure features, especially at or near the ligand binding pocket, that are common to highly structured RNAs including base stacking, long-range docking interactions, and tight turns in the RNA backbone.

Two reagents, N-methylisatoic anhydride (NMIA) and 1-methyl-6-nitroisatoic anhydride (1M6), proved especially promising. NMIA, one of the first reagents used in the SHAPE approach (Merino et al. 2005), reacts slowly with RNA and can be used to identify nucleotides that undergo

local conformational changes on slow timescales (Fig. 2.1B) (Gherghe et al. 2008). These nucleotides are usually in the relatively rare C2′-endo conformation and, in some cases, govern the folding of entire RNA domains (Mortimer and Weeks 2009). The second reagent, 1M6, differs from NMIA by a single nitro ($-NO_2$) group on the double ring system (Fig. 2.1B). This modification changes the chemical behavior of 1M6 in two ways relative to that of NMIA. Addition of the electron-withdrawing group increases the electrophilicity of the reactive center (Fig. 2.1A, red circle), and consequently 1M6 reacts more rapidly than NMIA. Second, the $-NO_2$ group significantly changes the electronic distribution of the reagent ring system which, we will show below, allows 1M6 to stack with RNA nucleobases.

When the folded, ligand-bound TPP riboswitch was allowed to react with NMIA, the observed reactivities agreed with the known structure for the ligand-bound TPP riboswitch (Figs. 2.2A). When this RNA was treated with 1M6, the overall SHAPE reactivity profile was very similar to that for NMIA (Figs. 2.2A). In particular, all base-paired nucleotides were unreactive and many single-stranded nucleotides exhibited similar reactivity towards both reagents. Critically, a few nucleotides exhibited strongly enhanced reactivity towards one of the two reagents (Fig. 2.2A, asterisks). SHAPE chemistry is quantitative; therefore, reagent-specific reactivities can be identified by simply subtracting one profile from another. After excluding nucleotides that participate in crystal contacts or that have poorly-defined electron densities in the previously determined crystal structure (Fig. 2.2B, gray columns), we identified six nucleotides that exhibited statistically significant differential reactivities to NMIA versus 1M6 (Fig. 2.2B).

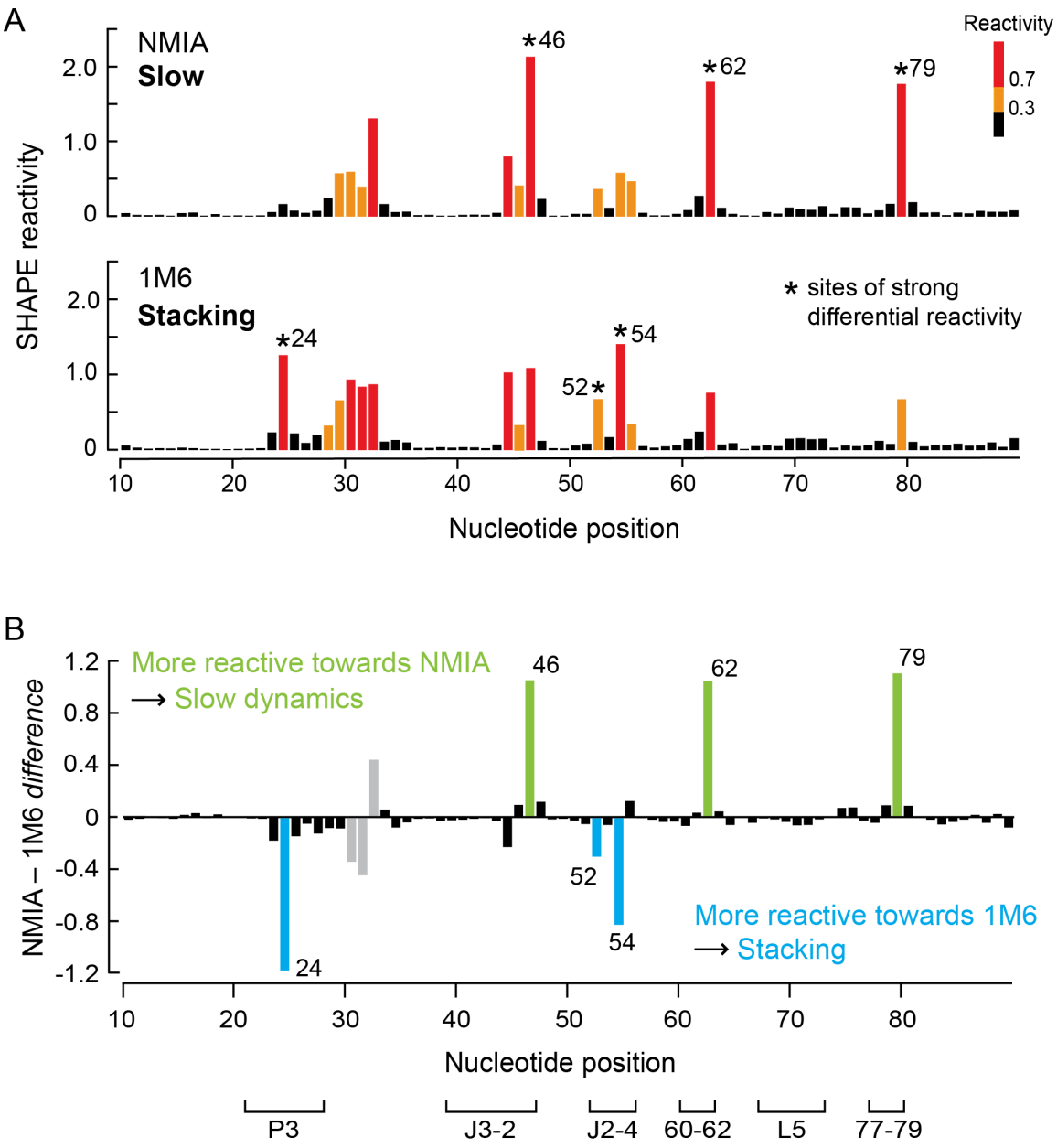

Figure 2.2: SHAPE analysis of the ligand-bound state of the TPP riboswitch. (A) Absolute SHAPE reactivities resulting from reaction with NMIA (top) and 1M6 (bottom). Columns are colored by nucleotide reactivities. Asterisks indicate sites of strong differential reactivity. (B) Differential SHAPE reactivities calculated by subtracting the 1M6 profile from that of NMIA. Columns corresponding to nucleotides that exhibit statistically significant differential reactivity (absolute reactivity difference ≥ 0.3 SHAPE units and a *p*-value < 0.05,calculated using the Student's *t*-test) are colored in green (for NMIA) and blue (for 1M6). Gray columns represent nucleotides with differential reactivity that are involved in crystal contacts or have poorly-defined electron density.

Stacking interactions between reagent and nucleobase influence SHAPE reactivity

Each of the three nucleotides that reacted preferentially with NMIA (Fig. 2.3A) adopts the relatively rare C2′-endo conformation, consistent with previous studies(Gherghe et al. 2008). However, the mechanism by which nucleotides might react preferentially with 1M6 has not been previously explored. The three nucleotides that reacted preferentially with 1M6 are located in diverse local structural environments but share the characteristic that one *face* of the nucleobase is available for π-π stacking interactions with a small molecule like 1M6 (Fig. 2.3B). This conformation is unusual because, both in A-form helices and in most highly folded RNAs, base-base stacking is nearly fully saturated (Leontis et al. 2006; Butcher and Pyle 2011). Only a few nucleotides in special structural contexts – especially at bulges, turns and the termini of some helices – form "one-sided" stacking interactions.

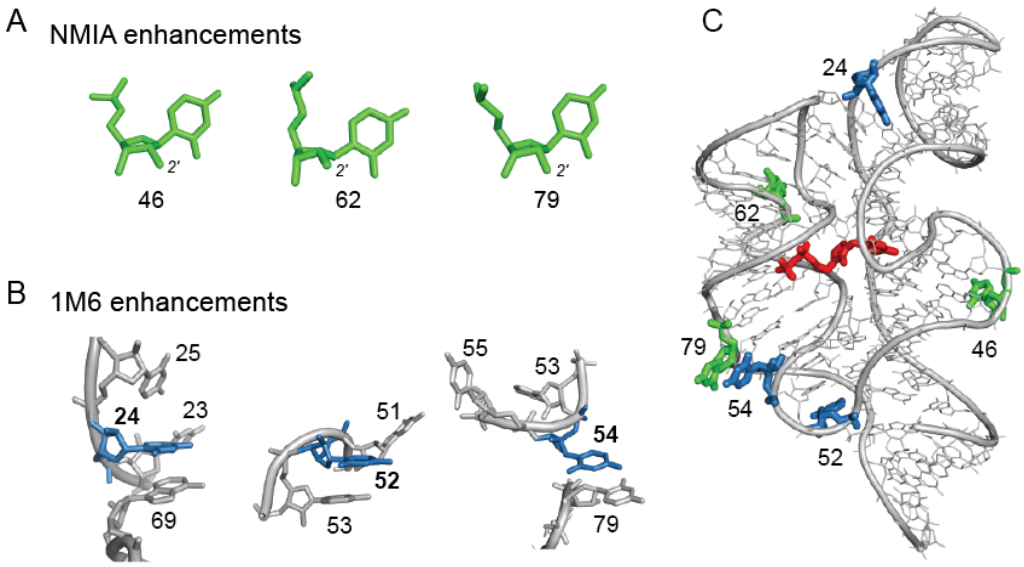

Figure 2.3: Conformations and structural contexts for nucleotides exhibiting differential reactivities in the ligand-bound state of the TPP riboswitch. (A) Sites of enhanced reactivity towards NMIA correspond to nucleotides in the C2′-endo ribose conformation. (B) Sites of 1M6 enhancement reflect one-sided stacking conformations. (C) NMIA (green) and 1M6 (blue) enhancements superimposed on a three-dimensional model for the TPP riboswitch aptamer domain with the bound ligand shown in red (2gdi).

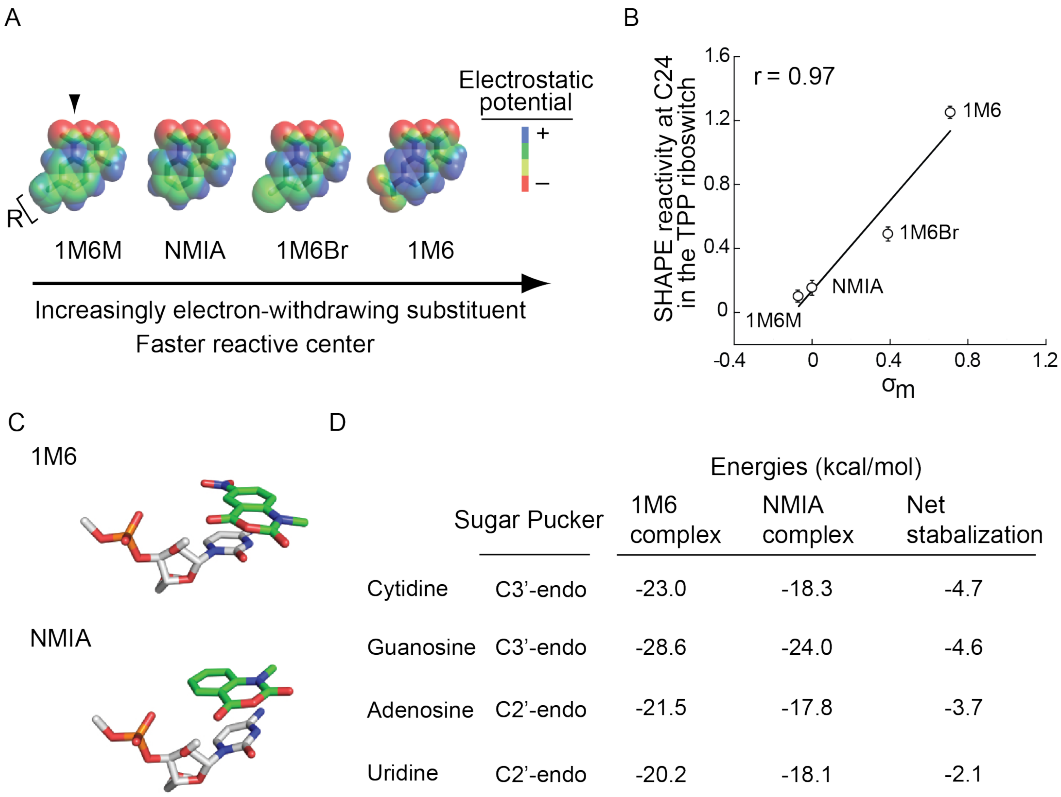

Figure 2.4: Effect of varying the substituents on SHAPE reactivity and electronic structure calculations for the 1M6- and NMIA-nucleotide complex stabilities. Increasing the electron-withdrawing ability of the functional group results in a more electrophilic reactive center and changes in the overall electrostatic profile of the reagent. (A) Electrostatic potential maps for each reagent. (B) Correlation between SHAPE reactivity at C24 (which presents an available open stacking face; see Figure 3B) and A45 (which has pre-existing stacking interactions at both nucleobase faces) in the TPP riboswitch and the electron withdrawing potential as measured by the Hammett coefficient (σ_m) of the reagent R-group. (C) The most stable stacking conformations for the cytidine-1M6 and -NMIA complexes. (D) Representative open-faced stacking complex energies and net stabilization energy as a function of nucleotide, ribose conformation, and reagent.

We hypothesized that the $-NO_2$ substituent polarizes the two-ring system which stabilizes the 1M6-nucleobase stacking interaction. We evaluated this model experimentally by varying the electron-withdrawing ability of the ring functional group of the reagent from a methyl group (slightly electron-donating), to bromine (moderately electron-withdrawing), to a nitro group (strongly electron-withdrawing) (Fig.s 2.1B). The SHAPE reactivities of the "one-sided" stacking nucleotide C24 increased monotonically with increasing electron-withdrawing ability of the reagent substituent as reflected by the Hammett coefficient (Hammett 1937) for each functional group (Pearson's linear *r* =

0.97; Fig. 2.4B). In contrast, this trend was not observed for A45, which is also reactive towards SHAPE reagents but forms stacking interactions on both sides of the adenine base. These reactivity patterns are consistent with the formation of increasingly favorable reagent-nucleobase stacking interactions (Mignon et al. 2005), which are possible at C24 but not A45.

Since the effect of electron-withdrawing groups on the stacking interaction and resulting SHAPE reactivity is quantitative, we estimated representative electronic contributions associated with this interaction from first principles for the four RNA nucleotide types (Fig. 2.4C). Complexes formed between 1M6 and the four RNA nucleotides were -2 to -5 kcal/mol more stable than those formed with NMIA (Fig. 2.4D). These values are significant when compared to the approximate net stabilization energy of a two base pair stack of 2-3 kcal/mol (Petersheim and Turner 1983) and likely reflect the upper limit of favorable interaction. Favorable stacking appears to enhance 1M6 reactivity by increasing the effective local concentration of the reagent at nucleotides where one face is available for the one-sided stacking interaction.

HMX overview

In the first step of the HMX strategy, an RNA of interest is modified with a *2'*-hydroxyl selective reagent under denaturing conditions such that modifications are distributed roughly equally and sparsely among all nucleotides in the RNA population. Second, the RNA is allowed to fold under conditions that favor the native, functional state; and, third, the RNA is subjected to a selection step to partition the RNA into active and inactive components. An experiment with an unmodified control is performed in parallel. The RNAs in this analysis were modified using *N*-methylisatoic anhydride (NMIA) (Merino et al. 2005). NMIA is particularly well-suited for this application because it modifies RNA at high temperatures, to form a simple bulky (but not too bulky) adduct in the RNA backbone. Under denaturing conditions (95 °C at low ion concentrations) NMIA modifies all positions in an RNA at the 2'-hydroxyl position. Some adducts will have no or small structural consequences, whereas other adducts will prevent native folding of the RNA. We partitioned the

natively folded from the unfolded structures based on mobility in non-denaturing acrylamide gels, although many other selection strategies are compatible with this approach. After partitioning folded and unfolded populations, positions of modified nucleotides were detected as stops using reverse transcription- mediated primer extension. Adducts that disrupted folding were identified by comparing the profiles of the unfolded and folded RNA at each position. HMX scores for each nucleotide were calculated as the difference between the normalized profiles for the folded and unfolded RNA.

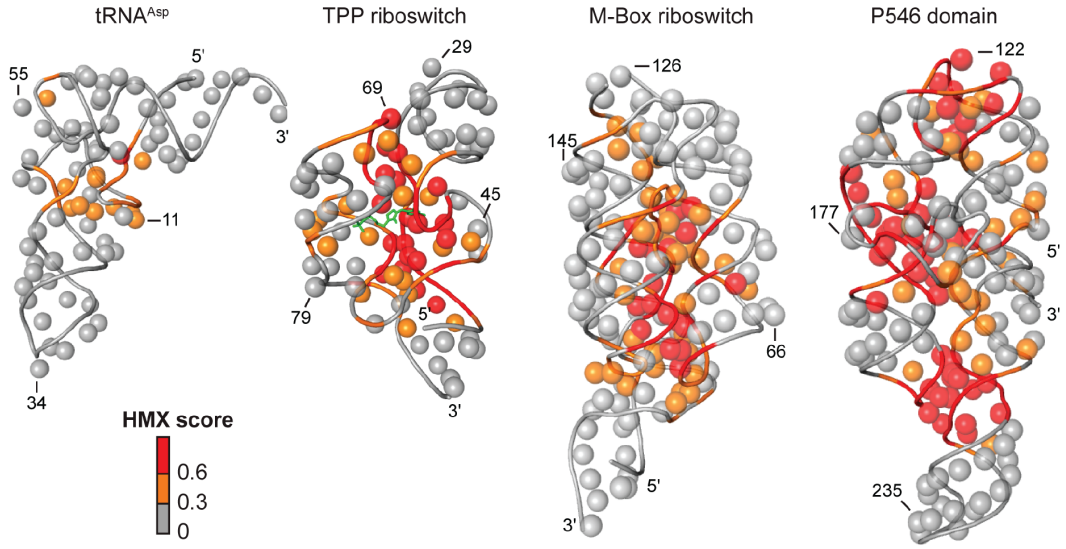

Figure 2.5: Visualization of HMX interference information on accepted three-dimensional structures. The 2'-OH group for each nucleotide is shown as a sphere and the phosphoriboos backbone as a tube. Nucleotides are colored by HMX score; the TPP ligand is colored green.

After partitioning, sites of modification were identified in the folded and unfolded populations by reverse transcription-mediated primer extension. 2'-O-ester adducts that prevent folding were over represented in the unfolded band and underrepresented in the folded band. The resulting modified RNA 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 the separation of unfolded and folded populations using *2'-O*-adduct molecular interference is imperfect and that there is some noise in the separated signals. Positions with medium and high interference scores were visualized on the known three- dimensional structures of each RNA (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 (Lesnik et al. 1993; Lesnik and Freier 1998), interfering positions corresponded almost exclusively to higher-order interactions and not to canonically base-paired nucleotides (Fig. 2.5).

Molecular overlap model for HMX intensities.

Because molecular interference appeared to correlate so strongly with RNA tertiary interactions, we sought to understand the molecular basis of this correlation. To do so we first defined a pseudo-atom, representing the 2'-O-ester adduct, described by two parameters: *L*, the length of the pseudo-atom vector extended from the *2'*-carbon–*2'*-oxygen bond, and *r*, the radius of the pseudo-atom. Using the accepted three- dimensional RNA structures we calculated the degree to which surrounding nucleotide atoms intersected the defined pseudo-atom shell, based on their van der Waals radii (Fig. 2.6A). The pseudo-atom bond length and atomic radius were determined by calculating a correlation coefficient between the simulated and 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).

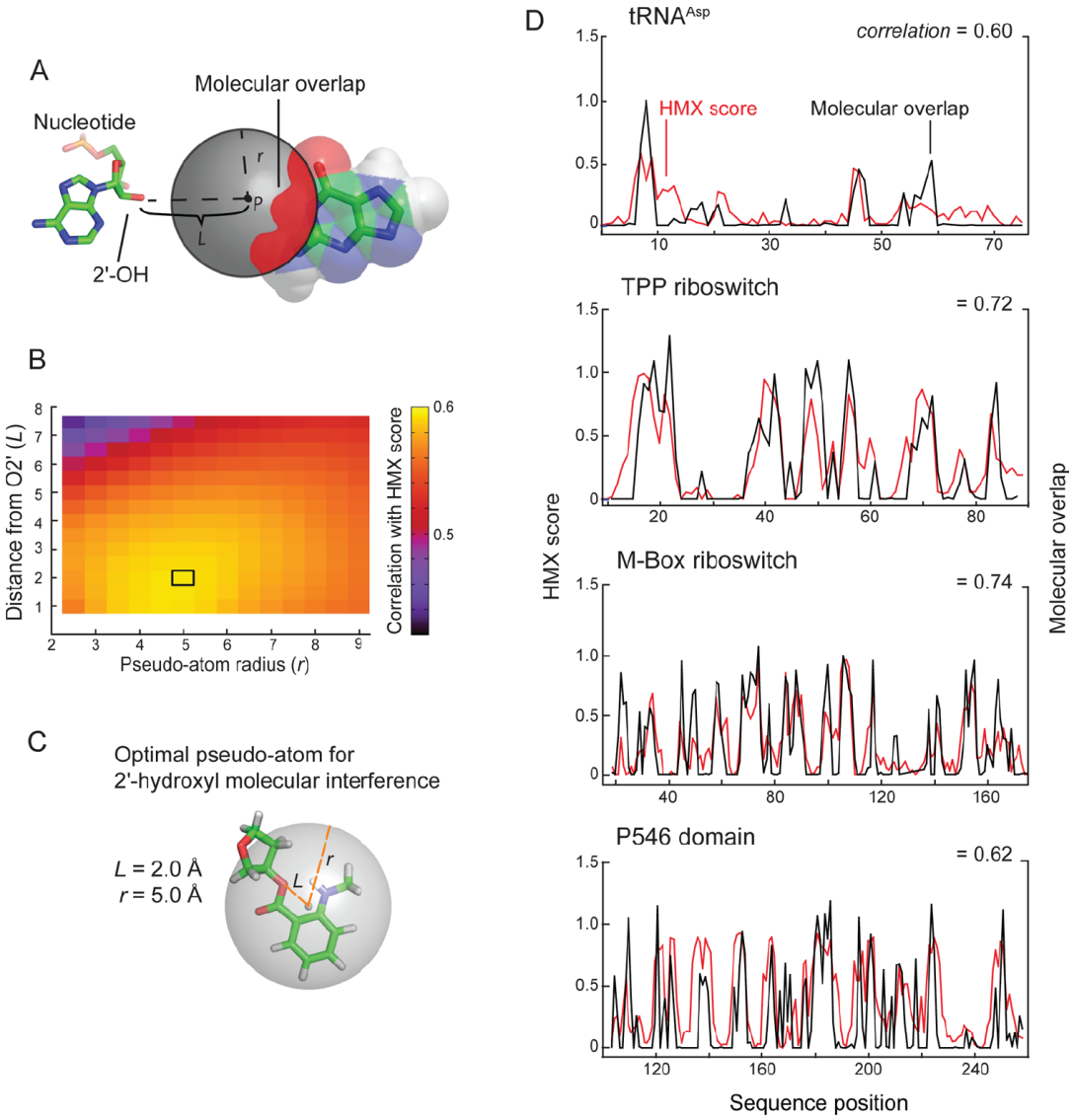

Figure 2.6: Physical model for 2'-hydroxyl molecular interference. (A) Model for interference by molecular overlap in which adducts are represented by a pseudo-atom (grey) at a distance (*L*) from the O2' position at radius (*r*). (B) Analysis of optimal pseudo-atom bond length and atomic radius. Maximum 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) Relationship between HMX scores and molecular overlap for tRNAAsp, the TPP and M-Box riboswitches, 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.

The correlations between the experimental interference scores and the molecular overlap calculations for each RNA are high (Fig. 2.6D), indicating that the *2'-O*-ester adduct disrupts RNA structure by sterically blocking RNA interactions in crowded regions of the RNA. For example, interfering 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. Critically, as judged by visualizing interfering positions in three dimensions (Fig. 2.5) and from molecular overlap analysis (Fig. 2.6D), HMX analysis is exquisitely sensitive to higher-order molecular interactions in RNA, with essentially no detection of false positive interactions.

Discussion

Differential SHAPE outlook

Differential SHAPE analysis uses a dual-reagent detection scheme that takes advantage of two distinct, fundamental features of local RNA structure: (1) some nucleotides are constrained in a special structural environment such that they become reactive on slow timescales and (2) although the vast majority of nucleotides in folded RNAs possess fully "saturated" base stacking interactions, a few nucleotides form unusual conformations that leave one face of the nucleobase accessible for a binding interaction. To a first-order approximation, these two features are selectively detectable by enhanced reactivity towards NMIA and 1M6, respectively. The longer lifetime of NMIA in solution, prior to degradation by hydrolysis, provides a longer window for nucleotides experiencing slow nucleotide dynamics to achieve a SHAPE-reactive conformation {Mortimer:2009gi} (Fig. 2.1). Enhanced reactivity towards 1M6 reflects preferential reagent binding via stacking at accessible nucleobase sites, a model supported by direct visualization in high-resolution structures (Figure

2.3B), the correlation between strength of electron-withdrawing substituent and SHAPE reactivity (Fig. 2.4B), and high-level density functional theory calculations (Figure 2.4D).

HMX experiment and outlook.

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 (Lesnik et al. 1993; Lesnik and Freier 1998). Thus, the *2'-O*-ester molecular interference measurement is exquisitely and specifically sensitive to interactions that govern RNA tertiary folding. For the five 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 closely to the densely packed interior of these structures (Fig. 2.5). This relationship is quantitative. Molecular interference by the 2'-O-ester group was highly correlated with a sphere of defined location relative to the RNA ribose group (Fig. 2.6). We anticipate that *2'-O*-ester mediated molecular interference will prove broadly useful in evaluating higher-order RNA packing in the context of large RNAs and RNA-protein complexes.

HMX is a simple, information-rich, and highly quantitative approach for analysis of the tertiary structure architecture of functionally important RNAs and provides a unique view of internal and closely packed RNA tertiary structure. Here, RNAs were partitioned based on size using gel electrophoresis; however, any strategy that separates functional from non-functional RNAs could be used, allowing HMX analysis to be implemented based on the ability of an RNA to interact with proteins or with other RNAs, or to perform catalysis and other functions.

Methods

1M6 and NMIA free energy calculations

Models of SHAPE reagents and nucleic acids were created separately using the modeling program *Avogadro* (Hanwell et al. 2012). The simplified nucleic acid model system consisting of a nucleoside with a phosphate at the 3' position and an alcohol at the 5' position of the sugar was created for each of the four RNA bases. The charge of the base was neutralized at the phosphate to simplify the gas phase optimization. Nucleotide and SHAPE molecule complexes were based off of modeling of the 1M6 stacking interaction at C24 in the TPP riboswitch. The orientation of the SHAPE molecule was chosen to maximize the overlap of the ring systems and substituent interactions between the molecules, which has been shown prior to enforce the stacking interactions (Florian et al. 1999).

Structures were optimized using the default method implemented in the Gaussian 09 package (Frisch et al. 2009) in gas phase using M06-2X/6-311G*. The M06-2X functional has been shown to be robust enough to model stacking interactions of aromatic systems (Churchill and Wetmore 2011) and consistently recover the CCSD(T) CBS π-π interaction energy (Rutledge and Wetmore 2010). This 6-311G* basis set was chosen for its computational efficiency for a system of this size for optimizing the geometry of the system. Once the models were optimized, a high level single point energy calculation on the optimized structure using M06-2X/6-311+G(2d,p) was used to obtain more accurate structure energies.

Interaction energies for the different SHAPE complexes were calculated using the SCF energy of the structures to get the energy difference at infinite distance allowing for direct comparisons between different complexes:

$$E_{interaction} = E_{complex} - (E_{SHAPE} + E_{Nucleoside})$$

Modeling of adduct disruption of native RNA tertiary structure (HMX).

The *2'-O*-ester adducts were modeled as spheres (Fig. 2.6A). Hydrogen atoms were added using the Molprobity web service (Chen et al. 2010) and the RNA model was extracted from a pdb file. 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. Atoms from the originating and directly adjacent 5' and 3' nucleotides were excluded from the calculation. Clashes between atoms of the RNA and the center of the adduct sphere were assumed to be most disruptive. Thus, points for the Monte Carlo integration were sampled from a normal distribution with σ defined as the radius of the adduct; points are thus concentrated at the center of the adduct sphere. Points falling within the van der Waals radii of atoms in the PDB were scored as hits. Volume integrals converged after sampling 200,000 points at each nucleotide position.

REFERENCES

Butcher SE, Pyle AM. 2011. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc Chem Res* **44**: 1302–1311.

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**: 12–21.

Churchill CDM, Wetmore SD. 2011. Developing a computational model that accurately reproduces the structural features of a dinucleoside monophosphate unit within B-DNA. *Phys Chem Chem Phys* **13**: 16373–16383.

Clarke PA. 1999. RNA footprinting and modification interference analysis. *Methods Mol Biol* **118**: 73–91.

Conway L, Wickens M. 1989. Modification interference analysis of reactions using RNA substrates. *Meth Enzymol* **180**: 369–379.

Florian J, Šponer J, Warshel A. 1999. Thermodynamic parameters for stacking and hydrogen bonding of nucleic acid bases in aqueous solution: ab initio/Langevin dipoles study. *The Journal of Physical Chemistry B* **103**: 884–892.

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, et al. 2009. Gaussian 09.

Gherghe CM, Mortimer SA, Krahn JM, Thompson NL, Weeks KM. 2008. Slow conformational dynamics at C2'-endo nucleotides in RNA. *J Am Chem Soc* **130**: 8884–8885.

Haller A, Altman RB, Soulière MF, Blanchard SC, Micura R. 2013. Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. *Proc Natl Acad Sci* **110**: 4188–4193.

Hammett LP. 1937. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J Am Chem Soc* **59**: 96–103.

Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. 2012. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform* **4**: 17.

Holbrook SR. 2008. Structural principles from large RNAs. *Annu Rev Biophys* **37**: 445–464.

Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16**: 279–287.

Lesnik EA, Freier SM. 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.

Lesnik EA, Guinosso CJ, Kawasaki AM, Sasmor H, Zounes M, Cummins LL, Ecker DJ, Cook PD, Freier SM. 1993. Oligodeoxynucleotides containing 2'-O-modified adenosine: synthesis and effects on stability of DNA: RNA duplexes. *Biochemistry* **32**: 7832–7838.

Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* **127**: 4223–4231.

Mignon P, Loverix S, Steyaert J, Geerlings P. 2005. Influence of the pi-pi interaction on the hydrogen bonding capacity of stacked DNA/RNA bases. *Nucleic Acids Res* **33**: 1779–1789.

Mortimer SA, Weeks KM. 2009. C2'-endo nucleotides as molecular timers suggested by the folding of an RNA domain. *Proc Natl Acad Sci* **106**: 15622–15627.

Petersheim M, Turner DH. 1983. Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry* **22**: 256–263.

Rutledge LRRR, Wetmore SDWD. 2010. The assessment of density functionals for DNA–protein stacked and T-shaped complexes. *http://dxdoiorg/101139/V10-046* **88**: 815–830.

Ryder SP, Strobel SA. 1999. Nucleotide analog interference mapping. *Methods* **18**: 38–50.

Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. 2006. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**: 1167–1171.

Sharp PA. 2009. The centrality of RNA. *Cell* **136**: 577–580.

Sokoloski JE, Godfrey SA, Dombrowski SE, Bevilacqua PC. 2011. Prevalence of syn nucleobases in the active sites of functional RNAs. *RNA* **17**: 1775–1787.

Steen K-A, Siegfried NA, Weeks KM. 2011. Selective 2'-hydroxyl acylation analyzed by protection from exoribonuclease (RNase-detected SHAPE) for direct analysis of covalent adducts and of nucleotide flexibility in RNA. *Nature Protocols* **6**: 1683–1694.

Strobel SA. 1999. A chemogenetic approach to RNA function/structure analysis. *Curr Opin Struct Biol* **9**: 346–352.

Weeks KM. 2010. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol* **20**: 295–304.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1**: 1610–1616.

CHAPTER 3: RNA SECONDARY STRUCTURE MODELING AT CONSISTENT HIGH ACCURACY USING DIFFERENTIAL SHAPE[2]

Introduction

RNA is a central information carrier in biology (Sharp 2009). Information is encoded in RNA at two distinct levels: in its primary sequence and in its ability to fold into higher order structures (Leontis et al. 2006; Dethoff et al. 2012). The most fundamental level of higher order structure is the pattern of base pairing or secondary structure. Defining the secondary structure of an RNA is also a critical first step in tertiary structure modeling (Hajdin et al. 2010; Weeks 2010; Bailor et al. 2011). The structures of RNA molecules modulate the numerous functions of RNA and the interactions of RNAs with proteins, small molecules, and other RNAs in splicing, translation, and other regulatory machineries (Mauger et al. 2013).

Accurate, de novo modeling of RNA secondary structure is challenging: In the absence of experimental restraints, current algorithms predict base pairing patterns that contain, on average, 50-70% of the canonical (G-C, A-U, and G-U) pairs in secondary structures established through phylogenetic analysis or high-resolution experimental methods (Mathews et al. 2004; Hajdin et al. 2013). The modeling challenge results from the fact that there are only four RNA nucleotides and that these nucleotides have the potential to arrange into many, often energetically similar, RNA secondary structures, even though many RNAs adopt a few or only single structures (Tinoco and Bustamante 1999). Features that are difficult to extract solely from the sequence – such as kinetic pathways, protein facilitators, and ligand binding – also influence RNA folding. Identification of the correct RNA secondary structure also becomes much more difficult as the length of the RNA increases.

[2] This chapter previously appeared as an article in *RNA*. The original citation is as follows: Rice, G. M., Leonard, C. W., & Weeks, K. M. (2014). RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA*, *20*(6), 846–854. doi:10.1261/rna.043323.113

Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) reagents can be used to interrogate the flexibility of nearly every nucleotide in an RNA (Merino et al. 2005; McGinnis et al. 2012). Reactivity at the 2'-hydroxyl toward the reagent 1-methyl-7-nitroisatoic anhydride (1M7) measures local nucleotide flexibility. Because base-paired nucleotides are also structurally constrained, SHAPE reactivity is roughly inversely proportional to the probability that a nucleotide is paired. Incorporation of SHAPE reactivity information into RNA folding algorithms results in accuracies above 90% for most RNAs including those with pseudoknots (Deigan et al. 2009; Hajdin et al. 2013). SHAPE has been used to create nucleotide-resolution models for the viral genomes of HIV-1 (Watts et al. 2009) and STMV (Archer et al. 2013) and to analyze conformational changes in HIV-1 (Wilkinson et al. 2008) and the Moloney murine leukemia virus (Grohman et al. 2013). Although SHAPE-directed folding yields near-perfect models for many RNAs, there remain a few RNAs whose structures are difficult to recover using a single structure probing experiment (Cordero et al. 2012; Leonard et al. 2013). These "hard" RNAs are modeled with sensitivities in the 75-85% range.

The usefulness of secondary structure models at different accuracies can be summarized on a multi-point scale (Fig. 3.1), analogous to those used in other fields (Munroe 2012). Models with prediction sensitivities below 60% contain large errors in gross structure and are not generally useful for generating biological hypotheses. Computational-only algorithms achieve median prediction accuracies of about 70%. An individual model that recovers 70% of the accepted base pairs will have some correct helices and also critical errors (Fig. 3.1, second structure from bottom). Although approaches that recover 70% of the accepted base pairs include both correct and incorrect pairs, it is generally difficult to determine which helices are correct and which are not. Using SHAPE-directed modeling, the predicted structures for the most challenging RNAs contain 80-85% of accepted base pairs. In some cases, the incorrectly predicted base pairs are scattered throughout the RNA such that the overall model is quite good. In other cases, errors are located in structural elements that are known to be functionally important (Fig. 3.1, middle structure).

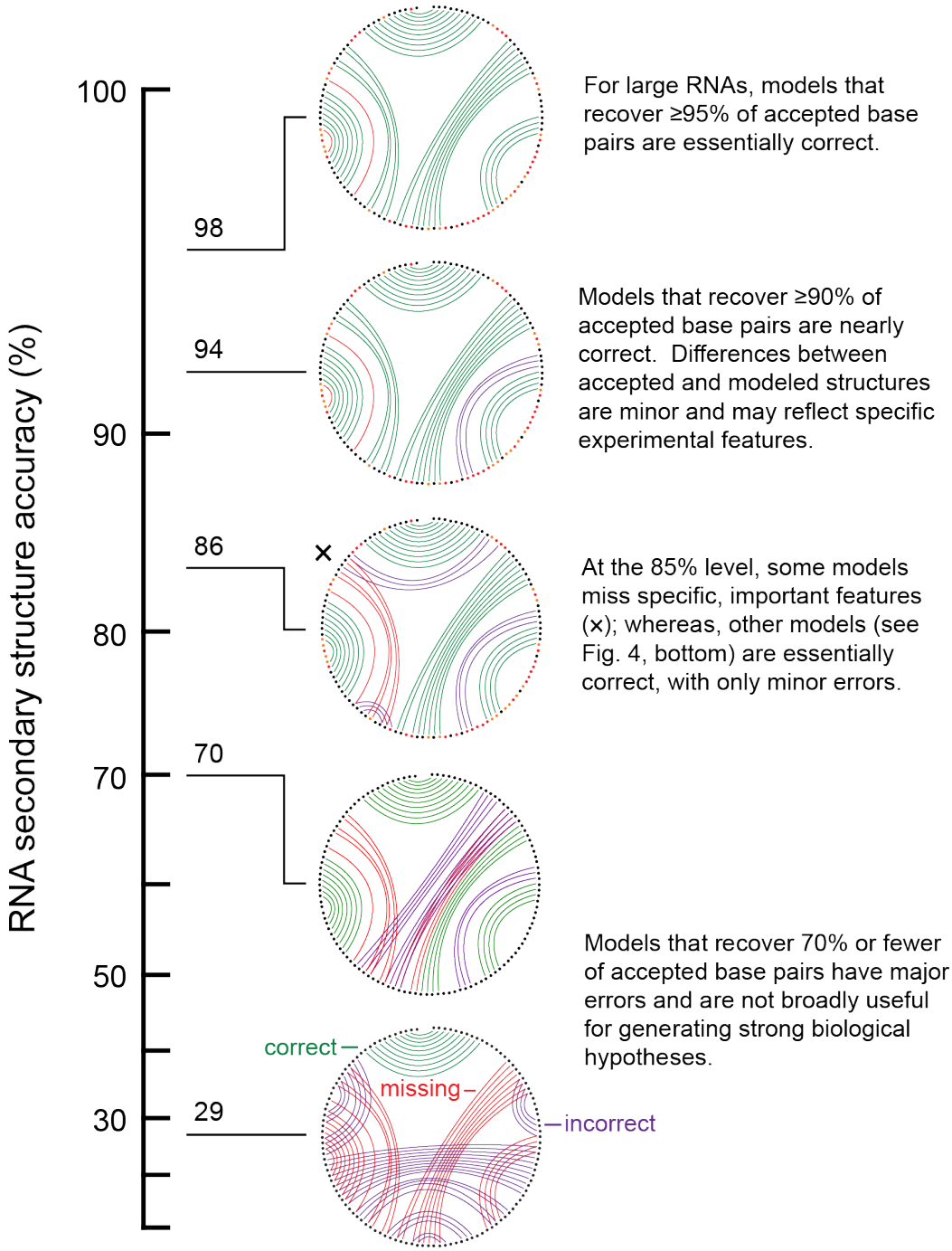

Figure 3.1: Accuracy of an RNA structure model and its usefulness for understanding structure-function interrelationships. Representative structures for the E. coli 5S rRNA are shown. Accuracy is represented as the sensitivity and plotted on a reverse-logarithmic scale to emphasize the increasing level of difficulty as the standard for recovery of accepted base pairs increases. For all secondary structure (circle plot) diagrams, correct base pairs are shown in green, missing base pairs are shown in red, and extra base pairs relative to the accepted structure are shown in purple.

On average, SHAPE-directed modeling currently recovers approximately 93% of accepted base pairs in challenging sets of RNA molecules. This level of sensitivity is sufficient for generation of robust biological hypotheses and for three-dimensional structure modeling. Many of the models generated at this level of accuracy differ from the accepted models by a few base pairs and should be considered nearly perfect (Fig. 3.1, upper structures). Improving accuracies to above the 90% level for all RNAs is the current challenge in experimentally-directed secondary structure modeling. Inclusion of additional comprehensive and information-rich biochemical information could further inform, and potentially solve, the RNA secondary structure modeling problem.

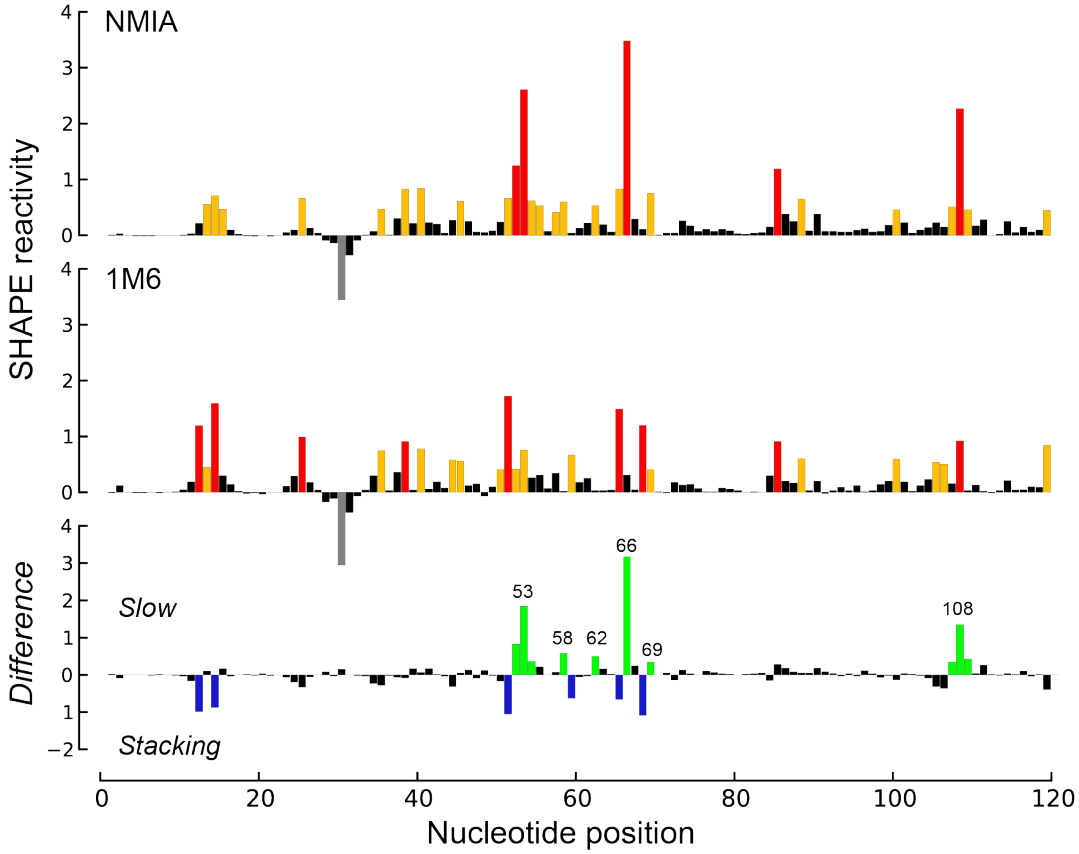

Figure 3.2: Differential SHAPE analysis of the E. coli 5S rRNA. Normalized SHAPE reactivities from reactions with NMIA (top) and 1M6 (middle) are colored by nucleotide reactivity. Differential SHAPE reactivities (Steen et al. 2012) (bottom) were calculated by first scaling 1M6 to NMIA reactivities over a moving window and then subtracting 1M6 from NMIA reactivities. Strong differential reactivity enhancements (>|0.3| SHAPE-units) are colored green for NMIA and blue for 1M6. These sites correspond to nucleotides with slow dynamics and those with a face available for stacking, respectively. Nucleotide positions showing strong positive-amplitude (favoring NMIA) differential reactivities are labeled.

We recently described an approach that we call *differential* SHAPE that reveals local non-canonical and tertiary structure interactions based on simple biochemical probing experiments (Steen et al. 2012). In this strategy, the position-specific reactivities of two reagents, N-methylisatoic anhydride (NMIA) and 1-methyl-6-nitroisatoic anhydride (1M6), are compared. The first reagent, NMIA, has a relatively long half-life in solution and reacts preferentially with nucleotides that experience slow dynamics. Often these nucleotides are in the rare C2'-endo ribose conformation and have been implicated as molecular timers capable of governing folding in large RNAs (Gherghe et al. 2008; Mortimer and Weeks 2009). For the second reagent, the nitro group of 1M6 makes the two-ring system electron-poor, and this reagent is able to stack with RNA nucleobases that are not protected by interactions with other nucleotides in an RNA structure (Steen et al. 2012). This conformation is unusual since most nucleobases stack with other bases on both faces (Leontis et al. 2006). By taking the difference in reactivity profiles for these two 2'-hydroxyl selective reagents, nucleotides involved in structurally distinctive interactions within an RNA structure can be identified (Fig. 3.2). Because the differential SHAPE analysis is specifically sensitive to non-canonical and tertiary interactions in RNA (Steen et al. 2012), this approach can help to identify nucleotides that are constrained (and thus unreactive to 1M7-SHAPE) but do not participate in canonical base pairing. Here we develop a pseudo-free energy term that includes information from the slow and stacking differential SHAPE reactivities to yield nearly perfect secondary structure models in a concise experiment that scales to RNAs of any size.

Results

Selection of a challenging test set.

To evaluate the utility of incorporating differential SHAPE data into a modeling algorithm, we chose a set of diverse RNAs with well-established secondary structures for which single-reagent SHAPE-directed secondary structure prediction remains challenging (Table 3.1). These included six

riboswitch aptamer domains that require ligand binding to fold into their correct structures (the TPP, adenine, glycine, cyclic-di-GMP, M-Box, and lysine riboswitches); four RNAs longer than 300 nucleotides, including several domains of the E. coli 16S and 23S ribosomal RNAs; four pseudoknot-containing RNAs; and every other RNA of which we are aware that contains up to one pseudoknot for which the single-reagent 1M7 prediction accuracy is less than 90% (Table 3.1) (Cordero et al. 2012; Leonard et al. 2013; Hajdin et al. 2013).

		SHAPE	–		+		+		
		differential	–		–		***NMIA–1M6***		
	length (nts)		sens	ppv	sens	ppv	sens	ppv	
TPP riboswitch, *E. coli*	79		77.3	85.0	96.5	91.3	95.5	100.0	responsive
cyclic-di-GMP riboswitch, *V. cholerae*	97		75.0	77.8	89.2	86.2	96.4	93.1	
5S rRNA, *E. coli*	120		28.6	25.0	85.7	76.9	94.3	91.7	
Glycine riboswitch, *F. nuleatum*	158		70.0	60.9	55.0	48.9	95.0	95.0	
Domain III of 23S rRNA, *E. coli*	372		46.9	43.1	82.7	74.3	90.8	83.2	
Group I intron, *T. thermophila*	425		83.3	75.0	93.2	91.2	84.9	89.7	
3' domain of 16S rRNA, *E. coli*	478		26.7	21.2	89.5	77.6	97.1	86.1	
Average			**58.3**	**55.4**	**84.5**	**78.1**	**93.4**	**91.2**	
Adenine riboswitch, *V. vulnificus*	71		100.0	100.0	100.0	100.0	100.0	100.0	non-responsive
tRNA phe, *E. coli*	76		100.0	91.3	100.0	75.0	100.0	77.8	
M-Box riboswitch, *B. subtilis*	154		87.5	91.3	83.3	90.9	83.3	93.0	
Lysine riboswitch, *T. maritime*	174		75.8	84.8	84.9	90.3	84.9	90.3	
Group II Intron, *O. iheyensis*	412		88.0	97.5	93.2	96.9	92.5	98.4	
5' domain of 16S rRNA, *E. coli*	530		61.3	57.9	97.8	91.8	97.8	91.8	
Domain II of 23S rRNA, *E. coli*	685		87.6	78.6	97.8	87.4	96.8	88.2	
Average			**85.7**	**85.9**	**93.9**	**90.3**	**93.6**	**91.4**	
Overall Average			**72.0**	**70.7**	**89.2**	**84.2**	**93.5**	**91.3**	

Table 3.1: RNA secondary structure modeling accuracies with 1M7 and differential SHAPE information. All well-folded RNAs containing up to one pseudoknot, of which we are aware, for which single-reagent 1M7-restrained secondary structure prediction results in less than 90% sensitivity are included in this table. RNAs are listed based on whether or not modeling is responsive to differential reactivity information: (top) predictions that improve and (bottom) predictions that show small or no changes. RNAs were judged to be responsive to differential SHAPE data if either the sens or ppv changed by at least 3%. Averages were calculated separately for each class and for all RNAs together.

Incorporation of differential SHAPE into secondary structure modeling.

SHAPE experiments were performed with 1M7, NMIA, and 1M6 on RNAs pre-incubated in the presence of cognate ligand if appropriate but without protein. Based on pilot work on three short RNAs, SHAPE reactivity signals from NMIA and 1M6 correlate strongly at most positions (Steen et al. 2012). We therefore used a windowed scaling algorithm to locally normalize NMIA and 1M6 SHAPE profiles to each other (see Methods) and then subtracted the normalized profiles to generate differential SHAPE reactivity traces (Fig. 3.2).

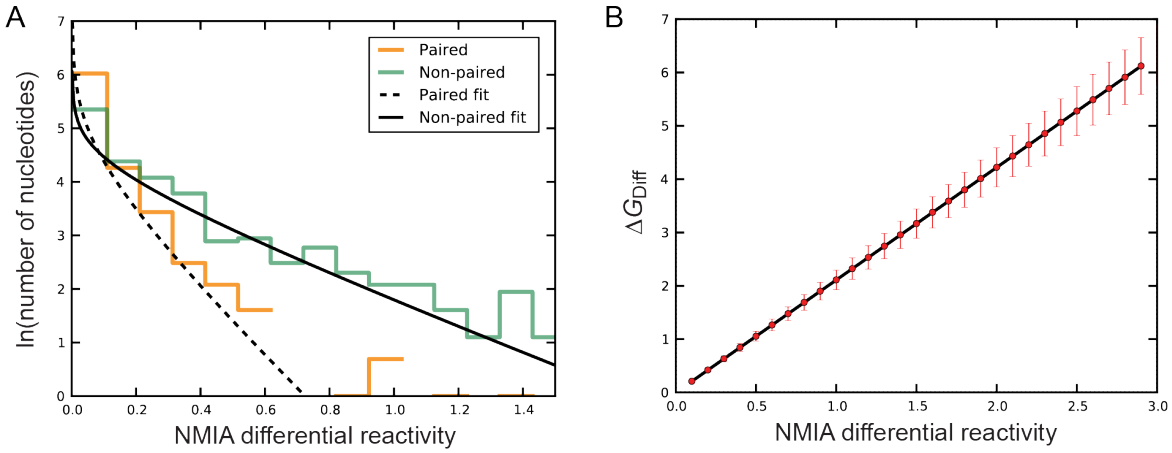

Figure 3.3: Statistical determination of the ΔGDiff free energy change penalty. (A) Differential reactivities were binned as a function of base pairing status in the accepted structure. Paired and non-paired nucleotides were each fit to a gamma distribution. (B) Final ΔG_{diff} energy function calculated from a linear fit of the Gibbs free energy derived from the ratio of paired and non-paired histogram fits. Error bars (red lines) show the standard error of fitting using a leave-one-out jackknife analysis.

We used a statistical potential approach (Rohl et al. 2004; Cordero et al. 2012) to evaluate the differential SHAPE signals. This approach infers a free energy from the difference in the distributions of paired and non-paired nucleotides. The energy function was linear and proved robust when subjected to a leave-one-out jackknife analysis (Fig. 3.3). During fitting, we evaluated both positive and negative differential signals from NMIA and 1M6 (Fig. 3.2, bottom panel; green and blue bars, respectively). The negative-amplitude signal from 1M6 was not as highly correlated with single-stranded character at the sites of differential reactivity as was the positive-amplitude signal. The differential reactivity pseudo-free energy change term for each nucleotide was taken as:

$$\Delta G_{\text{Diff}} = d \times (\text{positive amplitude differential signal}) \quad (1)$$

where d is 2.11 kcal/mol. This energy penalty was added to the standard 1M7-based pseudo-free energy as implemented in ShapeKnots (Low and Weeks 2010; Hajdin et al. 2013); inclusion of this penalty improved predictions for many RNAs. For each RNA model, we report the accuracy of a secondary structure prediction in terms of it sensitivity (sens, fraction of base pairs in the accepted structure predicted correctly) and positive predictive value (ppv, the fraction of predicted pairs that occur in the accepted structure).

Impact of ΔG_{Diff} on structure modeling.

In the absence of experimental restraints, the mfold algorithm predicts only 10 of the 35 base pairs (29%) in the accepted structure of the *E. coli* 5S rRNA (Fig. 3.4, left structure). Addition of 1M7-SHAPE constraints yielded a substantial improvement: 86% of the accepted base pairs were present in the SHAPE-directed model. As is common for predictions at this level of accuracy, most of the structure is modeled correctly. The exceptions are base pairs in one element, a helix at a three-way junction (Fig. 3.4, middle structure, positions 102-107). When differential SHAPE data were added as constraints, a substantially improved structural model was obtained (Fig. 3.4, right structure). The errors in the differential SHAPE-based model are minor and involve the addition of a few base pairs in the second helix of the structure near nucleotide 30. These base pairs may in fact form under our probing conditions, given that this RNA was probed in the absence of ribosomal subunits and proteins.

Addition of differential SHAPE information also improved the accuracy of prediction of the glycine riboswitch structure (Fig. 3.5, top). With data from 1M7 only, the predicted model for the glycine riboswitch had 55% sens and 49% ppv. The major error in the model is the prediction of a false pseudoknot that then propagates other errors (Fig. 3.5, top, left-hand structure). Inclusion of the differential SHAPE penalty resulted in sens and ppv of 95%. In this case, use of the differential reactivity penalty corrected major errors (for example, the differential reactivities at positions 12-13

and 112) and eliminated the false positive pseudoknot. In addition, lower magnitude differential reactivities shifted the folding landscape of nucleotides 39-49 to result in agreement of the predicted and accepted structures.

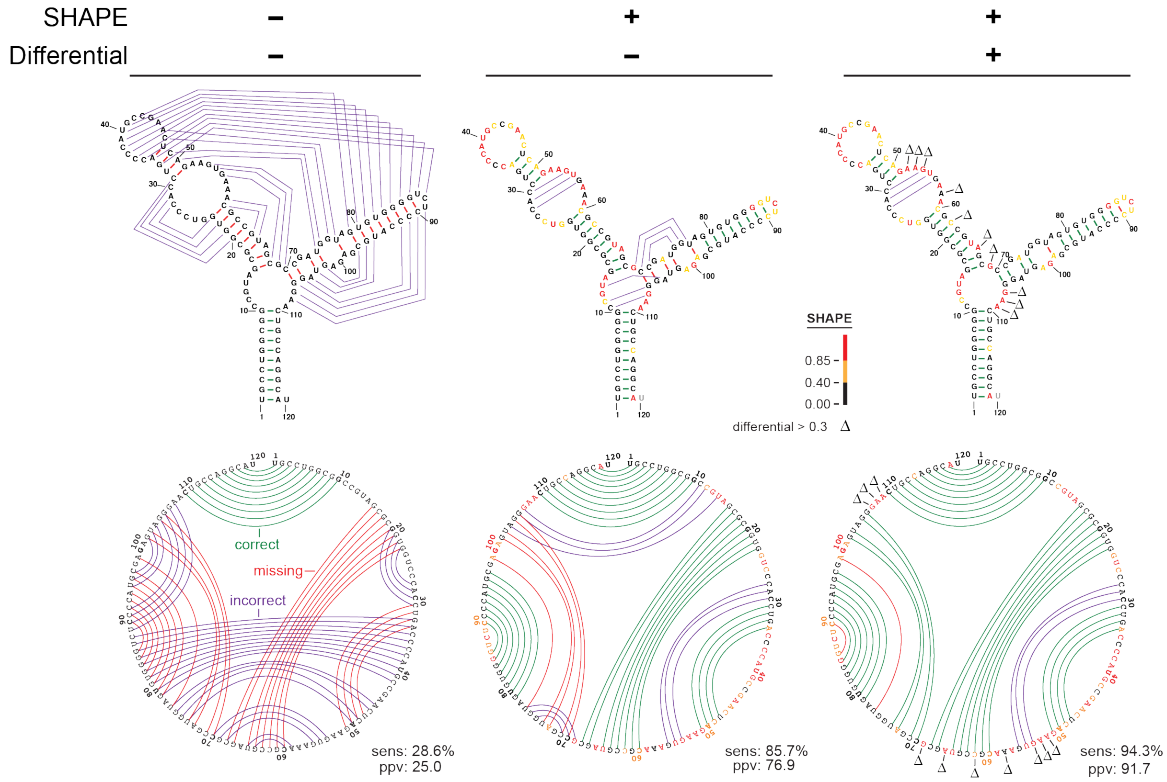

Figure 3.4: Representative secondary structure modeling for the 5S rRNA without and with SHAPE data. Base pair predictions are illustrated with colored lines (green, purple, and red denoting correct, incorrect, and missing base pairs, respectively) on conventional secondary structure representations (top) and circle plots (bottom). Nucleotides are colored according to their SHAPE reactivity on a black, yellow, red scale for low, medium, and strong reactivity. Nucleotides showing strong preferential reactivity with NMIA (>0.3 units) are indicated with a delta symbol.

The predicted structure of the M-Box riboswitch, at 83% sensitivity (Table 3.1), was formally the lowest quality model in the test set. Differential reactivity constraints improved the prediction by a single base pair relative to the structure predicted using 1M7 data only (Fig. 3.5, bottom). The overall topology of the M-Box RNA is largely correct regardless of the inclusion of differential SHAPE information: The three-helix junction and all major helices are predicted correctly. The largest difference between the modeled and accepted structures occurs at the P1 helix connecting the

5' and 3' ends of the RNA (Fig. 3.5, bottom left). Nucleotides in this helix are moderately reactive toward SHAPE reagents, suggesting that the P1 helix is not especially stable under the conditions used for structure probing. In the crystal structure that is the basis for the accepted model, the P1 helix is stabilized by three G-C base pairs (Dann et al. 2007) that were not present in the transcript analyzed by SHAPE. SHAPE data suggest that the native sequence P1 helix is conformationally dynamic. For the sequence of RNA probed in this work, the SHAPE-constrained structure is essentially correct.

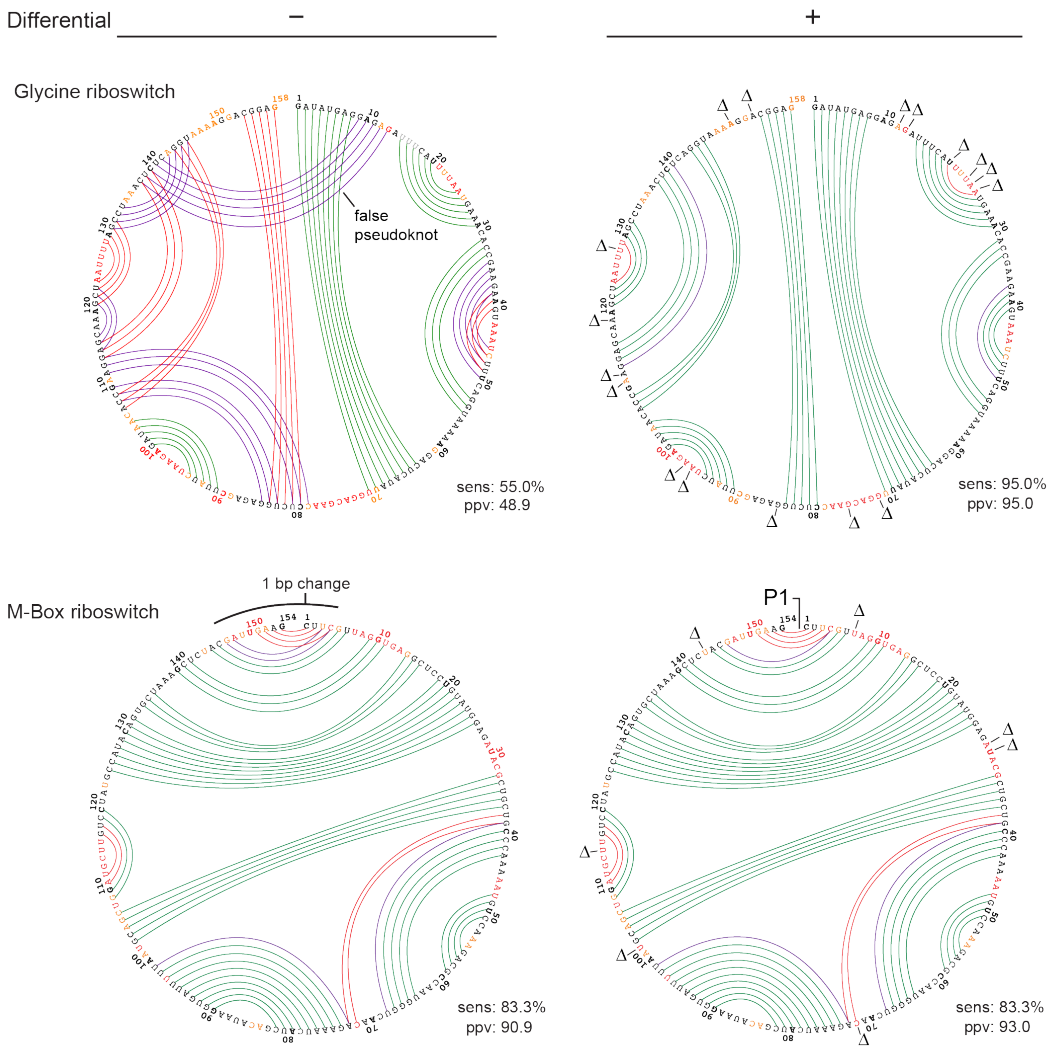

Figure 3.5: Circle plots illustrating SHAPE-directed structure modeling. Glycine (top) and M-Box (bottom) riboswitches with 1M7 SHAPE data (left) and with 1M7 and differential reactivity data (right). Scheme for illustrating base-pair accuracy (relative to crystallographic structures) and nucleotide SHAPE reactivities are as outlined in Figure 3.3; positions with positive-amplitude (favoring NMIA) differential reactivities are indicated with a delta symbol.

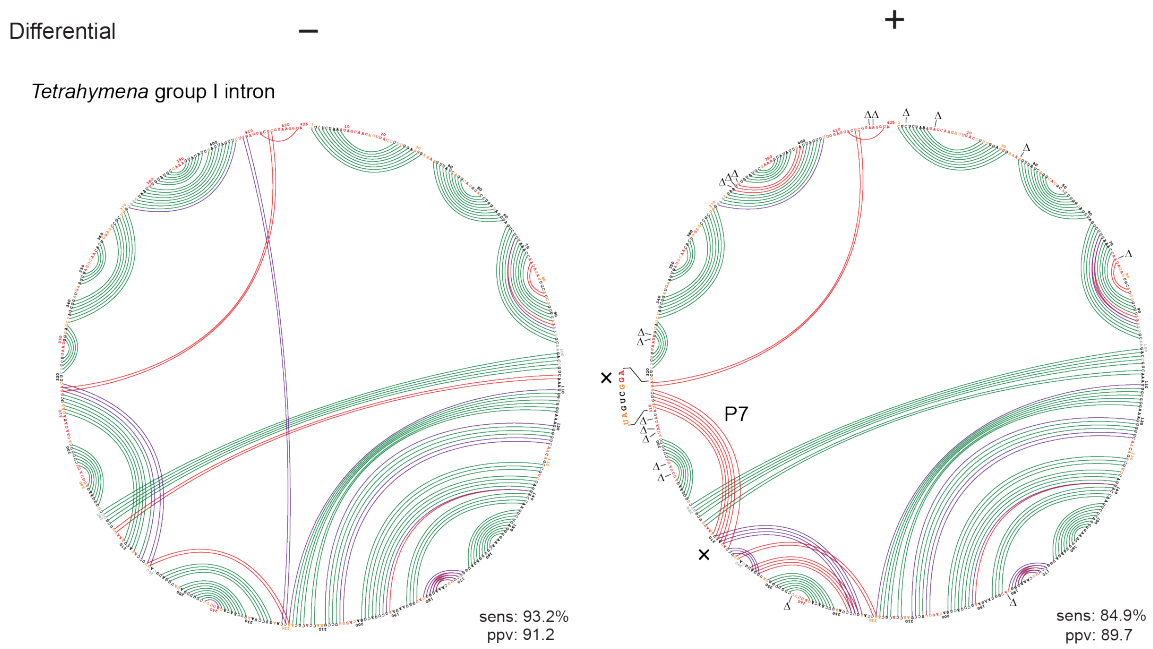

Figure 3.6: Circle plots illustrating SHAPE-directed structure modeling for Tetrahymena group I intron. 1M7 SHAPE data (left) and with 1M7 and differential reactivity data (right). Reactive nucleotides in the P7 helix are shown in an expanded view (right); × symbols indicate structurally significant mis-predictions relative to the accepted structure. Scheme for illustrating base-pair accuracy (relative to crystallographic structures) and nucleotide SHAPE reactivities are as outlined in Figure 3.3; positions with positive-amplitude differential reactivities (favoring NMIA) are indicated with a delta symbol.

Responsive and non-responsive RNAs.

For the RNAs in our test set, predictions either significantly improved with the addition of differential SHAPE data or were only modestly affected. We define structural improvement as significant if the sensitivity or ppv or both increased by at least 3%. Seven RNAs in our dataset showed significant improvement by this criterion (Table 3.1, top; responsive RNAs). The predicted structures for these RNAs increased in sensitivity from an average 84.5% to an average of 93.4%. Improvement in positive predictive value (ppv) was even more substantial: from 78.1% to 91.2%. Of the RNAs in the less responsive category, four of the eight showed small improvements in sensitivity or ppv (Table 3.1, middle), and the changes in the lowest free-energy structure involved relatively minor adjustments in base pairing relative to structures predicted using 1M7 data only. Notably, although predictions for multiple RNAs were improved by the addition of differential SHAPE

restraints, none of the predictions became substantially worse with the exception of the *Tetrahymena* group I intron (Table 3.1).

The modeled structure for the *Tetrahymena* group I intron became less like the accepted structure upon inclusion of differential reactivity information: The sensitivity decreased from 93% to 85% (Table 3.1 and Fig. 3.6). The P7 helix comprises a pseudoknot in the accepted RNA structure. One strand of the P7 helix is reactive by SHAPE and is not present in the SHAPE-directed model (Fig. 3.6). Our data suggest that the P7 helix is conformationally dynamic under the solution probing conditions used in this work.

Discussion

Developing accurate secondary structure models of long RNAs is an absolute prerequisite for understanding the role of RNA structure and RNA-ligand interactions in most phases of gene regulation (Mauger et al. 2013). Moreover, an accurate secondary structure model is critical for and can dramatically facilitate tertiary structure modeling (Hajdin et al. 2010; Bailor et al. 2011). The ideal approach for RNA structure modeling should balance high accuracy with concise and scalable experimentation. The nearest-neighbor thermodynamic model developed by Turner and colleagues (Mathews and Turner 2006) provides a critical foundation for secondary structure modeling. However, there are features of RNA folding that are difficult to extract from sequence including ligand and protein binding effects, non-canonical and long-range tertiary structure interactions, and the kinetic history of the RNA folding reaction. Inclusion of single-reagent experimental structure probing data provides a substantial improvement in modeling accuracy for many RNAs (Deigan et al. 2009; Hajdin et al. 2013), but this improvement was not enough to yield accurate secondary structure models for all RNAs in our test set (Figs. 3.4 and 3.5). Here we demonstrated that inclusion of information from a differential SHAPE experiment substantially increases the sensitivity and positive predictive value of secondary structure models for an RNA test set designed to be as challenging as possible (Table 3.1). The consistent, monotonic, trend in accuracy improvement observed suggests

that each set of restraints – nearest neighbor parameters, 1M7-SHAPE, and differential SHAPE – provides information that is orthogonal to the others, roughly corresponding to local secondary structure, non-nearest neighbor interactions, and non-canonical and tertiary interactions, respectively.

Source		This work*		Cordero *et al.* 2012†				Hajdin *et al.* 2013*	
Method		Three-reagent		NMIA		NMIA + DMS		1M7	
	length (nts)	sens	ppv	sens	ppv	sens	ppv	sens	ppv
Adenine riboswitch, *V. vulnificus*	71	100.0	100.0	100.0	91.3	100.0	91.3	100.0	100.0
tRNA phe, *E. coli*	76	100.0	77.8	100.0	95.5	100.0	95.5	100.0	84.0
TPP riboswitch, *E. coli*	79	95.5	100.0	--	--	--	--	95.5	87.5
cyclic-di-GMP riboswitch, *V. cholerae*	97	96.4	93.1	96.2	92.5	96.2	92.5	89.3	86.2
5S rRNA, *E. coli*	120	94.3	91.7	85.3	76.3	85.3	76.3	85.3	76.3
M-Box riboswitch, *B. subtilis*	154	83.3	93.0	--	--	--	--	87.5	91.3
Glycine riboswitch, *F. nuleatum*	158	95.0	95.0	94.9	86.0	97.5	92.8	--	--
Lysine riboswitch, *T. maritime*	174	84.9	90.3	--	--	--	--	87.3	88.7
Domain III of 23S rRNA, *E. coli*	372	90.8	83.2	--	--	--	--	--	--
Group II Intron, *O. iheyensis*	412	92.5	98.4	--	--	--	--	93.2	97.6
Group I Intron, *T. thermophila*	425	84.9	89.7	--	--	--	--	93.9	91.2
3' domain of 16S rRNA, *E. coli*	478	97.1	86.1	--	--	--	--	--	--
5' domain of 16S rRNA, *E. coli*	530	97.8	91.8	--	--	--	--	--	--
Domain II of 23S rRNA, *E. coli*	685	96.8	88.2	--	--	--	--	--	--

Table 3.2: RNA secondary structure modeling accuracies comparing three-reagent differential SHAPE to related recent works. Approaches that allow pseudoknots are indicated with an asterisk. Methods that used parameters optimized using small datasets are indicated with a dagger.

The information content of three-reagent SHAPE-directed RNA structure modeling appears to exceed that of previously described chemical probing approaches. Addition of dimethyl sulfate (DMS) and *N*-cyclohexyl-*N*'-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate (CMCT) reactivity information in the context of a dataset of six small RNAs yielded improvement of roughly three base pairs in one RNA (Kladwang et al. 2011b; Cordero et al. 2012) (Table 3.2). In contrast, the differential SHAPE experiment yielded large, structurally significant improvements in seven RNAs (Table 3.1, top) and less dramatic improvements in four other RNAs (Table 3.1, middle) over and above single-reagent 1M7-directed modeling. Large improvement was observed for the 5S rRNA, which was not improved with addition of DMS and CMCT data (Cordero et al. 2012). In addition, models developed using three-reagent SHAPE probing have prediction accuracies that equal or exceed that of approaches that involve probing of comprehensive sets of mutants (Kladwang et al.

2011a). The differential SHAPE data thus have high information content that is obtained in a concise experiment that scales easily to large RNAs.

Using differential SHAPE for RNA secondary structure prediction represents a significant advance in RNA structure modeling. With differential SHAPE information, the structures of some of the RNA molecules that were previously viewed as the most challenging, including the 5S rRNA, the glycine riboswitch, and some ribosomal domains, were modeled in nearly perfect agreement with the accepted structures (Table 3.1). An intriguing trend was that the RNAs that were most responsive to the differential reactivity penalty were those with structures predicted most poorly in the absence of differential SHAPE experimental. We speculate that RNAs in this class have non-canonical interactions that are incompletely described by the nearest neighbor algorithm or single-reagent data. SHAPE-driven predicted structures that disagree with the accepted structures – those of the M-Box and lysine riboswitches and the *Tetrahymena* group I intron – "errors" appear to reflect differences between in-crystal and in-solution conformations for these RNAs.

Limitations and Perspective

There are limitations to the experimentally-based RNA structure probing approach outlined here. By far, the most important of these is the restriction of having only a small database of RNAs with well-defined accepted structures (Rivas et al. 2012; Leonard et al. 2013). There are currently very few large RNAs with complex structures whose structures are well verified. This is an especially critical problem now that three-reagent SHAPE-direct structure modeling has reached a high level of accuracy for RNAs of known structure. Second, approaches for modeling pseudoknots have advanced significantly (Hajdin et al. 2013) but accurate modeling of more than a single pseudoknot in an RNA remains a challenge, both due to limitations in current energy models and due to the computational requirements for many algorithms. Third, this work has focused on canonical base pairs and does not explicitly model non-canonical pairs, although in many cases these can be inferred from their lack of reactivity towards 1M7. Fourth, SHAPE-directed folding algorithms

currently restrict base pairing partner to within 600 nucleotides. In general, this is a good assumption and, for example, allows full-length ribosomal RNAs to be modeled at high accuracy (Deigan et al. 2009). However, there are important RNA-RNA interactions that occur over distances of a thousand nucleotides or more (Alvarez et al. 2005; Jin et al. 2011) that will not be detected with the current approach. Finally, SHAPE reactivities always reflect the structural ensemble present in solution at the time of probing. If an RNA is partially misfolded or samples multiple conformations, the resulting SHAPE profile will reflect these contributions.

The highly accurate RNA secondary structure modeling reported here involves straightforward experiments with three reagents 1M7 (Mortimer and Weeks 2007), 1M6, and NMIA (Steen et al. 2012). In this work we examined complex RNA structures, including more than 3800 nucleotides, and specifically focused on those RNAs thought to comprise the most difficult known modeling challenges. The limitations outlined above notwithstanding, we believe that three-reagent SHAPE is approaching the upper limit that solution-phase RNA structure probing can accomplish. Three-reagent SHAPE structure probing is experimentally concise, yields consistently accurate RNA structural models, and can be applied to RNAs of any complexity and size, including complete viral genomes and the constituents of entire transcriptomes.

Methods

Chemical probing by differential SHAPE.

Differential SHAPE data for the aptamer domains of the *E. coli* thiamine pyrophosphate (TPP) riboswitch, *V. vulnificus* adenine riboswitch, and *T. maritime* lysine riboswitch were reported previously (Steen et al. 2012). DNA templates (IDT) for *E. coli* 5S rRNA and the tRNAPhe, *F. nucleatum* glycine riboswitch, *B. subtilis* M-Box riboswitch, *T. thermophila* group I intron, and the *O. iheyensis* group II intron RNAs were encoded in the context of flanking 5' and 3' structure cassettes (Wilkinson et al. 2006), amplified by PCR, and transcribed into RNA using T7 RNA polymerase.

RNAs were purified using denaturing polyacrylamide gel electrophoresis, excised from the gel, and passively eluted overnight at 4 °C. 16S and 23S ribosomal RNAs were isolated from DH5α cells during mid-log phase using non-denaturing conditions (Deigan et al. 2009). RNAs were refolded in 100 mM HEPES, pH 8.0, 100 mM NaCl, and 10 mM $MgCl_2$ (Steen et al. 2012). The glycine aptamer RNA was incubated with 5 μM final glycine during folding. After folding, all RNAs were modified in the presence of 8 mM SHAPE reagent and incubated at 37 °C for 3 min (1M6 and 1M7) or 22 min (NMIA). No-reagent controls, containing neat DMSO rather than SHAPE reagent, were performed in parallel.

Following modification and precipitation with ethanol, reagent and control RNAs were subjected to reverse transcription with Superscript III (Invitrogen) using fluorescently labeled primers (VIC dye, Invitrogen) that targeted the 3' structure cassette (Wilkinson et al. 2006). A second, internal primer was used for the group II intron to read through the end of the RNA. A reverse transcription sequencing reaction using ddC and a NED-labeled primer was also performed to allow sequence alignment. Reagent or no-reagent control reactions were combined with sequencing reactions and analyzed using an ABI 3500 capillary electrophoresis instrument. Resulting data were processed using *QuShape* (Karabiber et al. 2013). The ribosomal RNAs were analyzed by a new approach, SHAPE-MaP, which will be described in an independent communication. For all RNAs, 1M7 SHAPE reactivities were normalized using the boxplot approach (Hajdin et al. 2013). In this approach, reactivities were first sorted, and reactivities above either 1.5 × interquartile range or the 90th percentile, whichever value was greater, were excluded as outliers. Next, a normalization factor was calculated by the averaging the next 10% of SHAPE reactivities. The original data set was then divided by the newly calculated normalization factor to yield the final processed data.

Differential SHAPE data analysis.

NMIA and 1M6 SHAPE reactivities were normalized by excluding the top 2% of reactivities and dividing by the average of the next 8% of reactivities. 1M6 reactivities were then scaled more

precisely to NMIA reactivities by minimizing the reactivity difference over a 51-nt sliding window. The scaled 1M6 reactivities were subtracted from NMIA reactivities to yield a differential SHAPE profile (Fig. 3.2). This algorithm, implemented in a python program, is included in the Supplemental Materials.

Differential SHAPE pseudo-free energy change penalty.

RNAs with secondary structures derived from high-resolution methods (crystallography or NMR) were used to classify the conformation of nucleotides as either paired (G-C, A-U or G-U) or non-paired. Next, a histogram of differential reactivities (NMIA reactivity minus 1M6 reactivity) for each category was created using a bin-width of 0.2 SHAPE units. Positive and negative differential SHAPE reactivities were treated separately. A ΔG_{Diff} statistical energy potential was then fit using an approach analogous to those used extensively for protein modeling (Rohl et al. 2004) and recently for RNA modeling (Cordero et al. 2012). Histograms of paired and non-paired differential nucleotides from all RNAs were pooled and fit to a gamma distribution (Fig. 3.3A). A free energy at a temperature (T) of 310 K was calculated using the Gibbs relationship:

$$\Delta G_{\mathrm{Diff}} = -k_{\mathrm{b}} T \ln\left(\frac{P(x)paired}{P(x)nonpaired}\right)$$

P(x)paired and *P(x)nonpaired* are the probabilities that a nucleotide is paired or non-paired at SHAPE reactivity x, respectively; kb is the Boltzmann constant; and ΔG_{Diff} is the resulting free change energy penalty that should be applied to a particular differential SHAPE reactivity, x. The resulting function was linear with an intercept near zero. To simplify the calculation and to make the energy function continuous for all differential reactivities, ΔGDiff was fit to a linear equation with an intercept of zero. A standard error measurement of the fit was estimated by a leave-one-out jackknife approach; the resulting fit was a line with a slope of 2.11 kcal/mol and an intercept of zero (Fig. 3.3B).

		SHAPE	–		+		+		
		differential	–		–		***NMIA–1M7***		
	length (nts)		sens	ppv	sens	ppv	sens	ppv	
TPP riboswitch, *E. coli*	79		77.3	85.0	96.5	91.3	95.5	100.0	responsive
5S rRNA, *E. coli*	120		28.6	25.0	85.7	76.9	94.3	91.7	
Glycine riboswitch, *F. nuleatum*	158		70.0	60.9	55.0	48.9	97.5	95.1	
Group I intron, *T. thermophila*	425		83.3	75.0	93.2	91.2	82.6	88.7	
3' domain of 16S rRNA, *E. coli*	478		26.7	21.2	89.5	77.6	97.1	86.8	
5' domain of 16S rRNA, *E. coli*	530		61.3	57.9	97.8	91.8	88.3	83.4	
Average			**57.9**	**54.2**	**86.3**	**79.6**	**92.5**	**91.0**	
Adenine riboswitch, *V. vulnificus*	71		100.0	100.0	100.0	100.0	100.0	100.0	non-responsive
tRNA phe, *E. coli*	76		100.0	91.3	100.0	75.0	100.0	77.8	
cyclic-di-GMP riboswitch, *V. cholerae*	97		75.0	77.8	89.2	86.2	89.2	86.2	
M-Box riboswitch, *B. subtilis*	154		87.5	91.3	83.3	90.9	83.3	93.0	
Lysine riboswitch, *T. maritime*	174		75.8	84.8	84.9	90.3	84.9	90.3	
Domain III of 23S rRNA, *E. coli*	372		46.9	43.1	82.7	74.3	82.7	74.3	
Group II Intron, *O. iheyensis*	412		88.0	97.5	93.2	96.9	94.0	98.4	
Domain II of 23S rRNA, *E. coli*	685		87.6	78.6	97.8	87.4	97.8	87.9	
Average			**82.6**	**83.0**	**91.4**	**87.6**	**91.5**	**88.5**	
Overall Average			**72.0**	**70.7**	**89.2**	**84.2**	**91.9**	**89.5**	

Table 3.3: RNA secondary structure modeling accuracies for a two-reagent differential SHAPE experiment using 1M7 and NMIA. RNAs are listed based on whether or not structure prediction was sensitive to the NMIA–1M7 differential reactivities. The two-reagent experiment yielded significant modeling improvements relative to prediction with 1M7 data only, but improvements were not as large as those with the recommended three-reagent experiment (Table 3.1).

Exploration of simpler differential SHAPE energy potentials.

We explored the possibility of omitting the 1M6 experiment and calculating differential SHAPE reactivities based only on 1M7 and NMIA experiments. Reactivity differences between NMIA and 1M7 were calculated for each nucleotide using the difference subtraction algorithm outlined above. The relationship was linear with a slope of 2.91 kcal/mol. Standard errors resulting from a leave-one-out jackknife analysis were of similar magnitude to those of the relationship between NMIA and 1M6 reactivities. This two-reagent version of the differential SHAPE experiment yielded significant improvements to RNA secondary structure modeling (Table 3.3); however, the three-reagent analysis ultimately yielded more accurate structure models (compare Tables 3.1 and 3.3). Due to the higher information content of the NMIA–1M6 differential analysis, we recommend

using three reagents (1M7, 1M6, and NMIA) to achieve highest accuracies in secondary structure modeling.

During the course of fitting our new differential SHAPE data we also refit the 1M7 free energy potential using a statistical potential and our previously published RNA data set (Hajdin et al. 2013). Paired and non-paired nucleotide distributions were fit to a mixture of two gamma distributions and a free energy change term was calculated using the Gibbs relationship. The resulting free energy change function was comparable in magnitude and x-intercept to the prior grid-search optimized log function (Fig. 3.7). Thus, we have chosen to use the original log-function for incorporating 1M7 data into SHAPE-directed structure modeling.

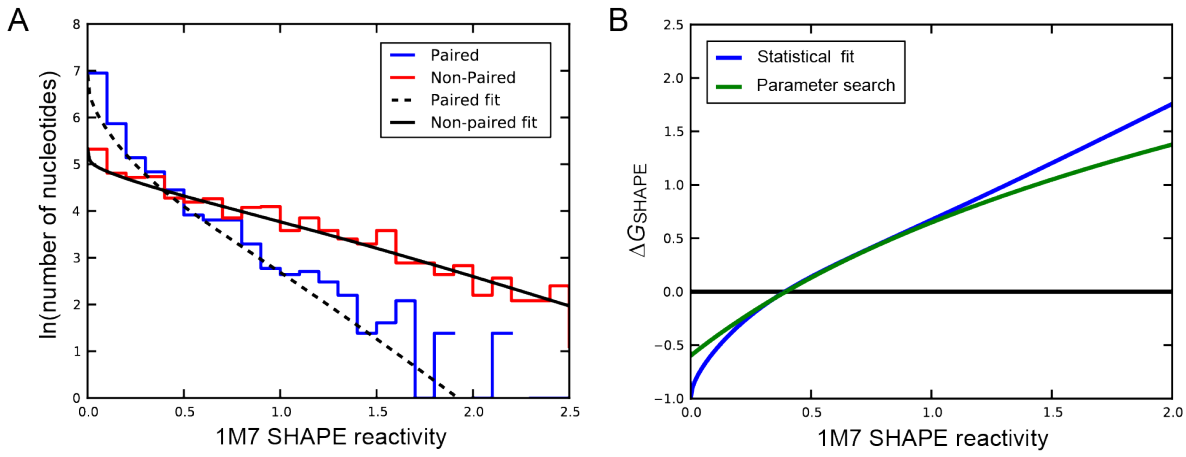

Figure 3.7: Comparison of the statistically determined pseudo-free energy change term with the grid-search optimized ln-form ΔGSHAPE. (A) 1M7-shape reactivities (Hajdin et al. 2013) were binned based on pairing status in the accepted structure and a histogram for each group was generated based on SHAPE reactivity. Histograms were fit to a double gamma distribution. (B) The resulting free energy change from the distribution fitting (blue line) compared to the parameter search optimized *ln*-form free energy change developed previously (Hajdin et al. 2013) [$\Delta G_{SHAPE} = 1.8 \times ln(\text{SHAPE} + 1) - 0.6$].

Implementation in RNAstructure Fold and ShapeKnots.

A modified SHAPE energy file was created for use in *RNAstructure* Fold (Reuter and Mathews 2010) and *ShapeKnots* (Hajdin et al. 2013) to incorporate the differential SHAPE information. Differential pseudo-free energy change values (ΔG_{Diff}) for each nucleotide were calculated from the positive-amplitude differential reactivities (d):

$$\Delta G(d)_{\text{Diff}} = \begin{cases} 2.11d & \text{if } d > 0 \\ 0 & \text{if } d \leq 0 \end{cases}$$

SHAPE pseudo-free energy changes were calculated from 1M7 reactivities using the log-form SHAPE equation (Hajdin et al. 2013):

$$\Delta G_{\text{SHAPE}} = 1.8\, ln(\text{SHAPE} + 1) - 0.6$$

These two free energies were summed, and a modified SHAPE reactivity file was calculated for use in *Fold* or *ShapeKnots* such that, when used with slope of 1.0 and an intercept of -1.0, the folding algorithm applies the appropriate pseudo-free energy change term:

$$\text{SHAPE} = e^{(\Delta G_{\text{SHAPE}} + \Delta G_{\text{Diff}} + 1)} - 1$$

Future versions of *ShapeKnots* and *Fold* will simplify this procedure and allow the 1M7 and differential-SHAPE magnitudes to be entered directly from a data file. For ShapeKnots, the optimized pseudoknot parameters (P1 = 3.5, P2 = 6.5) (Hajdin et al. 2013) were used. The *maxtracebacks* option was set to 100 and the *window* option was set to 0 to maximize the number of potential identified structures.

The calculation for folding RNAs using 1M7 rather than 1M6 as the differential reagent was performed in the same way, except that the differential slope was 2.91. The resulting folds are summarized in Table 3.3. In general, we recommend using *ShapeKnots* for RNA secondary structure modeling because of its ability to predict pseudoknots (Hajdin et al. 2013); at a practical level, this program is limited to RNAs under ~700 nts in length.

Plots and figures.

Secondary structure plots were constructed using *VARNA* (Darty et al. 2009) and circle plots were made using *CircleCompare*, a part of *RNAstructure* (Reuter and Mathews 2010). Model sens was calculated as the number of correct base pairs divided by the total number of base pairs in the correct structure; ppv was calculated as the number of correct base pairs divided by the total number of predicted base pairs. sens and ppv values for ribosomal domains were calculated after omitting regions (Deigan et al. 2009) in which SHAPE reactivities were clearly not consistent with the pattern of base pairing in the accepted secondary structure model.

REFERENCES

Alvarez DE, Lodeiro MF, Ludueña SJ, Pietrasanta LI, Gamarnik AV. 2005. Long-range RNA-RNA interactions circularize the dengue virus genome. *J Virol* **79**: 6631–6643.

Archer EJ, Simpson MA, Watts NJ, O'Kane R, Wang B, Erie DA, McPherson A, Weeks KM. 2013. Long-range architecture in a viral RNA genome. *Biochemistry* **52**: 3182–3190.

Bailor MH, Mustoe AM, Brooks CL, Al-Hashimi HM. 2011. Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Curr Opin Struct Biol* **21**: 296–305.

Cordero P, Kladwang W, VanLang CC, Das R. 2012. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* **51**: 7037–7039.

Dann CE, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. 2007. Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130**: 878–892.

Darty K, Denise A, Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**: 1974–1975.

Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106**: 97–102.

Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM. 2012. Functional complexity and regulation through RNA dynamics. *Nature* **482**: 322–330.

Gherghe CM, Mortimer SA, Krahn JM, Thompson NL, Weeks KM. 2008. Slow conformational dynamics at C2'-endo nucleotides in RNA. *J Am Chem Soc* **130**: 8884–8885.

Grohman JK, Gorelick RJ, Lickwar CR, Lieb JD, Bower BD, Znosko BM, Weeks KM. 2013. A guanosine-centric mechanism for RNA chaperone function. *Science* **340**: 190–195.

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci* **110**: 5498–5503.

Hajdin CE, Ding F, Dokholyan NV, Weeks KM. 2010. On the significance of an RNA tertiary structure prediction. *RNA* **16**: 1340–1349.

Jin Y, Yang Y, Zhang P. 2011. New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. *RNA Biol* **8**: 450–457.

Karabiber F, McGinnis JL, Favorov OV, Weeks KM. 2013. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19**: 63–73.

Kladwang W, VanLang CC, Cordero P, Das R. 2011a. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem* **3**: 954–962.

Kladwang W, VanLang CC, Cordero P, Das R. 2011b. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* **50**: 8049–8056.

Leonard CW, Hajdin CE, Karabiber F, Mathews DH, Favorov OV, Dokholyan NV, Weeks KM. 2013. Principles for understanding the accuracy of SHAPE-directed RNA structure modeling. *Biochemistry* **52**: 588–595.

Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16**: 279–287.

Low JT, Weeks KM. 2010. SHAPE-directed RNA secondary structure prediction. *Methods* **52**: 150–158.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101**: 7287–7292.

Mathews DH, Turner DH. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* **16**: 270–278.

Mauger DM, Siegfried NA, Weeks KM. 2013. The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS Lett* **587**: 1180–1188.

McGinnis JL, Dunkle JA, Cate JHD, Weeks KM. 2012. The mechanisms of RNA SHAPE chemistry. *J Am Chem Soc* **134**: 6617–6624.

Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* **127**: 4223–4231.

Mortimer SA, Weeks KM. 2007. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc* **129**: 4144–4145.

Mortimer SA, Weeks KM. 2009. C2'-endo nucleotides as molecular timers suggested by the folding of an RNA domain. *Proc Natl Acad Sci* **106**: 15622–15627.

Munroe R. 2012. Star Ratings. *xkcd*. http://xkcd.com/1098/ (Accessed August 5, 2013).

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf* **11**: 129.

Rivas E, Lang R, Eddy SR. 2012. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **18**: 193–212.

Rohl CA, Strauss CEM, Misura KMS, Baker D. 2004. Protein structure prediction using Rosetta. *Meth Enzymol* **383**: 66–93.

Sharp PA. 2009. The centrality of RNA. *Cell* **136**: 577–580.

Steen K-A, Rice GM, Weeks KM. 2012. Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *J Am Chem Soc* **134**: 13160–13163.

Tinoco I, Bustamante C. 1999. How RNA folds. *J Mol Biol* **293**: 271–281.

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**: 711–716.

Weeks KM. 2010. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol* **20**: 295–304.

Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, Mathews DH, Giddings MC, Weeks KM. 2008. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol* **6**: e96.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1**: 1610–1616.

CHAPTER 4: AUTOMATED MOTIF DISCOVERY IN LARGE RNAS USING SHAPE-MAP AND *SUPERFOLD*[3]

Introduction

Higher-order structures govern most aspects of RNA function, modulating interactions with small molecule ligands, individual proteins, large multi-component complexes, and other small and large RNAs (Sharp 2009; Dethoff et al. 2012). There are numerous features of RNA structure that are difficult or impossible to determine from sequence-based analysis alone. Inclusion of data from chemical probing experiments, in which an RNA reacts with diagnostic chemical reagents in a structure-selective way, dramatically improves the accuracy of RNA structure modeling (Weeks 2010).

Substantial effort has therefore been directed toward developing high-throughput approaches to analyze RNA secondary structure. Recently reported approaches for RNA structure analysis that use massively parallel sequencing to read out the results of enzymatic, selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE), or dimethyl sulfate (DMS) probing have provided comprehensive support for large-scale comparative trends in transcript structure but have not been shown to yield accurate secondary structure models or enable novel motif discovery (Mathews et al. 2004; Kertesz et al. 2010; Mauger and Weeks 2010; Underwood et al. 2010; Lucks et al. 2011; Weeks 2011; Ding et al. 2014; Rouskin et al. 2014). In general, these "-seq" approaches are not well suited to recovering RNA structure probing information because they require complex RNA ligation and library preparation steps that result in substantial nucleobase and local structure biases. In

[3] This chapter has been previously published and represents a co-first author work. My contributions were performing SHAPE-MaP experiments on model RNAs, designing the *Superfold* RNA folding and motif discovery analyses, and collaborating with the other authors in interpreting the experiments and writing the manuscript. The original citation is as follows: Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, *11*(9), 959–965. doi:10.1038/nmeth.3029

addition, there is no known pathway for using enzyme or DMS probing data, which report on only a subset of nucleotides, to model complex RNAs accurately. Moreover, understanding many critical features of RNA folding mechanisms (Grohman et al. 2013), RNA-protein interactions (Wilkinson et al. 2008; Gherghe et al. 2010), and in-cell effects on RNA folding and structure (Tyrrell et al. 2013; McGinnis and Weeks 2014) require that all four RNA nucleotides be interrogated simultaneously. In this paper we describe approaches for both quickly generating thousands of nucleotides of accurate chemical probing information and for computer-aided discovery of novel RNA motifs and large-scale structure modeling.

Results

The MaP strategy

SHAPE (Mortimer and Weeks 2007; Weeks and Mauger 2011; Rice et al. 2014) experiments use 2'-hydroxyl-selective reagents that react to form covalent 2'-*O*-adducts at conformationally flexible RNA nucleotides, both under simplified solution conditions (Merino et al. 2005; Wilkinson et al. 2008) and in cells (Spitale et al. 2013; Tyrrell et al. 2013; McGinnis and Weeks 2014). Recent innovations that include SHAPE data as restraints in RNA structure prediction algorithms consistently yield highly accurate secondary structure models for structurally complex RNAs (Hajdin et al. 2013; Rice et al. 2014). Here we quantify SHAPE chemical modifications (Merino et al. 2005; Mortimer and Weeks 2007; Steen et al. 2012; Rice et al. 2014) in RNA in a single direct step by massively parallel sequencing (Fig. 4.1). The approach exploits conditions that cause reverse transcriptase to misread SHAPE-modified nucleotides and incorporate a nucleotide non-complementary to the original sequence in the newly synthesized cDNA. The positions and relative frequencies of SHAPE adducts are thus immediately, directly, and permanently recorded as mutations in the cDNA primary sequence, thereby creating a SHAPE mutational profile (SHAPE-MaP). In a SHAPE-MaP experiment, the RNA is treated with a SHAPE reagent or treated with solvent only, and

the RNA is modified under denaturing conditions to control for sequence-specific biases in detection of adduct-induced mutations (Fig. 4.2a). RNA from each experimental condition is subjected to reverse transcription, and the resulting cDNAs are then prepared for massively parallel sequencing. Reactive positions are identified by subtracting data from the treated sample from data obtained from the untreated sample and by normalizing to data from the denatured control (Figs. 4.1 and 4.2b).

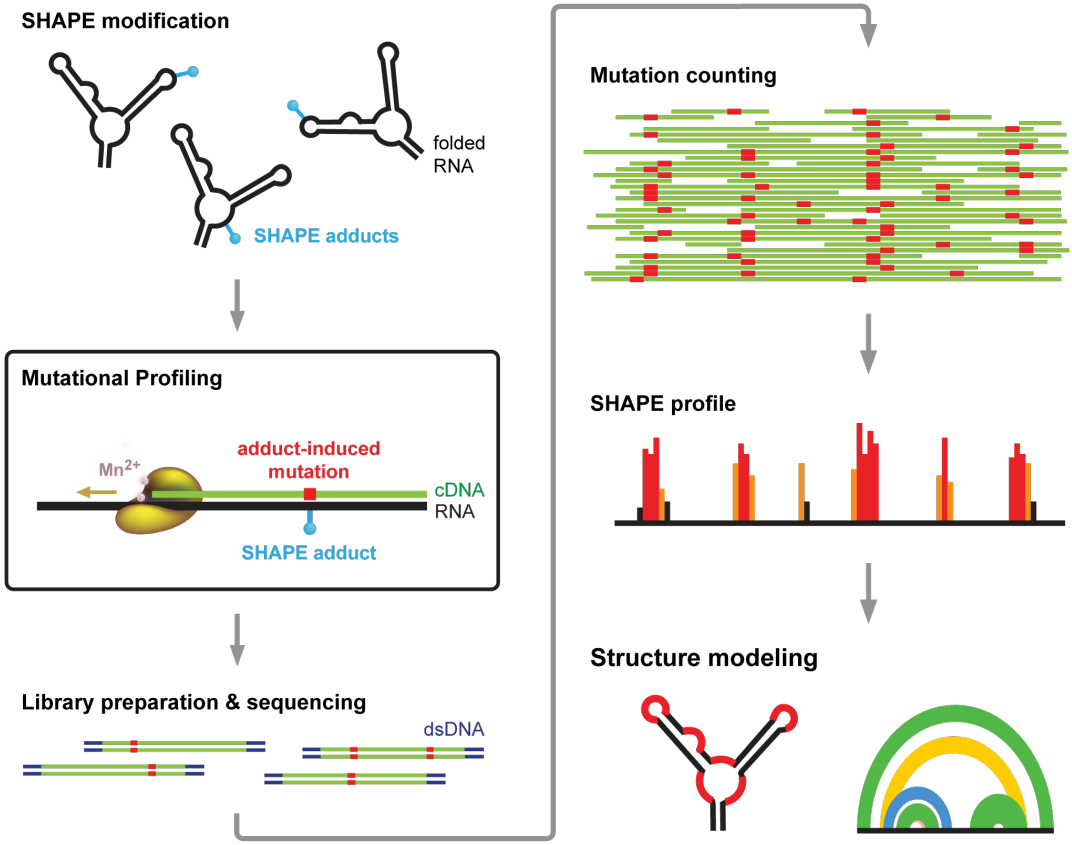

Figure 4.1: Overview of the SHAPE-MaP approach. RNA is treated with a SHAPE reagent that reacts at conformationally dynamic nucleotides. During reverse transcription the polymerase reads through chemical adducts in the RNA and incorporates a nucleotide non-complementary to the original sequence (red) into the cDNA. The resulting cDNA is sequenced using any massively parallel approach to create mutational profiles (MaP). Sequencing reads are aligned to a reference sequence, and nucleotide-resolution mutation rates are calculated, corrected for background and normalized, producing a standard SHAPE reactivity profile. SHAPE reactivities can then be used to model secondary structures, visualize competing and alternative structures, or quantify any process or function that modulates local nucleotide RNA dynamics.

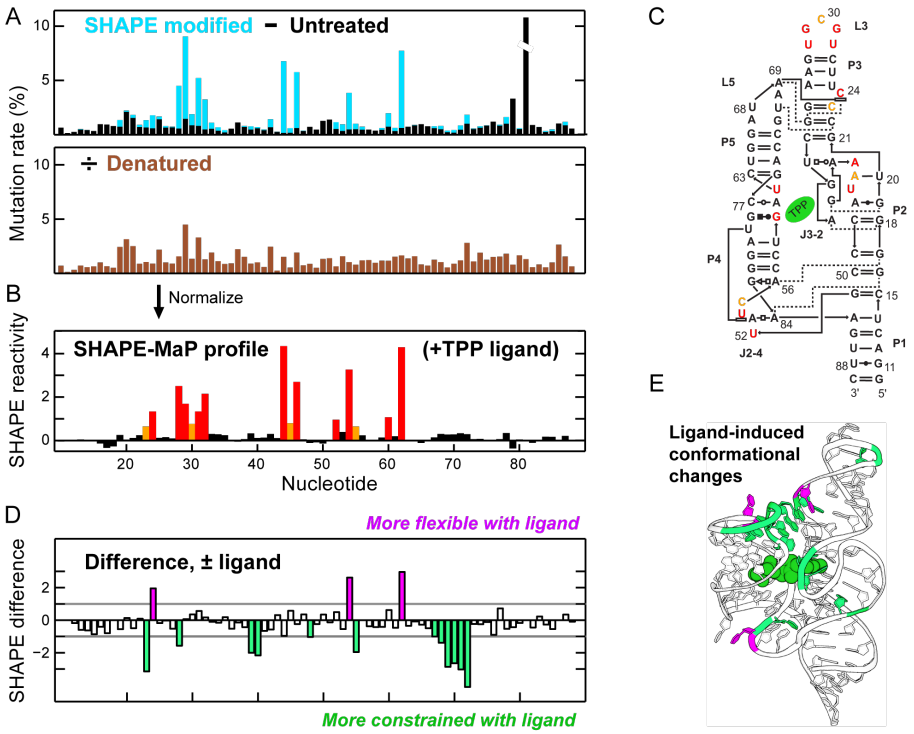

Figure 4.2: Nucleotide-resolution interrogation of RNA structure and ligand-induced conformational changes. (a) Mutation rate profiles for the SHAPE modified and untreated TPP riboswitch RNA in the presence of ligand (*top*) and for SHAPE modification performed under denaturing conditions (*bottom*). (b) Quantitative SHAPE profile obtained after subtracting the data from the untreated sample from data for the treated sample and normalizing by the denatured control. (c) SHAPE reactivities plotted on the accepted secondary structure of the ligand-bound TPP riboswitch. Red, orange, and black correspond to high, moderate, and low reactivities, respectively. (d) Difference SHAPE profile showing conformational changes in the TPP riboswitch upon ligand binding. (e) Superposition of ligand-induced conformational changes on the TPP riboswitch structure. Data are representative of two biological replicates.

Structure modeling: validation

We initially examined the structure of the *E. coli* thiamine pyrophosphate (TPP) riboswitch aptamer domain in the presence and absence of saturating concentrations of the TPP ligand (Fig. 4.2). SHAPE-MaP profiles recapitulated the known reactivity pattern for the folded, ligand-bound RNA (Fig. 4.2b-c) and accurately reported nucleotide-resolution reactivity differences that occur upon ligand binding (Fig. 4.2d-e). These results, and an analysis of the 1542-nt *E. coli* 16S rRNA (Figs. 4.3, 4.4), demonstrate the ability of SHAPE-MaP to capture fine structural details for distinct RNA conformations at nucleotide resolution, accurately and reproducibly, and independently of nucleotide type. Because the SHAPE profiles are reconstructed from mutation frequencies derived from all

sequencing reads, uncertainties in SHAPE reactivities can be estimated from the Poisson distribution of mutation events (Figs. 4.3, 4.4).

Use of SHAPE data as pseudo-free energy change terms to constrain secondary structure modeling has been extensively benchmarked using RNA test sets specifically chosen to be challenging to conventional secondary structure modeling (Hajdin et al. 2013; Rice et al. 2014). To assess the accuracy of SHAPE-MaP, we probed a subset of these RNAs, ranging in size from 78 to 2,904 nucleotides, with the well-validated 1M7 reagent (Mortimer and Weeks 2007).

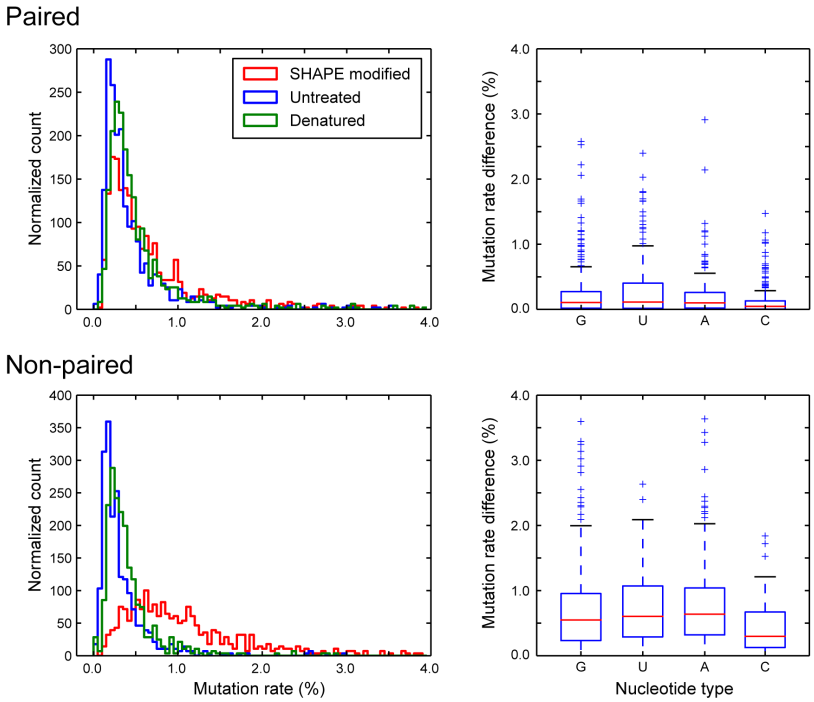

Figure 4.3: Mutation rate histograms for paired and non-paired nucleotides in the 16S rRNA. Nucleotides were separated into paired (*upper panels*) and non-paired (*lower panels*) groups based on their observed pairing in the *E. coli* 16S rRNA[38]. Mutation rate histograms for each experimental sample (SHAPE, untreated, and denatured) were calculated based on pairing status (*left*-hand panels). Distributions of mutation rates for the SHAPE-modified and untreated samples are similar for base-paired nucleotides; whereas nucleotides in non-paired conformations are much more reactive towards SHAPE probing. (*right*-hand panels) SHAPE-MaP reactivities are independent of nucleotide type.

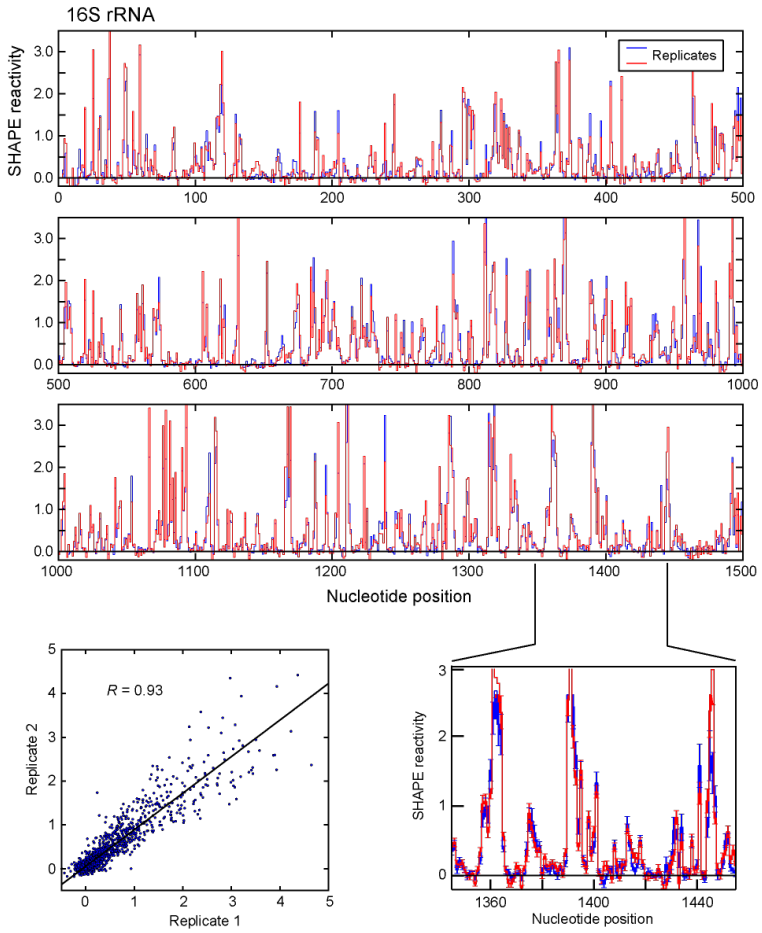

Figure 4.4: SHAPE-MaP replicates of *E. coli* 16S rRNA. Data correspond to full biological replicates performed six months apart by different individuals. The inset for nucleotides 1350-1450 (bottom right) shows standard errors.

We also evaluated the "differential" SHAPE experiment that uses two additional reagents – 1M6 and NMIA – to detect non-canonical and tertiary interactions and yields RNA structural models with consistent high accuracy, even for especially challenging RNAs (Steen et al. 2012; Rice et al. 2014). The overall accuracy of SHAPE-MaP directed RNA structure modeling using differential reactivities, measured in terms of sensitivity (sens) and positive predictive value (ppv), was similar to and often superior to that of conventional SHAPE reactivities based on adduct-mediated termination of primer extension detected by capillary electrophoresis. The accuracy for recovery of accepted, canonical base pairs exceeded 90% (Fig. 4.5a).

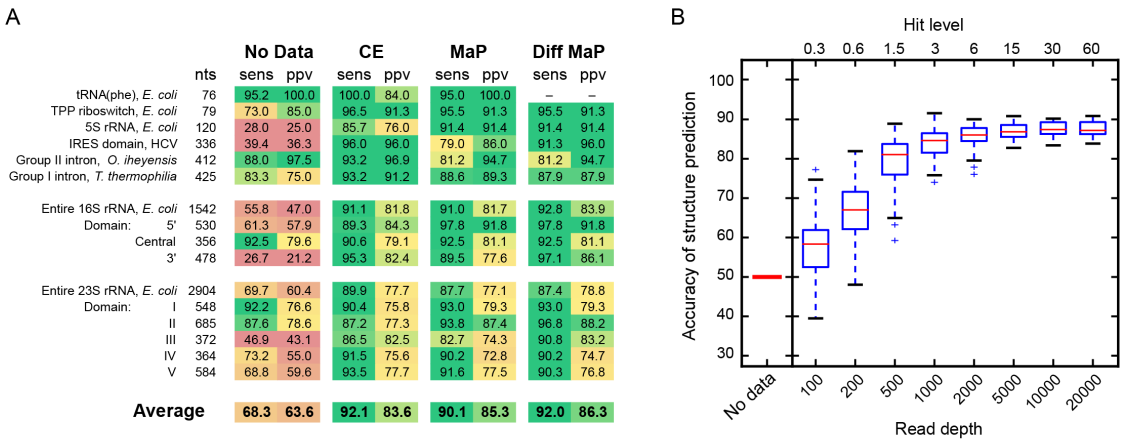

Figure 4.5: Accuracy of SHAPE-MaP-directed secondary structure modeling. (a) Secondary structure modeling accuracies reported as a function of sensitivity (sens) and positive predictive value (ppv) for calculations performed without experimental constraints (no data), with conventional capillary electrophoresis (CE) data, and with SHAPE-MaP data obtained with the 1M7 reagent (Deigan et al. 2009; Hajdin et al. 2013) or with three-reagent differential (Diff) data (Rice et al. 2014). Results are colored on a scale to reflect low (red) to high (green) modeling accuracy. (b) Relationship between sequencing read depth, hit level, and accuracy of RNA structure modeling. Model accuracy (vertical axis) is shown as the geometric average of the sens and ppv of predicted structures with respect to the accepted model (Deigan et al. 2009). Boxplots summarize modeling the secondary structure of the 16S ribosomal RNA as a function of simulated SHAPE-MaP read depth. At each depth, 100 folding trajectories were sampled. The line at the center of the box indicates the median value and boxes indicate the interquartile range. Whiskers contain data points that are within 1.5 times the interquartile range and outliers are indicated with (+) marks. Hit level is the total signal above normalized background per transcript nucleotide.

SHAPE reactivities obtained using the MaP strategy are measured as many individual events by massively parallel sequencing. Reliability depends on adequate measurement of mutation rates. We achieved accurate modeling of the 16S rRNA structure using a per-nucleotide read depth of 2,000-5,000. This corresponds to 6 to 15 modifications above background per ribosomal nucleotide on average (Fig. 4.4b). Although several prior studies (Kertesz et al. 2010; Ding et al. 2014; Rouskin et al. 2014) have been performed in which all of the RNAs in a given transcriptome were physically present during the probing phase of the experiment, only a few thousand nucleotides in each case were sampled at a depth that would allow full recovery of the underlying structural information. Importantly, we achieved accurate SHAPE-MaP directed modeling using the same parameters originally defined for capillary electrophoresis-based experiments and obtained comparable high accuracies using both RNA-specific and randomly primed experiments (Fig. 4.5a). Data were highly

reproducible between experimental replicates performed months apart by different individuals (Fig. 4.4), emphasizing the robustness of SHAPE-MaP.

A second-generation model for an HIV-1 RNA genome

We obtained single-nucleotide resolution structural information for the entire authentic HIV-1_{NL4-3} genomic RNA (~9,200 nts) in experiments and data analysis performed over roughly 2 weeks. The 1M7 and differential SHAPE-MaP data were processed to yield SHAPE reactivity profiles and secondary structure models using efficient and fully automated algorithms (Figs. 4.6, 4.7). Since our report in 2009 of a model for the HIV-1 RNA genome, we have made multiple, fundamental advances in SHAPE-directed RNA structure modeling. These innovations include improved energy models, the ability to model pseudoknots, and concise strategies for detecting tertiary and non-canonical interactions (Hajdin et al. 2013; Rice et al. 2014). The MaP approach, implemented in this work, yields nucleotide-resolution reactivity data for large RNAs that are equal or superior to the prior gold standard capillary electrophoresis data (Fig. 4.5). Thus, the HIV-1 genome structure presented here represents a higher resolution, second-generation model for well-defined elements in this RNA.

Development of SuperFold, a large RNA folding algorithm

SuperFold takes a windowing approach to break up the folding of large RNAs. For long RNAs, practical window size choices are roughly 1,200 nucleotides for a partition function calculation and 3,000 nucleotides for a minimum free energy calculation. Dividing the folding of a large RNA into smaller segments allows modern multi-core workstations to model RNA structures in a modest amount of clock-time. *SuperFold* runs in three main stages: partition function calculation, minimum free energy calculation, and structural analysis (Fig. 4.7).

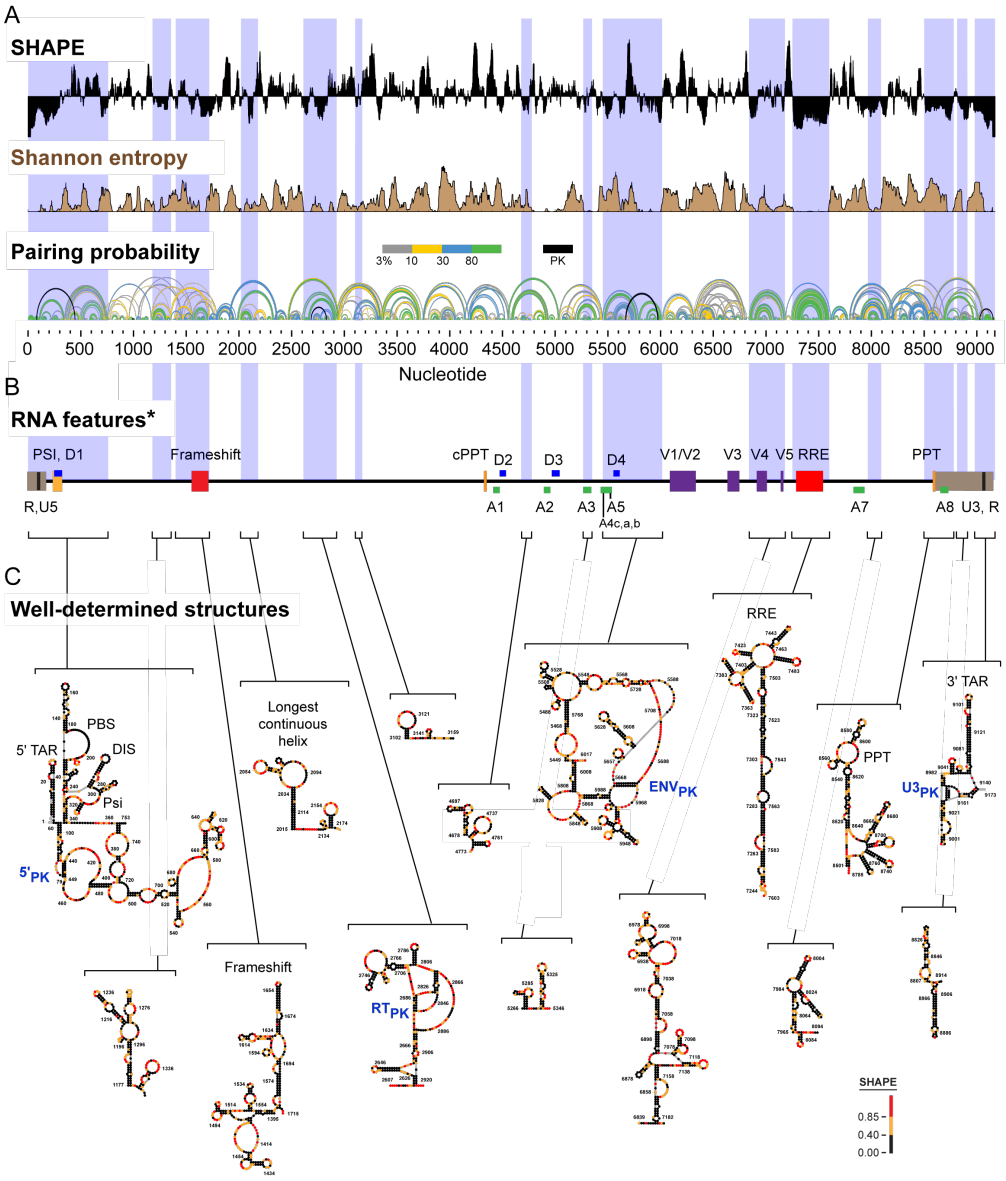

Figure 4.6: SHAPE-MaP analysis of the HIV-1 NL4-3 genome. (a) SHAPE reactivities, Shannon entropy and pairing probability for the NL4-3 HIV-1 genomic RNA. Reactivities are shown as the centered 55-nt median window, relative to the global median; regions above or below the line are more flexible or constrained than the median, respectively. Arcs representing base pairs are colored by their respective pairing probabilities, with green arcs indicating highly probable helices. Areas with many overlapping arcs have multiple potential structures. Pseudoknots (PK) are indicated by black arcs. (b) RNA regions identified as having biological functions. Brackets enclose well-determined regions and are drawn to emphasize locations of these regions relative to known RNA features in the context of the viral genome. Regions correspond to low SHAPE-low Shannon entropy domains and are extended to include all intersecting helices from the lowest predicted free-energy secondary structure. 5' and 3' UTRs are brown; splice acceptors and donors are green and blue, respectively; polypurine tracts are yellow; variable domains are purple; and the frameshift and RRE domains are red. (c) Secondary structure models for regions, identified *de novo*, with low SHAPE reactivities and low Shannon entropies. Nucleotides are colored by SHAPE reactivity and pseudoknotted structures are labeled in blue.

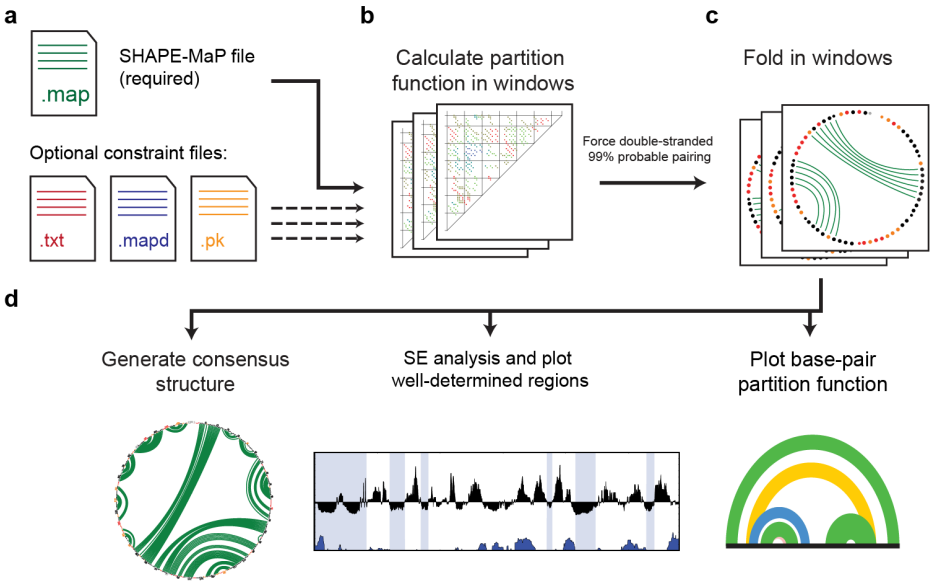

Figure 4.7: Overview of the *SuperFold* pipeline. (a) Input files are defined by the user. Only the ".map" file is required. Optional files allow for modeling of single-stranded interactions, pseudoknots, and inclusion of differential SHAPE reactivities (Rice et al. 2014) in the energy function. (b) The RNA is divided into overlapping windows. The partition function is calculated and positions within 300 nucleotides of edges are not used further. The partition function for the full-length RNA is calculated as the average across each window in which a given base pair is able to occur. (c) Base pairs with 99% probability are used to constrain a minimum free energy calculation in windows of 4,000 nucleotides. (d) In the final step, a consensus structure is generated by requiring that base pairs in step (c) occur in greater than one-half of windows. A Shannon entropy analysis is used to identify well determined regions, and the partition function of probable base pairs is plotted as arcs across the RNA.

The partition function and minimum free energy structure calculation are implemented using *RNAstructure* (Reuter and Mathews 2010), since *RNAstructure* has been written to directly incorporate SHAPE reactivity information (Deigan et al. 2009; Hajdin et al. 2013).

Two important assumptions are made in order to model RNA folding in windows: (*i*) RNA structure is predominately local in nature. A maximum pairing distance of 600 nts is currently implemented in *RNAstructure*. It is a practical, but imperfect, assumption that pairing does not occur outside this number of nucleotides (Deigan et al. 2009; Rice et al. 2014). In addition, a consequence of this implementation is that improperly choosing the "ends" of the RNA will introduce (potentially) cascading effects on nucleotide pairing. To mitigate this effect, predicted structures 300 nucleotides from the 5' and 3' ends of a given window are removed. (*ii*) The most stable structure will

predominate despite potentially poorly chosen 5' and 3' ends. *SuperFold* combines predicted pairs from many overlapping windows and requires base pairs that occur in more than half of potential cases to be retained in a minimum free energy secondary structure model.

The partition function computed over a given RNA can be informative for determining the regions of an RNA that form well-defined structures versus those regions that are likely to exist as an ensemble of structures (McCaskill 1990). *SuperFold* calculates the partition function of an RNA in windows of 1200 nucleotides. Interactions within 300 nucleotides of the window 5' and 3' ends are removed. Pairing probabilities are then averaged across each window in which a given base pair is able to occur. Additional partition function calculations are performed using the true 5' and 3' ends to reduce de-weighting of the partition function at the ends of an RNA.

The partition function can also be used to identify helices that are most likely to be modeled correctly. Nucleotides with predicted pairing probabilities above 0.99 appear to be correct more than 90% of the time (Mathews 2004). This observation is used to constrain the minimum free energy structure prediction using *RNAstructure Fold* in 3,000 nucleotide sliding windows. Highly likely pairs, based on the partition function within a folding window, are constrained to be base paired. Nucleotides split by an overlapping window are forced to be single stranded. The combination of these two constraints mitigates the effects of inadvertently poorly chosen ends.

De novo identification of well-determined structures

Almost any long RNA sequence will form some secondary structures (Doty et al. 1959), but not all of these structures are biologically important or well-defined. Therefore, we used SHAPE-directed modeling, whose underlying energy function yields highly accurate models for RNAs with well-defined secondary structures (Fig. 4.5), to calculate a probability for each base pair across all possible structures in the Boltzmann ensemble of structures predicted for the HIV-1 RNA. These probabilities were used, in turn, to calculate Shannon entropies (Huynen et al. 1997; Mathews 2004) (Fig. 4.6). Regions with higher Shannon entropies are likely to form alternative structures, and those

with low Shannon entropies correspond to regions with well-defined RNA structures or persistent single-strandedness, as determined by SHAPE reactivity.

The plot of pairing probability across the entire HIV-1 genome reveals both well-determined and variable RNA structures in the HIV-1 genomic RNA (Fig. 4.6a). Previously characterized structured regions such as the 5'-UTR, Rev response element (RRE), frameshift element, and polypurine tract (PPT) are well determined in the model (represented by green arcs). In contrast, there are also large regions – for example, from nucleotides 3200 to 4500 and from nucleotides 6100 to 6800 – that have high SHAPE reactivities and high Shannon entropy and are therefore likely to sample many conformations (shown as blue, yellow, and gray arcs). This visualization approach highlights regions with unique, likely stable structures and those regions where multiple structures are likely to be in equilibrium.

Critically, analysis of Shannon entropies and SHAPE reactivities provides an approach for *de novo* discovery of regions with well-defined structure in long RNAs. Fifteen regions in the HIV-1 genomic RNA had both low SHAPE reactivity values (indicating a high degree of RNA structure) and low Shannon entropies (providing confidence in a single predominant secondary structure) (Fig. 4.6a, b, shaded in purple). We created nucleotide-resolution structure models for each of these regions (Fig. 4.6c). The models of known, functionally important regulatory structures – RRE, 5' TAR, primer binding site (PBS), packaging element PSI structures, ribosomal frameshift element, and 3' TAR – agreed closely with previously proposed models for these regions. In addition, the longest continuous helix, the hairpins flanking the polypurine tract, and other features remain consistent between the prior(Watts et al. 2009) and this second-generation model. We next assembled a list of all regulatory elements likely to function via an RNA motif (Fig. 4.6b). We then compared the locations of these RNA structural elements with the highly structured and low entropy regions identified *de novo* by SHAPE-MaP. Functional RNA elements occur overwhelmingly in low SHAPE, low Shannon entropy regions (p-value = 0.002; Fig. 4.6), indicating that most RNA-mediated functions operate in the context of an underlying RNA structure. Several low SHAPE, low Shannon

entropy regions in the HIV-1 genome occur in regions not previously associated with known RNA functional elements: These regions are high-value targets for discovery of new RNA motifs.

Motif discovery and deconvolution of structural polymorphism

Pseudoknots appear to be rare in large RNAs and are difficult to identify; however, these motifs appear to be overrepresented in functionally important regions of many RNAs (Staple and Butcher 2005; Brierley et al. 2007). As a rigorous test of the current cumulative advances in SHAPE-directed structure modeling and of the high-throughput SHAPE-MaP data itself, we searched (Hajdin et al. 2013) for novel pseudoknots in the HIV-1 RNA genome. In our model, there are four pseudoknots in regions of low SHAPE reactivity and low Shannon entropy (Fig. 4.6c). The pseudoknot adjacent to the 5' polyadenylation signal in the HIV-1 RNA ($5'_{PK}$) was previously validated (Paillart et al. 2002; Wilkinson et al. 2008). The three additional, novel pseudoknots are predicted to form in the reverse transcriptase coding region (RT_{PK}), at the beginning of *env* (ENV_{PK}), and in the U3 region adjacent to the 3' polyadenylation signal ($U3_{PK}$). An additional pseudoknot predicted by the ShapeKnots algorithm that lies in a region of high SHAPE reactivity and Shannon entropy ($CA_{PK,}$ nt 961-1014) was analyzed as a negative control.

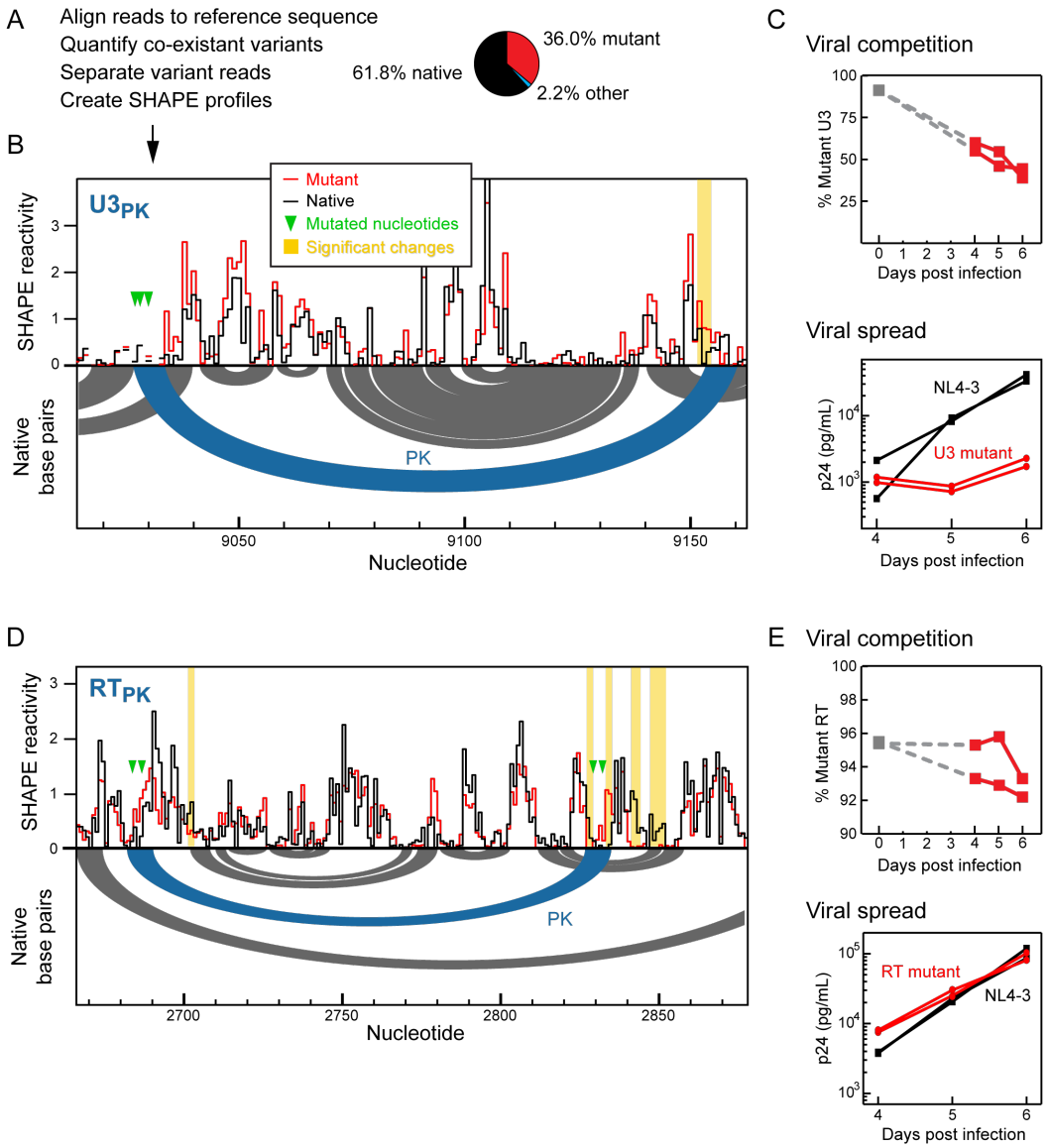

Figure 4.8: Functional and structural validation of newly discovered HIV-1 RNA motifs. (a) Scheme for simultaneous deconvolution and structural analysis of a mixture of native sequence and $U3_{PK}$ mutant genomes. (b) SHAPE profiles for the $U3_{PK}$ pseudoknot bridging U3 and R from a single experiment. The experiment simultaneously probed a mixture of viruses with native sequence and mutant $U3_{PK}$ RNAs. Secondary structure for the native sequence is shown as arcs below the y-axis intercept. Significant SHAPE reactivity differences are emphasized with yellow vertical lines. (c) Direct growth competition and viral spread for $U3_{PK}$ mutant and native sequence NL4-3 HIV-1 virions in Jurkat cells. Each line in the viral spread assay is a biological replicate from representative technical replicates. Percentage of mutant in the initial inoculum is presented as a grey square at day 0, single replicate. p24 levels correspond to the amount of HIV-1 capsid protein. (d) SHAPE profiles for the RT_{PK} pseudoknot within the reverse transcriptase coding region. SHAPE data were obtained in separate experiments for each virus. (e) Viral spread and direct growth competition for RT_{PK} mutant and native sequence NL4-3 HIV-1 virions in Jurkat cells. For the competition data, y-axes are shown on an expanded scale for clarity.

We introduced silent mutations designed to disrupt each pseudoknot into the full-length HIV-1 genome. Special features of the $U3_{PK}$ region illustrate the power of the MaP approach. U3 sequences occur at both the 5' and 3' ends of the viral genome in proviral HIV-1 DNA but only at the 3' end in the viral RNA. During transfection of the provirus-encoding plasmid, these sequences can undergo recombination. When we introduced a single mutant copy of the U3 sequence (at the 3' end) into the pNL4-3 provirus, we observed partial recombination with the native sequence U3 at the 5' end of the proviral DNA. SHAPE-MaP experiments revealed that both native and mutant sequences were present at the 3' ends of individual genomic RNAs in the mutant $U3_{PK}$ sample. Critically, because nucleotides are analyzed in the context of unfragmented RNA regions in the MaP approach, we were able to independently monitor both alleles in the same experiment, computationally separate them, and construct native and mutant SHAPE profiles (Fig. 4.8a, b). Notable SHAPE reactivity differences between native and mutant U3 were observed, produced by viruses in direct competition with each other and consistent with precise disruption of the $U3_{PK}$ structure (Fig. 4.8b). Strikingly, mutations introduced in the 5' side of the $U3_{PK}$ pseudoknot helix induced changes in the predicted 3' pairing partner, located over 100 nucleotides away (Fig. 4.8b). SHAPE-MaP is thus uniquely useful for structural analysis and motif discovery in systems that contain complex mixtures of RNAs and for detecting and deconvoluting structural consequences of single-nucleotide and other allelic polymorphisms.

All mutant constructs were analyzed using SHAPE-MaP and in cell-based assays for viral fitness. Mutations in $U3_{PK}$ reduced viral spread in Jurkat cells by ~10-fold relative to NL4-3 and reduced viral fitness in direct competition with NL4-3, with a mean relative fitness difference of –0.32 relative to NL4-3 (Fig. 4.8c) (Resch et al. 2002). This large effect on viral fitness by mutations in the $U3_{PK}$ is consistent with the general importance of 3'-UTRs in regulating mRNA stability and translation (Matoulkova et al. 2012) and, more specifically, with a role for specific higher-order spatial organization of the poly(A) signal and upstream sequence elements in assembly of the polyadenylation machinery (Gilmartin et al. 1992; Klasens et al. 1999). SHAPE changes in the RT_{PK}

mutant were also located directly in or immediately adjacent to the pseudoknotted helix (Fig. 4.8d). Mutations in RT_{PK} showed a smaller, but reproducible, decrease in viral spread and viral fitness, with a mean relative fitness of –0.14, compared to NL4-3 (Fig. 4.8e). We also observed changes in SHAPE reactivities at both the 5' and 3' sequences for the long-distance ENV_{PK} mutant, including changes extending 5' from the pseudoknotted helix, suggestive of local refolding caused by disruption of this pseudoknot (Fig. 4.9). Viral spread and viral fitness were not reduced for the ENV_{PK} mutant, which may reflect the challenge of detecting some features of HIV-1 replication in cell culture. The mutations in CA_{PK} (Fig. 4.9), which we analyzed as a negative control, did not support existence of a pseudoknotted structure at this location by SHAPE-MaP analysis, in agreement with its high Shannon entropy profile.

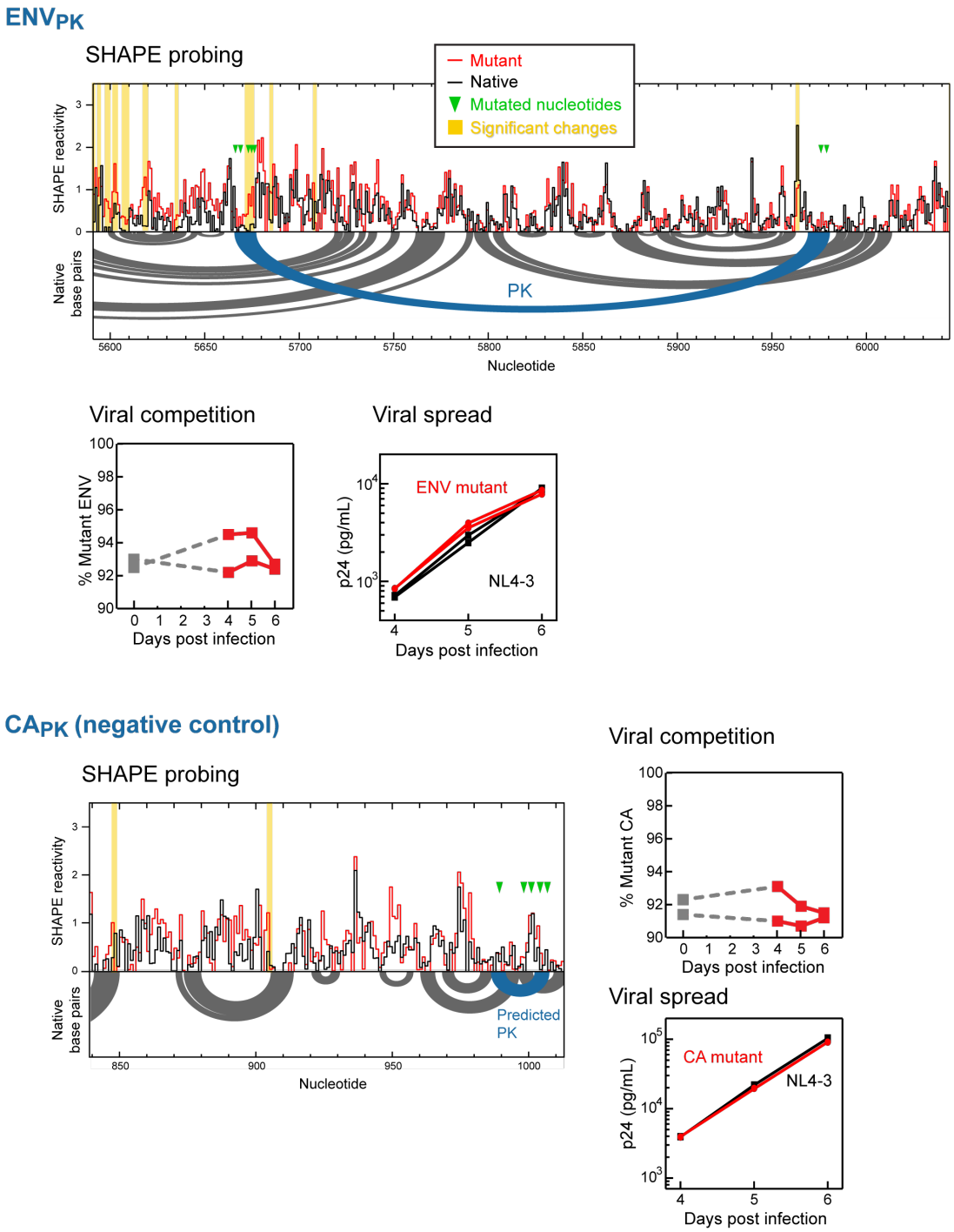

Figure 4.9: Pseudoknot SHAPE-MaP profiles for ENV_{PK} and CA_{PK}. (*upper panel*) SHAPE-MaP and structure profiles for ENV_{PK} and direct growth competition and viral spread data. (*lower panel*) SHAPE-MaP and structure profiles for CA_{PK}, located in a high entropy region of the RNA genome and thus served as a negative control. Also displayed are competition and viral spread assay data.

Discussion

This work defines an alternative strategy for reading out nucleic acid structure probing experiments by massively parallel sequencing. With mutational profiling, nucleic acid structural information is directly and concisely recorded in the sequence of the complementary cDNA and rendered insensitive to biases in library preparation and sequencing. MaP thus converts reverse transcription or DNA synthesis into a direct engine for nucleic acid structure discovery. MaP is fully independent of sequencing strategy and can therefore be used in any sequencing approach with a sufficiently low base call error rate to quantify chemical modifications in any low-abundance RNA detectable by reverse transcription. Detection of chemical adducts in RNA and DNA via direct read-through can be coupled with strategies for polymerase selection (Ghadessy and Holliger 2007; Chen and Romesberg 2014) to record, as mutational profiles or MaPs, a wide variety of post-transcriptional and epigenetic modifications. SHAPE-MaP data contain error estimates and are readily integrated into fully automated, vetted, algorithms for structure modeling (Figs. 4.6, 4.7).

In large- and genome-scale RNA structural studies, true functional elements must be identified in the background of the complex ensemble of structures that form in any large RNA. The combination of SHAPE-MaP analysis with analysis of pairing probabilities, calculated across large RNA regions, identified almost all known large-scale functional elements within the HIV-1 genome, with the exception of the central polypurine tract (cPPT; Fig. 4.6), which appears to have a conserved structure (Pollom et al. 2013). Thus, the sensitivity of functional element detection by SHAPE-MaP is very high. Moreover, despite the fact that the HIV-1 genome is one of most intensively studied RNAs in scientific history, quantitative and high-resolution SHAPE-MaP analysis nonetheless allowed rapid, *de novo* discovery and direct validation of new functional motifs, specifically three pseudoknots, a motif that has traditionally been challenging to predict. The positive predictive value of the approaches developed here is thus also correspondingly high. SHAPE-MaP is unique in its experimental simplicity and structural accuracy and can be scaled to RNA systems of any size and complexity.

Methods

SHAPE-MaP experimental overview.

SHAPE-MaP experiments use specialized conditions for reverse transcription that promote incorporation of nucleotides non-complementary to the RNA into the nascent cDNA at the locations of SHAPE adducts. Sites of RNA adducts thus correspond to internal mutations or deletions in the cDNA, relative to comparison with cDNAs transcribed from RNA not treated with SHAPE reagent. Reverse transcription can be carried out using gene-specific or random primers; both approaches are described below. Once cDNA synthesis is complete, RNA structural information is essentially permanently recorded in the sequence and thus independent of biases introduced during any multi-step library construction scheme. Library preparation is similar to that of an RNA-seq experiment, can be readily tailored to any sequencing platform, and allows multiplexing using sequence barcodes. Single-stranded breaks and background degradation do not intrinsically interfere with SHAPE-MaP experiments (in contrast to conventional SHAPE and other reverse transcriptase stop-dependent assays), as these are not detected during read-through sequencing. There is also no signal decay or drop-off in the MaP approach, which otherwise requires complex, partially heuristic, correction.

SHAPE-MaP development and efficiency

We determined the precise classes of adduct-induced misincorporation events by comparing substitution and deletion rates at non-paired and paired nucleotide positions in the 16S rRNA. Misincorporation trends were similar between all three SHAPE reagents [1M7 (Mortimer and Weeks 2007) and the "differential" reagents NMIA and 1M6 (Rice et al. 2014)]. Generally, the presence of a SHAPE adduct causes nucleotides to be misread as A, T, or deletion events, although there is significant information content in other misincorporation events (Fig. 4.10). Flexible nucleotides in a dinucleotide model substrate with a single reactive position (AddC) (Mortimer and Weeks 2007) are modified with an efficiency of ~2% by NMIA or 1M7 under conditions similar to those used here.

Mutation rates above background at flexible positions in the 16S rRNA are ≥0.5%, with many of the most reactive positions above 2% (Fig. 4.3). Given these boundary values, we estimate that the MaP strategy detects SHAPE adducts with an efficiency of ≥50%.

RNA folding and SHAPE probing of model RNAs. DNA templates (IDT) were synthesized for tRNAPhe, TPP riboswitch, E. coli 5S, hepatitis C virus IRES domain, T. thermophila group I intron, or O. iheyensis group II intron RNAs in the context of flanking 5' and 3' structure cassettes. Templates were amplified by PCR and transcribed into RNA using T7 RNA polymerase (Wilkinson et al. 2006). RNAs were purified by denaturing polyacrylamide gel electrophoresis, appropriate regions excised, and RNAs passively eluted from the gel overnight at 4 °C. 16S and 23S rRNAs were isolated from DH5α cells during mid-log phase using non-denaturing conditions (Deigan et al. 2009). For each sample, 5 pM of RNA was refolded in 100 mM HEPES, pH 8.0, 100 mM NaCl, and 10 mM MgCl2 in a final volume of 10 μL. After folding, RNAs were modified in the presence of 10 mM SHAPE reagent and incubated at 37 °C for 3 min (1M6 and 1M7) or 22 min (NMIA). No-reagent controls, containing neat DMSO rather than SHAPE reagent, were performed in parallel. To account for sequence-specific biases in adduct detection, RNAs were modified using NMIA, 1M7, or 1M6 under strongly denaturing conditions in 50 mM HEPES (pH 8.0), 4 mM EDTA, and 50% formamide at 95 °C. Following modification, RNAs were isolated using either RNA affinity columns (RNeasy MinElute; Qiagen) or G-50 spin columns (GE Healthcare).

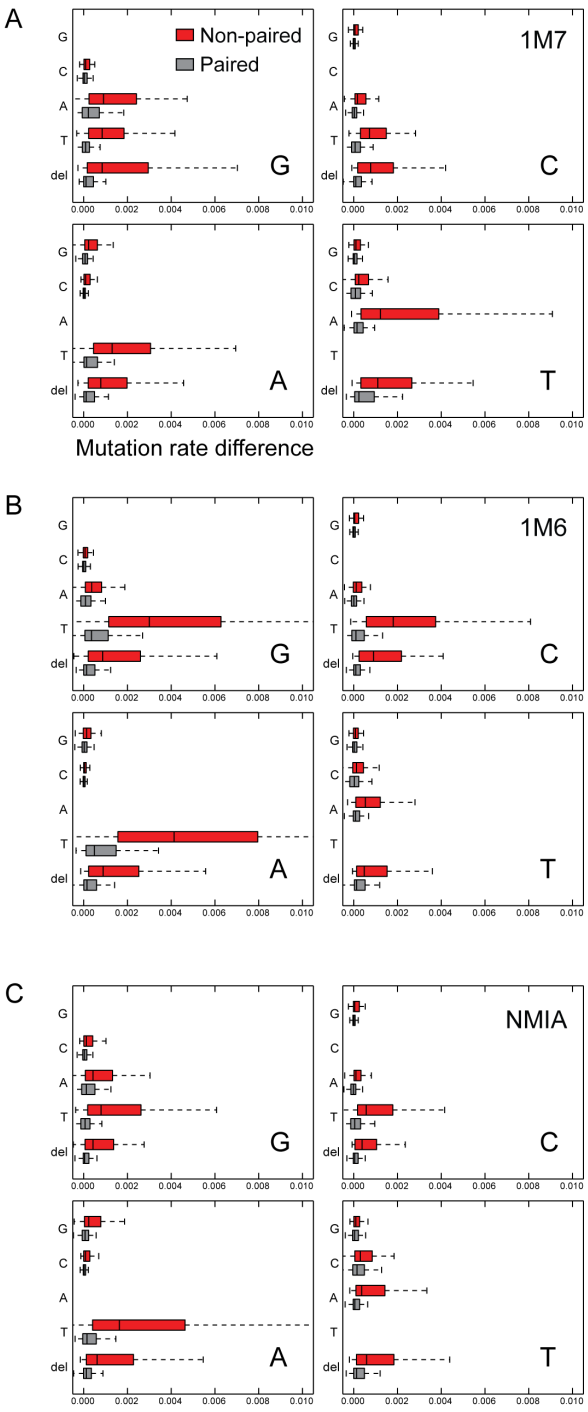

Figure 4.10: Detection of 2'-*O*-adducts by mutational profiling. Shown are rates for sequence changes and unambiguously aligned deletions, above background for the *E. coli* 16S rRNA. Nucleotides were defined as non-paired or paired based on the accepted secondary structure. The letter in the lower right of each panel indicates the expected nucleotide based on the coding strand, and the letters on the vertical axes indicate the nucleotide detected by sequencing or "del" for deletion. Rates are shown for the (a) 1M7, (b) 1M6, and (c) NMIA reagents. Nucleotide misincorporation and deletion rates were similar for the three SHAPE reagents.

SHAPE-MaP using fragmented samples.

Following SHAPE modification and purification, HIV-1, group II intron, HCV IRES, and ribosomal RNA samples were fragmented (yielding lengths of ~250-350 nts) by a 4 min incubation at 94 °C in a buffer containing 9 mM MgCl2, 225 mM KCl, 150 mM Tris HCl (pH 8.3). RNA fragments were desalted using G-50 spin-columns. Fragmented samples (250-500 ng total mass) were subjected to reverse transcription for 3 hours at 42 °C (using SuperScript II, Invitrogen). Reactions were primed using 200 ng random nonamer primers (NEB) for the ribosome, group II intron, and HCV IRES RNA or with custom LNA primers for HIV-1 RNA genomes. Reverse transcriptase buffer contained 0.7 mM premixed dNTPs, 50 mM Tris HCl (pH 8.0), 75 mM KCl, 6 mM MnCl2, and 14 mM DTT. Following reverse transcription, reactions were desalted using G-50 spin columns (GE Healthcare). Under these conditions (long incubation times and using 6 mM Mn2+ as the only divalent ion) the reverse transcriptase reads through sites of 2'-O-modification by a SHAPE reagent, incorporating a non-complementary nucleotide at the site of the adduct.

Double-stranded DNA libraries for massively parallel sequencing were generated using NEBNext sample preparation modules for Illumina. Second-strand synthesis (NEB E6111) of the cDNA library was performed using 100 ng input DNA, and the library was purified using a PureLink Micro PCR cleanup kit (Invitrogen K310250). End repair of the double-stranded DNA libraries was performed using the NEBNext End Repair Module (NEB E6050). Reaction volumes were adjusted to 100 μL, subjected to a cleanup step (Agencourt AMPure XP beads A63880, 1.6:1 beads-to-sample ratio), dA tailed (NEB E6053), and ligated with Illumina-compatible forked adapters (TruSeq) with a quick ligation module (NEB M2200). Emulsion PCR(Williams et al. 2006) (30 cycles) using Q5 hot-start, high-fidelity polymerase (NEB M0493) was performed to maintain library sample diversity. Resulting libraries were quantified (Qubit fluorimeter; Life Technologies), verified using a Bioanalyzer (Agilent), pooled, and subjected to sequencing using the Illumina MiSeq or HiSeq platform. Although only a single replicate of HIV-1, group II, and HCV IRES, fragmented RNAs

were performed, SHAPE reactivities agreed well for model RNAs with the known structure secondary structures (HCV IRES, and group II) and compare well with previously generated SHAPE-CE derived reactivities (Fig. 4.5). Additionally SHAPE reactivities of native sequence HIV-1 pseudoknotted regions probed using directed primers agree with randomly primed genomic reactivities.

SHAPE-MaP using targeted gene-specific primers

The tRNAPhe, TPP riboswitch, 5S rRNA, group I intron, and mutant HIV-1 construct RNAs were subjected to reverse transcription using a DNA primer specific to either the 3' structure cassette (5'-GAA CCG GAC CGA AGC CCG-3') for the small RNAs or to specific HIV-1 sequences flanking a pseudoknot using buffer and reaction conditions described in the previous section. Sequencing libraries were generated using a modular, targeted, two-step PCR approach that makes it possible to inexpensively and efficiently generate data for many different RNA targets. PCR reactions were performed using Q5 hot-start, high-fidelity DNA polymerase. The forward PCR primer (5'-GAC TGG AGT TCA GAC GTG TGC TCT TCC GATC NNNNN-gene-specific primer-3') includes an Illumina-specific region at the 5' end, followed by five random nucleotides to optimize cluster identification on the MiSeq instrument, and ends with a sequence complementary to the 5' end of the target RNA. The reverse primer (5'-CCC TAC ACG ACG CTC TTC CGA TCT NNNNN-gene-specific primer-3') includes an Illumina-specific region followed by five random nucleotides and a sequence that is the reverse complement of the 3' end of the target RNA. The cDNA library was 'tagged' by limited, 5-cycle PCR for amplicons or a longer 25 cycle PCR reaction when very low RNA concentrations were used. Excess primer, not used in the first few cycles, was removed (PureLink Micro PCR cleanup kit; Invitrogen). The second round of PCR added the remaining Illumina-specific sequences needed for on-flow cell amplification and barcoded the samples for multiplexing. The forward primer (CAA GCA GAA GAC GGC ATA CGA GAT [Barcode] GT GAC TGG AGT TCA GAC) contains a barcode and targets sequence in the forward primer from PCR 1.

The reverse primer (AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT T CCC TAC AC GAC GCT CTT CCG) contains an Illumina-specific sequence and targets the reverse primer from PCR 1. PCR 2 was performed for 25 or 5 cycles to generate the final library for sequencing (not exceeding 30 total cycles). Typical SHAPE-MaP experiments of mutant viruses used ~150 to 200 ng of RNA per experimental condition. However, when material is limiting, as little as 50 ng input RNA is sufficient.

Filtering by Z-factor for differential SHAPE data

SHAPE-MaP allows errors in SHAPE reactivity measurements to be estimated from a Poisson distribution describing the measured mutation rates at each nucleotide. The Poisson-estimated SHAPE reactivity error can be used to evaluate statistical significance when comparing two SHAPE signals. Significant differences between NMIA and 1M6 reactivity were identified using a Z-factor test (Zhang et al. 1999). This nucleotide-resolution test compares the absolute difference of the means with the associated measurement error:

$$Z_{factor} = 1 - \frac{3(\sigma_{NMIA} + \sigma_{1M6})}{|\mu_{NMIA} - \mu_{1M6}|} \quad (4)$$

Each nucleotide in a SHAPE-MaP experiment has a calculated reactivity μ and an associated standard error σ. The significance threshold for Z-factors was set at $Z > 0$, equivalent to a SHAPE reactivity difference for 1M6 and NMIA of at least three standard deviations. Differential nucleotide reactivities not meeting this significance criterion were set to zero.

Structure modeling

Secondary structure modeling for RNAs less than 700 nts in length was performed as described (Hajdin et al. 2013; Rice et al. 2014); differential SHAPE data were incorporated after filtering by Z-factor. For the HIV-1 RNA genome, we developed an automated windowed modeling approach, implemented in a *Superfold* Pipeline (Fig. 4.7), in which structure calculations were broken into stages designed to increase computational efficiency, generate realistic RNA structures, and

reduce end-effects caused by selecting a false 5' or 3' end from an internal fold in a window. This approach facilitated pseudoknot discovery, identification of probable base pairs, and generation of minimum free energy structures (Fig. 4.6). Representative calculations for folding of ribosomal subunits, performed using both one-step and windowed folding showed comparable, high degrees of accuracy and substantial reductions in computation 'wall time' using a typical desktop workstation for the windowed folding approach. For shorter RNAs, such as the 16S rRNA, there is a modest performance penalty for breaking the RNA into smaller windows. However, for RNAs longer than ~2000 nucleotides, computation time scales approximately linearly with length (Fig. 4.7).

Nearly all known and well-validated functional RNA structures are modeled identically in the current study and the prior 2009 investigation (Watts et al. 2009). Substantial improvements in digital (MaP) data acquisition, improved SHAPE-based energy functions (Hajdin et al. 2013; Rice et al. 2014) and automated data analysis (Figs. 4.5, 4.6 and 4.7) favor the the current second-generatation HIV-1 structure models over previous models in regions of disagreement. This work also reflects other innovations and analysis, notably that not all regions of an RNA are likely to form a single well-defined structure. As a result, an important component of the current work is the identification of regions in the HIV-1 RNA genome that do not form single well-defined structures.

Pseudoknot prediction. During the first stage, the full-length HIV-1 RNA genome was folded in 600-nt sliding windows moved in 100-nt increments using *ShapeKnots* with slope, intercept, P1, and P2 parameters set to previously defined values (1.8, -0.6, 0.35, 0.65) using 1M7- SHAPE data (Hajdin et al. 2013; Rice et al. 2014). Additional folds were computed at the ends of the genome to increase the number of windows that cover terminal sequences. Predicted pseudoknots were retained if the structure appeared in a majority of windows and had low SHAPE reactivity on both sides of the pseudoknotted helix. This list of pseudoknots was used for all later stages of modeling.

Partition function modeling. The partition function was calculated using *Partition* (Mathews 2004; Reuter and Mathews 2010) and included both 1M7 and differential SHAPE data in the free energy penalty. The max pairing distance was set to 500 nts. Partition was run in 1600-nt windows

with a step size of 375 nts. Two extra windows (lengths of 1550 and 1500 nts) were run on the 5' and 3' end sequences to increase sampling at the true ends and to reduce the effect of non-optimal cut site selection. Six sequences (the primer binding site, dimerization sequence, and four pseudoknots known to be involved in unusual or special interactions) were constrained as single stranded during partition function calculations. From the individual partition function files, the Shannon entropy of base pairing was calculated as:

$$H_i = -\sum_{j=1}^{J} p_{i,j} \log_{10} p_{i,j} \quad (5)$$

Where pi,j is the probability of pairing for nucleotides *i* and *j* over all potential J partners (Huynen et al. 1997). Following this calculation, 300 nts were trimmed from the 5' and 3' ends of each window that did not flank the true 5' and 3' ends of the RNA. This calculation retained more consistent internal values and discarded values skewed by end effects. Shannon entropy windows were combined by averaging, creating a single entropy file.

Individual probable pairs from each window were then trimmed using the same approach outlined for the Shannon entropy. Base pairs that formed with a probability less than 10^{-4} were removed to decrease computation time. Windows were combined, and all remaining pairs were averaged over all of the windows in which they could have appeared. A heuristic color scale was developed from the combined partition file to indicate relative likelihood of a pair appearing in the final structure. The resulting pairs were plotted as arcs (Fig. 4). Base pairs with a probability greater than 0.99 were used as double-stranded constraints in the next step.

Minimum free-energy modeling. A minimum free energy structure was generated using *Fold* (Reuter and Mathews 2010), 1M7 SHAPE data, and differential SHAPE data. A window size of 3,000 nts with a step size of 300 nts was used to generate potential structures over each window. Four folds (3100, 3050, 2950, and 2900 nts from the ends) were also generated to increase the number of structure models at the termini. These folds from overlapping windows were then combined into a complete structure by comparing base pairs common to each window and requiring that pairs in the

final structure appear in a majority of potential windows. As a final step, pseudoknotted helices were incorporated (Fig. 4.7).

Error analysis and determining a minimum number of reads required for accurate RNA structure modeling

The mutation rates for each of the contributing signals (SHAPE modified, untreated, denatured) were modeled using a Poisson distribution because discrete events from individual reads contribute to the overall signal. The variance of a Poisson distribution is equal to the number of observations; thus, the standard error of a 'true' rate can be modeled as:

$$\mathrm{SE}_{\mathrm{rate}} = \frac{\sqrt{\lambda}}{\mathrm{reads}} = \frac{\sqrt{\mathrm{rate}}}{\sqrt{\mathrm{reads}}} \qquad (6)$$

where λ is the number of events (mutations observed), *reads* is the read depth at the modeled nucleotide (both mutations and non-mutations), and *rate* is the number of events per read. As expected, bootstrapping of the standard error of SHAPE reactivity showed an x± power relationship as a function the read depth.

Using a deeply sequenced RNA (greater than 50,000 reads for each nucleotide), the number of expected mutation events at much lower read depths is known with high precision. Mutation events can be sampled from a Poisson distribution across the entire RNA to create profiles of plausible SHAPE data. To determine a minimum threshold for number of reads necessary for accurate SHAPE-directed secondary structure modeling, we examined the 16S rRNA because it is modeled poorly in the absence of experimental data (~50% sensitivity). For each simulated read depth, we created 100 SHAPE trajectories based on the expected Poisson variance at the simulated read depth and modeled it using RNAstructure *Fold* (Fig. 4.5b). As expected, modeling accuracy improved as read depth increased. For accurate nucleotide resolution structure modeling, we recommend at least 5000 reads; however, even at 500 reads, the measurement is useful for structure modeling (Fig. 4.5b).

Algorithmic discovery of HIV-1 regions with low Shannon entropy and low SHAPE reactivity

Overlaps of regions with both low SHAPE reactivity and low Shannon entropy were used to identify regions likely to have a single well-determined structure. First, local median SHAPE reactivity and Shannon entropy were calculated over centered sliding 55-nt windows. Next, we selected regions in which the local median fell below the global median for more than 40 nts in both Shannon entropy and SHAPE reactivity. Regions were combined if they were separated by fewer than 10 nts. Finally, regions were expanded to include nested secondary structures from the minimum predicted free-energy model.

To exclude the possibility that the algorithmically discovered structured regions overlapped known RNA elements merely by chance, we generated a randomized pool of segments and calculated the expected distribution of overlapping nucleotides. We maintained the same number and length of segments but randomized their locations within the 9173-nt genome. Out of 10^5 trials, only 219 showed a larger overlap than we observed, corresponding to a p-value of 0.002.

HIV competition assays

Mutant and native sequence virus were mixed at a 10:1 ratio, respectively, and used to infect 5 × 105 Jurkat cells in 1 mL total volume in 12-well plates. Infections were performed using half as much mutant and 20-fold less wild-type virus relative to the viral replication assays. Competition experiments were carried out in duplicate. Medium was initially harvested at 2 dpi to represent the initial inoculum. The medium was harvested at 3, 4, 5, and 6 dpi, and p24 (capsid protein) was quantified in medium (AlphaLISA HIV p24 kit). We required that p24 levels increase exponentially through day 6 to ensure that uninfected cells were in excess through the infections. Viral RNA was purified from medium (QIAamp viral RNA mini kit, Qiagen) and reverse transcription using SuperScript III (Life Technologies) was carried out using Primer ID primers(Jabara et al. 2011) to barcode each cDNA produced and eliminate population biases introduced during PCR. Subsequent

sample preparation was performed as described above for SHAPE-MaP using targeted gene-specific primers.

After sequencing, paired-end reads were merged into longer synthetic reads using FLASh (Fast Length Adjustment of Short reads)(Magoč and Salzberg 2011). Next, synthetic reads were aligned to the expected NL4-3 sequence for the targeted regions using *Bowtie2* (Staple et al. 2012) (using default parameters). A consensus read was built for each PrimerID based on a Phred score voting metric. IDs matching either native or mutant sequences were required to have the expected point mutations in all locations in order to be considered. The fraction of mutant IDs was expressed as the number of mutant IDs out of the sum of mutant and native IDs. Relative fitness of mutant viruses was determined from the rate of change of the ratio of mutant to NL4-3 measured over time (Resch et al. 2002).

Calculation of differences in SHAPE reactivities in pseudoknot mutants

Standard error measurements of SHAPE reactivities, estimated from the Poisson distribution, are dependent on the number of reads obtained for each sample. The observation that standard error decreases with the inverse square of read depth was used to derive a scaling equation that normalizes to a common depth of 8000 reads to account for differences in sequencing depth between samples. The standard error scaling factor, f_0, was calculated for each sample based on the average read depth, rave, of the lowest sequenced component (SHAPE modified, untreated, and denaturing conditions) contributing to the SHAPE reactivity profile:

$$f_0 = \frac{(r_{ave})^{\frac{1}{2}}}{(8000)^{\frac{1}{2}}} \quad (8)$$

After scaling standard errors to a common read depth, significance for each point was calculated using a modified z-factor test (Zhang et al. 1999) requiring differences to be greater than 1.96 times the sum of the standard errors. Z-factor scores greater than zero were considered significant:

$$Z_{factor} = 1 - \frac{1.96(\sigma_{PK}+\sigma_{WT})}{|\mu_{PK}-\mu_{WT}|} \quad (9)$$

Isolated reactivity changes can be viewed as noise in the context of a global structure shift resulting from disruption of a pseudoknot. Therefore, in addition to the z-factor test, differences were required to be consecutive.

REFERENCES

Brierley I, Pennell S, Gilbert RJC. 2007. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Micro* **5**: 598–610.

Chen T, Romesberg FE. 2014. Directed polymerase evolution. *FEBS Lett* **588**: 219–229.

Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106**: 97–102.

Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM. 2012. Functional complexity and regulation through RNA dynamics. *Nature* **482**: 322–330.

Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**: 696–700.

Doty P, Boedtker H, Fresco JR, Haselkorn R, Litt M. 1959. Secondary structure in ribonucleic acids. *Proc Natl Acad Sci* **45**: 482–499.

Ghadessy FJ, Holliger P. 2007. Compartmentalized self-replication: a novel method for the directed evolution of polymerases and other enzymes. *Methods Mol Biol* **352**: 237–248.

Gherghe C, Lombo T, Leonard CW, Datta SAK, Bess JW, Gorelick RJ, Rein A, Weeks KM. 2010. Definition of a high-affinity Gag recognition structure mediating packaging of a retroviral RNA genome. *Proc Natl Acad Sci* **107**: 19248–19253.

Gilmartin GM, Fleming ES, Oetjen J. 1992. Activation of HIV-1 pre-mRNA 3' processing in vitro requires both an upstream element and TAR. *EMBO J* **11**: 4419–4428.

Grohman JK, Gorelick RJ, Lickwar CR, Lieb JD, Bower BD, Znosko BM, Weeks KM. 2013. A guanosine-centric mechanism for RNA chaperone function. *Science* **340**: 190–195.

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci* **110**: 5498–5503.

Huynen M, Gutell R, Konings D. 1997. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol* **267**: 1104–1112.

Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci* **108**: 20166–20171.

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**: 103–107.

Klasens BI, Thiesen M, Virtanen A, Berkhout B. 1999. The ability of the HIV-1 AAUAAA signal to bind polyadenylation factors is controlled by local RNA structure. *Nucleic Acids Res* **27**: 446–454.

Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA, Arkin AP. 2011. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci* **108**: 11063–11068.

Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10**: 1178–1190.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101**: 7287–7292.

Matoulkova E, Michalova E, Vojtesek B, Hrstka R. 2012. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol* **9**: 563–576.

Mauger DM, Weeks KM. 2010. Toward global RNA structure analysis. *Nat Biotechnol* **28**: 1178–1179.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**: 1105–1119.

McGinnis JL, Weeks KM. 2014. Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry* **53**: 3237–3247.

Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* **127**: 4223–4231.

Mortimer SA, Weeks KM. 2007. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc* **129**: 4144–4145.

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

Pollom E, Dang KK, Potter EL, Gorelick RJ, Burch CL, Weeks KM, Swanstrom R. 2013. Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs. *PLoS Pathog* **9**: e1003294.

Resch W, Ziermann R, Parkin N, Gamarnik A, Swanstrom R. 2002. Nelfinavir-resistant, amprenavir-hypersusceptible strains of human immunodeficiency virus type 1 carrying an N88S mutation in protease have reduced infectivity, reduced replication capacity, and reduced fitness and process the Gag polyprotein precursor aberrantly. *J Virol* **76**: 8659–8666.

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf* **11**: 129.

Rice GM, Leonard CW, Weeks KM. 2014. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA* **20**: 846–854.

Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**: 701–705.

Sharp PA. 2009. The centrality of RNA. *Cell* **136**: 577–580.

Spitale RC, Crisalli P, Flynn RA, Torre EA, Kool ET, Chang HY. 2013. RNA SHAPE analysis in living cells. *Nat Chem Biol* **9**: 18–20.

Staple DW, Butcher SE. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **3**: e213.

Staple DW, Langmead B, Langmead B, Butcher SE, Salzberg SL, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**: 357–359.

Steen K-A, Rice GM, Weeks KM. 2012. Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *J Am Chem Soc* **134**: 13160–13163.

Tyrrell J, McGinnis JL, Weeks KM, Pielak GJ. 2013. The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry* **52**: 8777–8785.

Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, Haussler D. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Meth* **7**: 995–1001.

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**: 711–716.

Weeks KM. 2010. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol* **20**: 295–304.

Weeks KM. 2011. RNA structure probing dash seq. *Proc Natl Acad Sci* **108**: 10933–10934.

Weeks KM, Mauger DM. 2011. Exploring RNA structural codes with SHAPE chemistry. *Acc Chem Res* **44**: 1280–1291.

Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, Mathews DH, Giddings MC, Weeks KM. 2008. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol* **6**: e96.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1**: 1610–1616.

Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD. 2006. Amplification of complex gene libraries by emulsion PCR. *Nat Meth* **3**: 545–550.

Zhang J, Chung T, Oldenburg K. 1999. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen* **4**: 67–73.

CHAPTER 5: RIBOSOME DYNAMICS VISUALIZED BY CORRELATED CHEMICAL PROBING IN LIVING CELLS

Introduction

Messenger RNAs (mRNAs) are translated into proteins by the hybrid RNA-protein complex called the ribosome. The complete bacterial ribosome (70S) is composed of the small and large subunits (30S and 50S, respectively) that have a combined mass of ~2.5 MDa (Frank and Gonzalez 2010; Noeske and Cate 2012). Prior to translation, the ribosomal RNA must properly fold and assemble with many protein partners. In order to accomplish translation, the ribosome must adopt several distinct structural conformations. Through great effort, the structures of several intermediate states of translation have been revealed using Cryo-EM and X-ray crystallography (Zhang et al. 2009; Dunkle et al. 2011; Agirrezabala et al. 2012). These experiments were performed with ribosomes removed from their cellular context, potentially obscuring interactions that take place in cells (Tyrrell et al. 2013). I sought to examine the structural conformations of the ribosome in living cells by taking advantage of chemical probing experiments described in previous chapters. Dimethyl sulfate (DMS) reacts at unpaired adenine and cytosine nucleotides to form adducts at the base pairing face (Fig. 5.1a) (Ehresmann et al. 1987). As DMS is able to penetrate the cellular membrane, it can provide a structural snapshot of RNA base pairing inside living cells (Ding et al. 2014).

When high DMS concentrations are used, multiple modifications are made to an individual RNA strand. Since massively parallel sequencing is able to report on the sequences of single RNA template, massively parallel sequencing is in effect a single-molecule experiment (Bentley et al. 2008). To identify the sites of reaction of structure-selective reagents such as DMS, I used a mutational profiling (MaP) readout approach during reverse transcription. Using this method, the

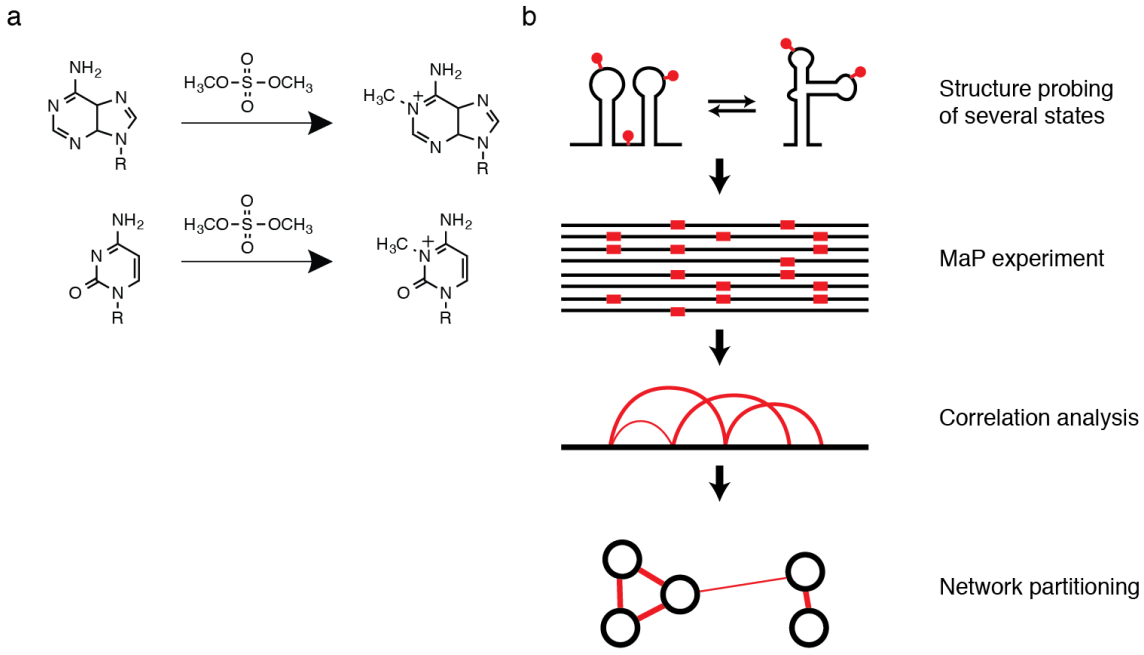

Figure 5.1: Overview of DMS modification and RING-MaP experiment. (A) Reaction of DMS with adenine and cytosine nucleobases showing the formation of a covalent adduct on the base-pairing face. (B) Overview of the RING-MaP experiment. In the first step, RNA (which may be an ensemble of multiple conformations) is modified with a structure-selective reagent. Next, the location of adducts is incorporated into the cDNA as mutations during reverse transcription. Following library construction and sequencing, correlations are detected from mutations occurring within the same sequencing read. Last, a network analysis is applied to the correlations, separating nucleotides into communities based on the interconnectedness of correlations.

location of chemical adducts on RNA is stored as a mutation in the growing cDNA (Siegfried et al. 2014). Using the MaP approach these events can be visualized in the same massively parallel sequencing read. The presence of multiple events on the same read coupled with statistical analysis allows the discovery of correlated mutation events in an experiment called RING-MaP (RNA interacting groups analyzed by mutational profiling) (Homan et al. 2014). The RING-MaP experiment provides information on the through-space interactions of nucleotides that are modified simultaneously.

Despite the power of the RING experiment to obtain structural information, it has several limitations. The first version of the RING-MaP experiment required the entire length of the RNA of interest be sequenced in a single read, preventing the RING analysis from being applied to RNAs longer than ~500 nucleotides in length. Additionally, after DMS probing, the RNA is heavily modified with adducts that hinder the processivity of the reverse transcription enzyme. Here I extend the RING-MaP experiment for use with random priming by optimizing both the experimental protocol and the computational analysis in order to characterize the structural states of ribosomes within the context of living cells and apply network analysis tools to uncover structural communication within the small subunit of the ribosome (Fig. 5.1b).

Results

Multisite dimethyl sulfate reactivity of the ribosome in distinct structural conformations

Two antibiotics were chosen to perturb the structural states of the ribosome in cells: rifampicin and spectinomycin. The antibiotic rifampicin is a transcription inhibitor that binds to RNA polymerase and prevents the production of new RNA. Rifampicin has been shown to bias the steady state structure of the 30S subunit in cells to that of the fully assembled state (McGinnis and Weeks 2014). Spectinomycin binds the small ribosomal subunit in a single location at helix 34 in the head

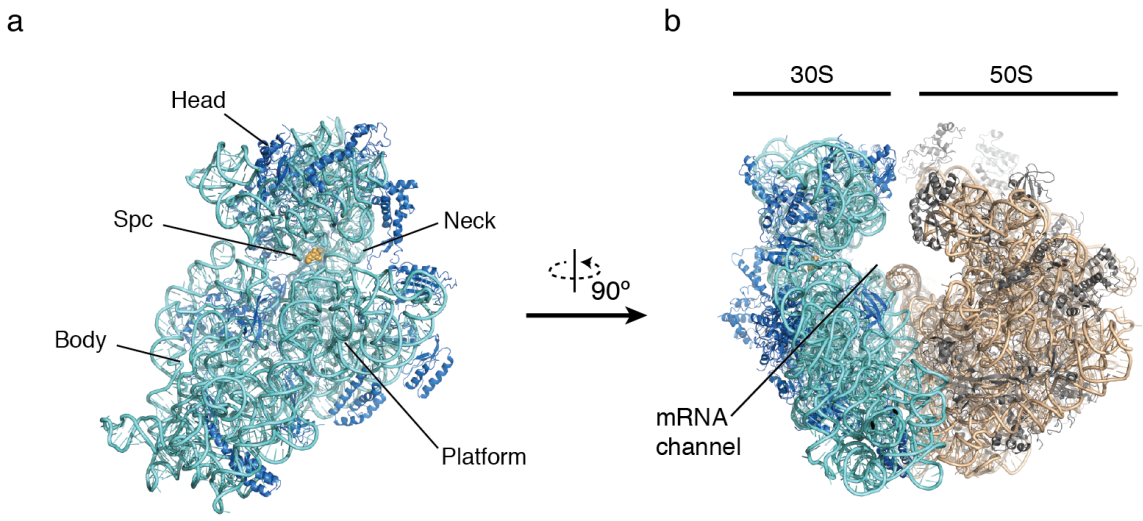

Figure 5.2: Structural organization of the bacterial ribosome. (A) Small subunit (30S) visualized through the large ribosomal subunit. RNA is depicted in light blue with proteins in dark blue. Regions of the small subunit are labeled according to their location. The antibiotic spectinomycin (yellow) is shown binding to the bottom of the head. (B) Organization of the small and large subunits.

domain (Fig 5.2) (Borovinskaya et al. 2007) trapping it in an intermediate state of rotation (Mohan et al. 2014). In my work, I treated cells with rifampicin prior to spectinomycin treatment in order to ensure that the ribosome targeting antibiotic treatment affected fully assembled ribosomes and not assembly intermediates.

Intact cells were treated with DMS to probe the structure of ribosomal RNA (rRNA) under native (*in vivo*), transcription stopped (+Rif), and transcription and translation stopped (+Rif/+Spc) conditions. A gently deproteinized (*ex vivo*) sample of rRNA was also evaluated. Using conditions first developed for RING-MaP, reverse transcription was extremely inefficient with heavily modified RNA. Using the highly modified *ex vivo* RNA, new reverse transcription conditions were screened in order to optimize complementary DNA (cDNA) yield and product length from mutation profiling. High concentrations of betaine, coupled with changes to the primer annealing protocol and temperature cycling, increased the average length of the reverse transcription products by 20% and the total cDNA yield by 4.5 fold. The shift to larger products and the increase in total yield allowed for the enrichment of larger inserts (and thus more chances to observe synchronous modification events) in the final sequencing libraries.

Following deep sequencing and read alignment, mutation rates were obtained for each of the samples (Fig. 5.3). The rRNA from the *ex vivo* sample was much more highly modified than were other samples (Fig. 5.3a). This observation is consistent with the fact that ribosomal proteins stabilize rRNA structure *in vivo*. DMS mutation rates at many positions in the *ex vivo* rRNA sample were above 5%, indicating that each read likely contained many structurally informative mutations. The *ex vivo* mutation rates were the most different from those of RNA probed *in vivo* (Pearson's R coefficient of approximately 0.44, Fig. 5.3b). Mutation rates of all in-cell samples were similar to each other, with Pearson coefficients above 0.99. Comparing the mutation rates +Rif to +Rif/+Spc revealed protection at C1192, consistent with spectinomycin binding and protection at helix 34 (Fig. 5.3c).

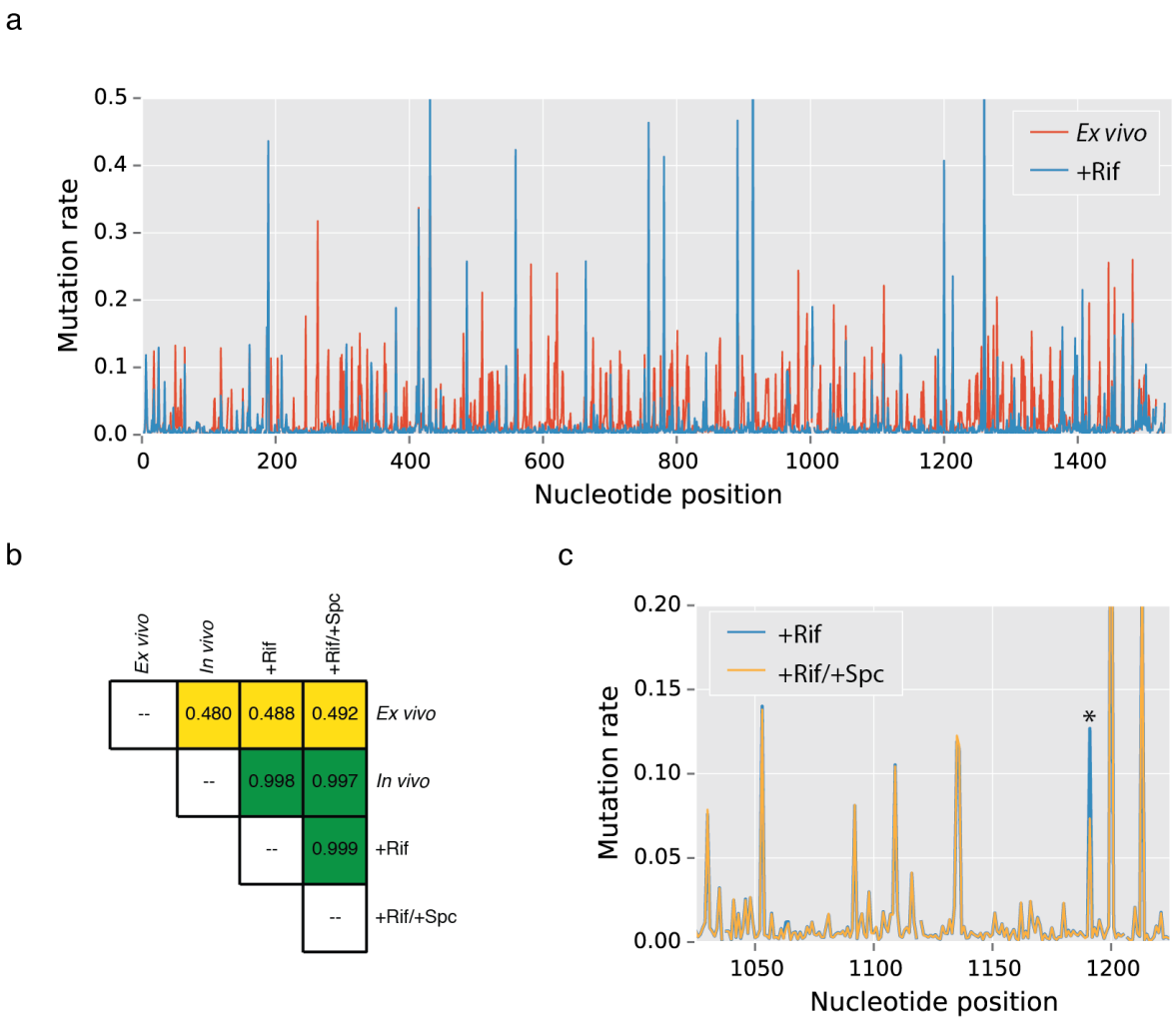

Figure 5.3: DMS reactivity across the small ribosomal subunit. (A) Comparison of the *ex vivo* (protein free) state to the in-cell, +Rif state. The overall reactivity of the *ex vivo* state is much higher across the length of the small subunit RNA. (B) Pairwise Pearson R correlations show that the *ex vivo* state is very different compared to that of RNA in ribosomes probed in cells. All in-cell probed states (*in vivo*, +Rif, and +Rif/+Spc) are highly similar with correlation values above 0.99. (C) Addition of spectinomycin results in protection at the expected nucleotides (asterisk at C1192). Other known spectinomycin-interacting nucleotides were either already unreactive or were not detectable using DMS probing.

Development of a general analysis framework for randomly primed reads.

The previous RING analysis approach required that all sequencing reads be stored in computer memory in order to perform association analyses. This approach is appropriate for small RNAs, but it rapidly becomes impractical as the length of the RNA and the number of available reads increase. Rather than retaining all sequencing reads in memory, it is possible to create a simplified representation of alignment and mutation location information (Fig 5.4a). This representation is a two-dimensional array with each element containing a contingency table of the counts and kinds of interactions (Fig 5.4b). Information about pairwise observations of mutations within each read can thus be counted and stored independently. Using this representation, the total amount of memory needed for analysis depends only on the RNA length and not on the number of sequencing reads. During analysis of sequencing data, only reads that meet Phred quality cutoffs are included. Since interactions are stored as *i-j* interaction pairs, long stretches of incomplete information (i.e., gaps between sequencing reads relative to the reference sequence) are allowed, with each *i-j* point in the matrix representing the contingency table for all reads that contained both nucleotides *i* and *j*. Using the contingency table, a Yates chi-squared statistic and Pearson correlation (phi) can be calculated (Fig 5.4c). Correlations with Yates chi-square values above 20 are considered significant. Based on this threshold for chi-squared statistics, the probability that correlated nucleotides are independent is less than 0.00001.

In paired-end sequencing, both ends of the DNA library are sequenced even though they may be separated by several hundred nucleotides. Modern sequencing platforms keep paired reads "together," effectively allowing detection of interactions at a distance up to the size of the inserts of the sequencing library. In this work, large DNA fragments were selected; sequencing libraries were constructed with insert sizes between 500 and 700 nucleotides. Approximately 50,000 reads between

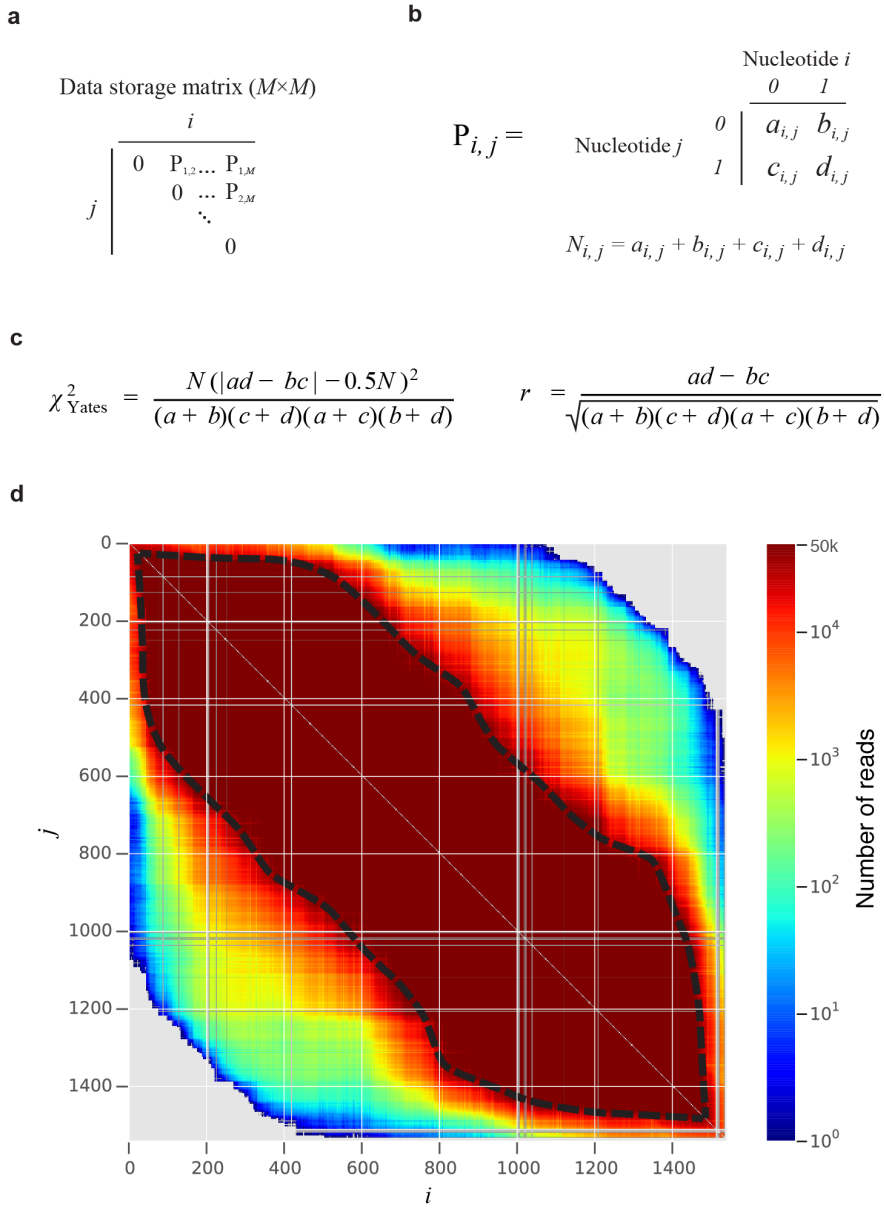

Figure 5.4: Computational approach for analyzing randomly primed read data. (A) A data storage matrix is used to store counts of within-read interactions. The matrix has a size equal to the length of the aligned RNA sequence. The index values *i* and *j* are the compared positions within the RNA. (B) Each element (P) of the data storage matrix contains a contingency table of counts across the entire data set for the number of times nucleotide *i* and *j* are not mutated together ($a_{i,j}$), are mutated together ($d_{i,j}$), or one is mutated but not the other ($b_{i,j}$ and $c_{i,j}$) within the same read. The total number of times nucleotide *i* and *j* are read together ($N_{i,j}$) is the sum of all the elements in the contingency table. (C) After read counting is performed, the significance of each interaction is tested using the χ^2_{Yates} test, and the strength of the interaction is measured using the Pearson's R metric. (D) Representative example of the pairwise read depth (*N*) for a single replicate of the +Rif samples. Colors, shown on a log scale, represent the number of times each pairwise interaction occurred. The scale is clipped at 50,000 reads showing approximately the minimum number of reads for robust measurement of correlations. The region contained within the dashed black line shows the scope of detectable interactions. Other samples (not shown) have similar pairwise read depths.

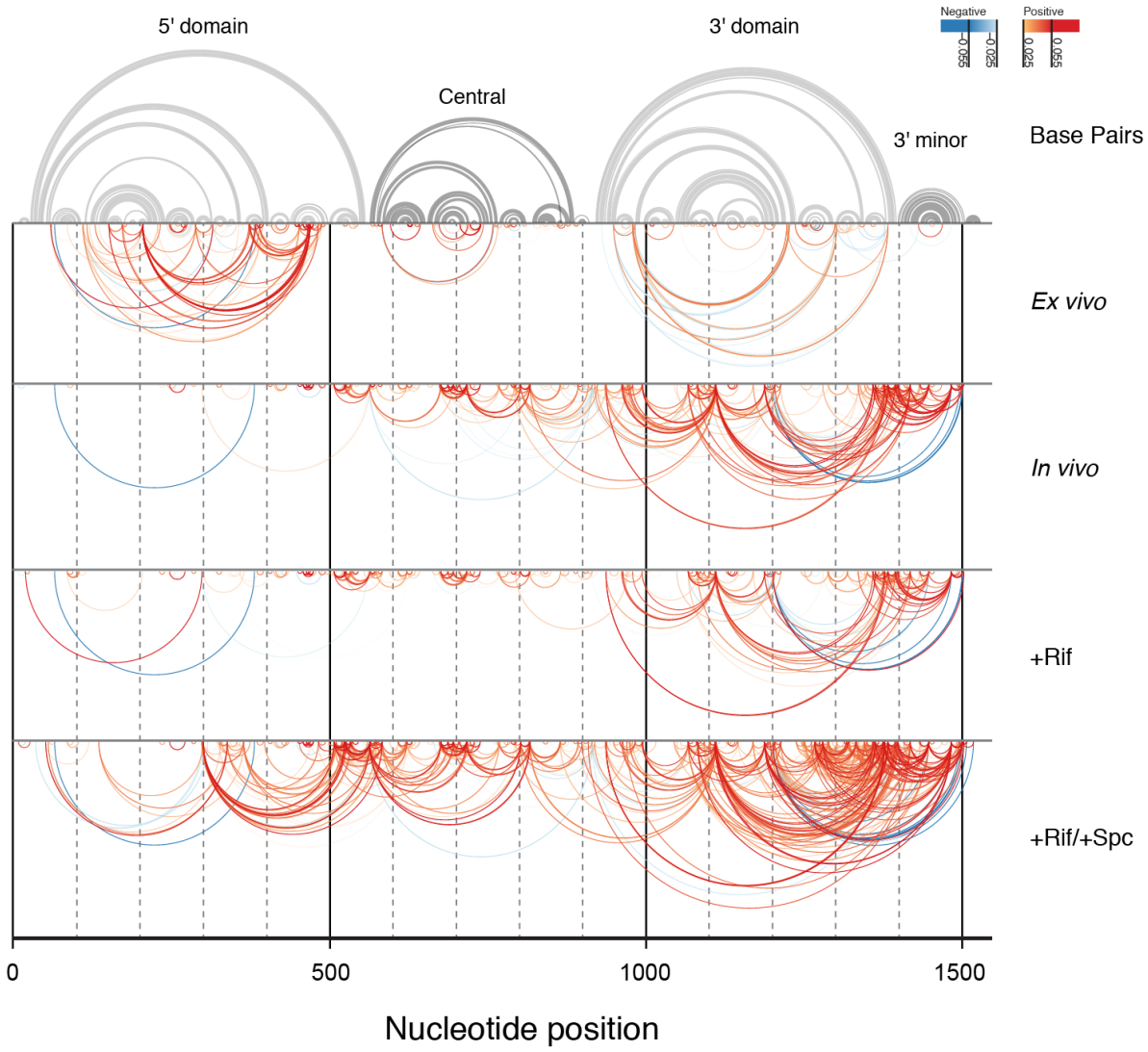

Figure 5.5: RING-MaP correlations within the small ribosomal subunit. Base pairs present in the structure established by covariation analysis are shown as gray arcs in the top panel. Below the base pairs, correlations for each of the merged conditions are shown as blue (negative correlation) and red (positive correlation) arcs. Correlations in the *ex vivo* sample are most dense within the 5' domain; correlations in the central and 3' domains in the *ex vivo* sample mirror the secondary structure of the small subunit. The *in vivo* and +Rif samples show similar patterns of correlations, most concentrated at the 3' end of the small subunit. Upon addition of spectinomycin to rifampicin-treated cells (+Rif/+Spc), the number and strength of correlations increases and spreads further into the 5' and central domains.

two locations of the RNA are needed to reliably detect correlated interactions. At this sequencing depth, for each of our samples, we can reliably detect interactions between 450 and 650 nucleotides in sequence space, with the specific number depending on biases in random priming (Fig 5.4d).

Correlated chemical probing reveals distinct structural networks

The improved correlation analysis was applied to each of the ribosomal data sets. Correlations between biological replicates were high, but some correlations were present in only a single data set. In order to increase the accuracy, correlation networks in biological replicates were analyzed separately and then merged into a single representative sample; the merged correlation coefficients are the average of the two measurements.

Analysis of the correlations in the *ex vivo* condition revealed a large number of interactions within the 5' domain of the small subunit (Fig 5.5). Many of the correlations in the central and 3' domains mirrored the known secondary structure. In contrast, correlations observed between nucleotides modified *in vivo* were shifted to the 3' end of the molecule, with a large number of the correlations centered at helix 44. The correlation networks of the *in vivo* and +Rif samples were highly similar. In the spectinomycin-containing sample (+Rif/+Spc), both the numbers and strengths of the interactions increased significantly when compared to the other in-cell states. Correlations at the 3' end observed in the +Rif state and a large number of correlations that span the central and 5' domain were observed in the +Rif/+Spc state.

Network analysis reveals distinct communities in the small subunit with structural hubs

Many of the correlations appear to be linked in the +Rif and +Rif/+Spc states, with one nucleotide interacting with several other nucleotides in the RNA. We hypothesize that these linkages represent several conformational states present as an ensemble of RNA structures. In order to test this hypothesis, we treated the RNA and correlations as a network graph, with nucleotides representing nodes and correlations representing edges connecting nodes. Nucleotides included in the graph were

connected to at least three other nucleotides with correlation strengths greater than 0.015. Modularity, a metric that measures the interconnectivity of nodes, was used to separate nucleotides into communities – distinct groups of highly connected nucleotides (Fig. 5.6). The number of connections made by each nucleotide is reflected in the size of the node (circle), with larger nodes representing more connections. The thickness of the edge (line connecting two nodes) indicates the strength of the correlation connecting two nucleotides, with thicker lines representing stronger correlations. In the +Rif sample, ten communities were detected; six of these communities include the majority of the interactions (Fig 5.6a). The overall structure of the network of the +Rif/+Spc sample is similar to that of the +Rif sample; however, the strengths and numbers of interactions increased and several of the communities merged to form a total of five communities (Fig 5.6b).

The read depth is sufficient to detect interactions between communities, and the flow of structural information goes through specific paths. This observation is especially true in the +Rif state. Nucleotides comprising the light blue community are closer in sequence space to those in the red community, yet are connected through the dark blue community. Several nucleotides have a large number of connections to other nucleotides in the network; these nucleotides represent hubs in the correlation network. In both the +Rif and +Rif/Spc networks, A1111 is a central hub with a large number of both in-community and out-community connections. Additionally, A1111 appears to be a lynchpin nucleotide that is representative of its entire community (Fig. 5.6a and b, red cluster). Interactions bridging communities may be of special interest, since these nucleotides may be involved in structural switches in an RNA.

When the +Rif/+Spc communities are mapped onto the accepted secondary structure for the small subunit, they do not fall into the canonical ribosome domains (Fig. 5.7, colored nucleotides). The green community has nucleotides located in both the 5' and central domains. Similarly, the blue community spans the 3' and 3' minor domains. The strength of the out of network linkages connecting communities is also not uniform (Fig. 5.7, colored lines). The strength and number of linkages

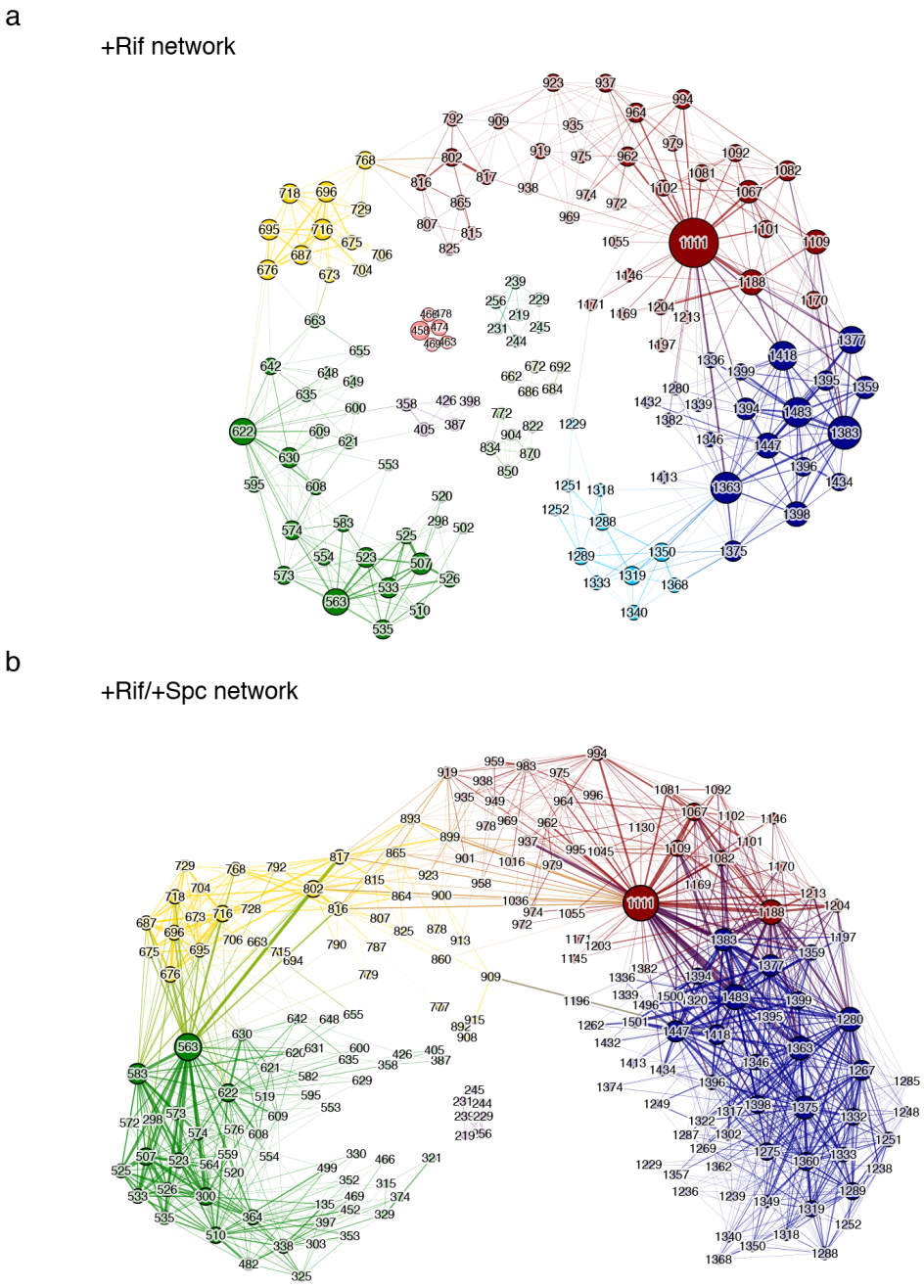

Figure 5.6 Correlation network diagrams separated into communities. Nucleotide positions were treated as nodes and correlations were treated as edges to create a network model of interactions. The relative size of the circle number indicates the number of correlations a nucleotide has with others in the network. The weight of the line connecting nucleotides indicates the strength of the correlation. Network modularity was used to separate nucleotide groups into communities based on connections between nodes. (A) Nucleotides in the +Rif sample separate into ten distinct communities with the most inter-connected communities made up of nucleotides in the 3' domain of the ribosome. (B) Addition of spectinomycin to rifampicin treated cells (+Rif/+Spc) increases the strength and number of correlations. Overall the network is more highly connected and fewer communities are detected. Strongly connected nucleotides such as 1111 are important in both networks.

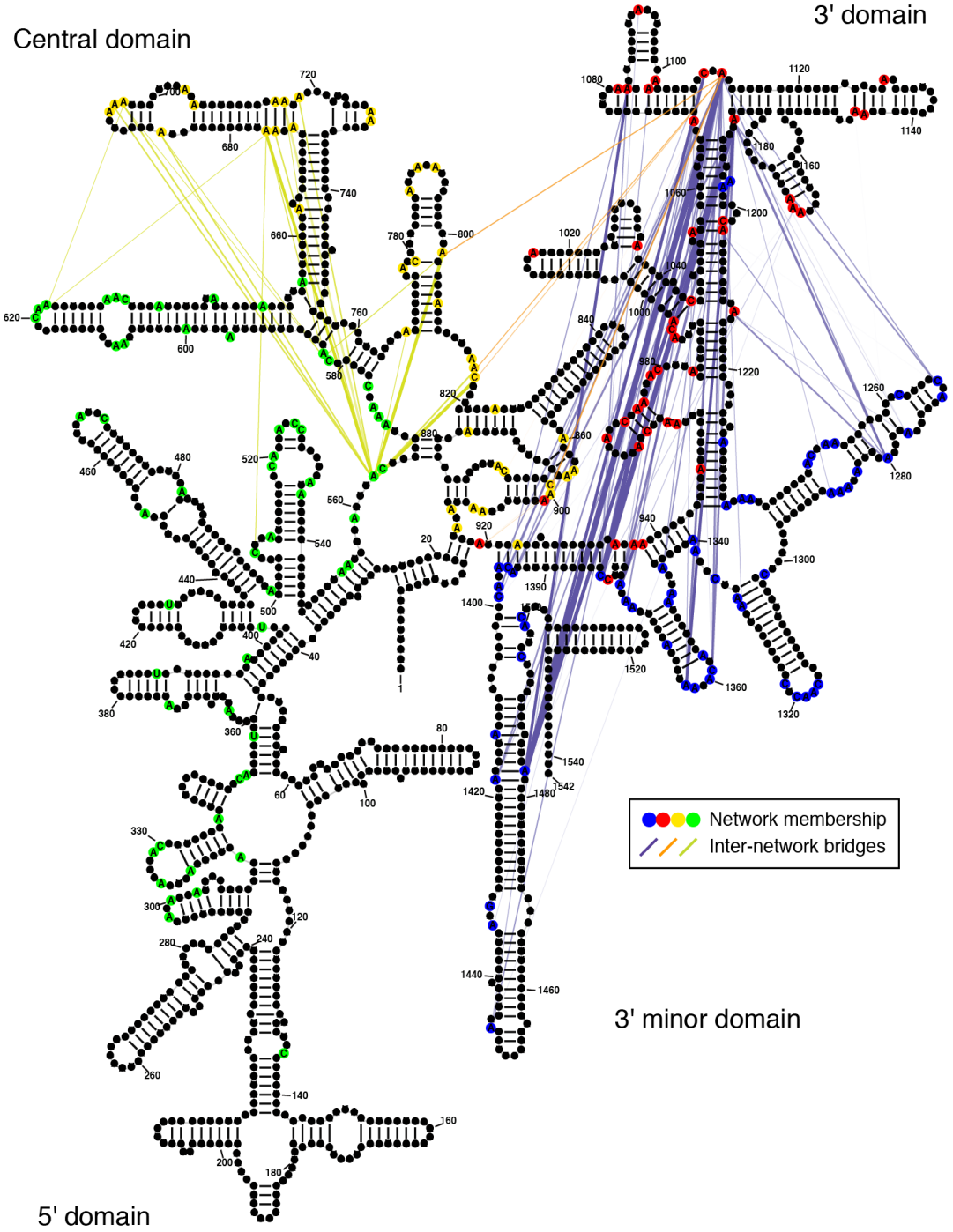

Figure 5.7 Bridging correlations connecting communities displayed on the small subunit for the rifampicin- and spectinomycin-treated ribosomes. A colored circle behind the nucleotide letter indicates community membership. Bridging correlations are indicated using lines between community members, the weight of the line indicates the strength of the interaction. Interactions within the same community are not shown for clarity. Communities do not span the canonical domains of the ribosome. Red and blue communities are highly connected (purple lines).

between the blue and red communities suggest a high degree of structural coordination. This is similarly true between the green and yellow communities. Yet there are few connections between the yellow and red communities, indicating structural independence.

Discussion

In the absence of proteins the correlation network is localized. There are a large number of interactions between nucleotides within the 5' domain (Fig. 5.5). This domain has been previously reported to fold in the absence of proteins into a quasi-native structure at high magnesium concentrations (Adilakshmi et al. 2005; 2008). Correlations in this domain map to known tertiary and secondary structure interactions. The RINGs seen in the 5' domain are consistent with those seen for structured small RNAs (Homan et al. 2014). In comparison the RINGs observed in the central and 3' domains match the secondary structure well. The lack of tertiary structure indicating RINGs in this region indicates that only the secondary structure likely forms in the 3' domain in the absence of protein.

In cells the small subunit of the ribosome is in a constant state of structural flux. It switches among confirmations present during translation and in the free 30S subunits. The presence of multiple structural states is likely to dilute the correlation analysis signal below significance. Since the antibiotic spectinomycin is thought to "trap" the head in a rotated conformation (Borovinskaya et al. 2007), the number of potential structural states that the ribosome can sample is reduced in the presence of this antibiotic. This reduction in competing structural states has the effect of increasing the strength and number of interactions that are detected in the +Rif/+Spc sample. Analysis of the structure of the ribosome trapped by binding of elongation factor G revealed the mechanism of 30S subunit head rotation (Mohan et al. 2014). Two helices were implicated as hinges involved in this rotation: h28 and nucleotides located in the junction above h34. The line drawn by these two hinges directly mirrors the strongest correlations connecting the red community at the top of h34 to blue-community nucleotides in h28 and h44 (Fig. 5.7). Several nucleotides such as A1111, A1188, and

C1383 act as network hubs and are central to many of the correlations observed in both the +Rif and +Rif/+Spc structural states (Fig. 5.6). These hub nucleotides could be key structural switches that control head rotation motions.

The extension of single-molecule chemical probing of RNA with random priming represents significant advance in detecting RNA interactions at moderate distance scales (400-600 nucleotides) within long RNAs. In the previous iteration of the RING-MaP experiment it was necessary to specifically target interesting regions limiting the approach to regions less than 400 nucleotides in length. Technical improvements were achieved in both the reverse transcription processivity and computational framework that now enable RING analysis on relatively large RNAs. In the evaluated rRNA, many of the interactions in naked RNA appear to correlate with secondary structure. This could provide a second level of experimental evidence for validating base pairs in RNAs that has not been possible before without the need to make tedious mutations or work with *in vitro* transcripts (Kladwang et al. 2011). With this advance it will be possible to quickly screen viral RNAs and other RNAs present in living cells for functionally important structural dynamics.

Methods

DMS modification of extracted ribosomal RNA

16S and 23S rRNAs (the RNA component of the 30S and 50S, subunits, respectively) were isolated from K12 MG1655 cells during mid-log phase (O.D. = 0.6) using non-denaturing conditions (Deigan et al. 2009). RNA was exchanged using a gravity-flow PD-10 sephadex column (GE Healthcare) into a folding buffer containing 300 mM cacodylate, pH 7.0, 200 mM potassium acetate, pH 7.0, and 10 mM $MgCl_2$ and incubated at 37 °C for 20 minutes.

Antibiotic treatment for in cell samples

In two lots, 2 mL of overnight culture was added to 48 mL of LB. Cells were incubated with shaking until the culture reached OD_{600} ~0.5 (approx. 30 min). To each culture, 5.55 mL of 10X

rifampicin (187.5 μg/mL = 10X) was added, and cells were incubated with shaking for 10 minutes. Following incubation, 27 mL of each culture was transferred to a new culture flask. To each culture, either 3 mL of water or 3 mL of 10X spectinomycin (494 μg/mL = 10X) was added. Cultures were allowed to incubate with shaking for 10 minutes. Cells were pelleted in 25 mL aliquots at 4000 *g* for 20 minutes. Supernatant was discarded, and the cell pellet was resuspended in 200 μL of folding buffer containing 300 mM cacodylate, pH 7.0, 200 mM potassium acetate, pH 7.0, and 10 mM $MgCl_2$ and incubated at 37 °C for 5 min.

Dimethyl sulfate treatment and purification of ribosomal RNAs

DMS was diluted in neat ethanol 1:5 to create a working DMS mixture. An aliquot of 90 μL of folded ribosome samples was added to 10 μL of working DMS mixture [(+) reaction] or 10 μL neat ethanol [(–) reaction] and incubated at 37 °C for 6 minutes. Following incubation, an equal volume (100 μL) of neat 2-mercaptoethanol was added to quench the DMS.

To each reaction, 1 mL of Trizol reagent (Invitrogen) was added and the reaction tubes were incubated at room temperature. After 5 minutes, 200 μL of cold chloroform was added, and tubes were shaken vigorously by hand for 15 seconds. Samples were incubated at room temperature for 2-3 minutes. Tubes were centrifuged at 12,000 *g* for 15 min at 4 °C. The aqueous upper layer was transferred to a new tube, and a 1.1X volume of isopropanol was added. Reactions were incubated at -20 °C for 30 minutes and then centrifuged at 15,000 *g* for 30 minutes at 4 °C. The supernatant was discarded, and pellets were carefully washed twice with 500 μL 80% ethanol, centrifuging five minutes at 15,000 *g* between washes. Following the washes, the supernatant was discarded, and pellets were air-dried for 5 minutes. Samples were then treated with DNaseI (Ambion) according to the manufacturer's protocol to remove any contaminating genomic DNA and purified using an RNeasy Mini Kit (Qiagen).

Reverse transcription screening for improved MaP conditions

In order to increase the efficiency of reverse transcription, various conditions were evaluated for those that improved double-stranded DNA yield after second-strand synthesis. Initial reaction conditions were selected from a subset of those screened for the Smart-seq2 protocol (Picelli et al. 2013). The following is the optimized protocol: To 700 ng RNA was added 200 ng of random nonamer primer, 2 μL of 10 mM dNTPs (Fermentas), and water to a final volume of 10 μL. Primers were annealed at 65 °C for 5 min followed by 4 °C for 2 minutes. Next, 9 μL of buffer master mix [2 μL of 10X NTP minus (10X = 500 mM Tris, pH 8.0, 750 mM KCl, 100 mM DTT), 2.76 μL water, 4 μL 5M betaine (Sigma), 0.24 μL 500 mM $MnCl_2$] was added to the annealed reaction mix. Samples were incubated at 25 °C for 2 min, 1 μL of SuperScript II (Invitrogen) was added, and samples were incubated according to a stepped primer extension protocol with the following program: 25 °C for 10 minutes, followed by 42 °C for 90 minutes, then 10 cycles of [2 minutes at 50 °C, 2 minutes at 42 °C]. An enzyme inactivation step was performed by incubating the samples at 70 °C for 10 minutes. Following primer extension, cDNA products were purified using RNAclean beads (Agencourt) using a 1.8 bead to sample ratio according to the manufacturer's protocol. Purified RNA was eluted from the beads in 68 μL of nuclease-free water and converted to double-stranded DNA using a second-strand synthesis enzyme mix (NEB) according to the manufacturer's instructions. Following second-strand synthesis, double-stranded DNA was purified using AmpureXP beads (Agencourt) using a 0.7:1 bead to sample ratio. Product sizes following second-strand synthesis were quantified using the Agilent Bioanalyzer 2100.

Library preparation and sequencing

For library preparation, 1 ng of each second-strand synthesis product was used in the NexteraXT (Illumina) sample preparation kit according to the manufacturer's directions. Final libraries were size-selected using AmpureXP beads (Agencourt) with a 0.5:1 bead to sample ratio. The libraries were quantified using an Agilent Bioanalyzer 2100 and QuBit high-sensitivity dsDNA

assay. Sequencing was performed on an Illumina NextSeq 500 system with a loading concentration of 1.4 pM, yielding approximately 400 million reads.

Data processing and alignment

Adapter sequences were removed from raw FASTQ files using the program *scythe* with default parameters (Buffalo 2011). Next, reads were trimmed for quality using *sickle* in paired-end mode with a Phred quality cutoff of 20 and a minimum length of 20 (Joshi and Fass 2011). Only pairs where both mates passed filtering were used in downstream stages. Following adapter removal and quality trimming, *ShapeMapper* (version 1.2) was used to map the processed FASTQ files to the 16S and 23S sequences (Siegfried et al. 2014). No further quality trimming was performed during *ShapeMapper*'s quality trimming stage. During the read alignment stage two additional flags were given to Bowtie2 in order to force concordant alignments: "--no-discordant" and "--no-mixed". The following options were changed from the defaults to optimize for long insert sizes. Parameters used were "maxInsertSize=1200", "minMapQual=30", and "minPhredToCount=30".

Correlation analysis of randomly primed reads

The "mutation strings" files from the *ShapeMapper* pipeline were used as input for randomly primed correlation analysis since they contain a simplified representation of the read alignment location, mutation locations, and sequencing instrument quality calls. A square matrix was constructed with a size equal to the length of the aligned RNA (Fig. 5.4a). In each element is a 2x2 contingency matrix containing the counts for possible outcomes comparing two nucleotides (Fig. 5.4b). In each read all *i-j* combinations of nucleotides are used to index the storage matrix. Mutations (scored as 1) and matching nucleotides (scored as 0) were used to index the contingency table. Only nucleotides with a phred score above 30 were counted. After all reads were processed, the storage matrix contained an easily indexed representation of the entire sequencing data set. Each *i-j* element in the matrix contains a snapshot of all the reads that span nucleotides *i* and *j*. Next each $i < j$ pair in

the read storage matrix was tested for significance using the Yates Chi-squared test with a significance criteria of 20 using the equation shown in Fig. 5.4c. Pearson's phi statistic was also calculated. After correlations were calculated for each of the samples, correlations from biological replicates were pooled requiring that each correlation pair must occur in both replicates. Correlation values in the final data set are the average of two biological replicates.

Network analysis of correlations in the small ribosomal subunit

Correlation values from the +Rif and +Rif/+Spc samples were fed into the network visualization software *gephi* using nucleotides as nodes and the correlation strength as edges in an undirected graph (Bastian et al. 2009). The graph was filtered with two requirements: first that the strength of the correlation must be greater than 0.015, and second that each node must have at least three other connections (k-core = 3). The graph layout was calculated using the "Force Atlas" algorithm, which rearranges the nodes such that those with stronger connecting weights arrange themselves closer together in space. Next, the graph modularity was calculated (sensitivity = 1.0) in order to detect possible communities of interactions among nucleotides (Blondel et al. 2008).

REFERENCES

Adilakshmi T, Bellur DL, Woodson SA. 2008. Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly. *Nature* **455**: 1268–1272.

Adilakshmi T, Ramaswamy P, Woodson SA. 2005. Protein-independent folding pathway of the 16S rRNA 5' domain. *J Mol Biol* **351**: 508–519.

Agirrezabala X, Liao HY, Schreiner E, Fu J, Ortiz-Meoz RF, Schulten K, Green R, Frank J. 2012. Structural characterization of mRNA-tRNA translocation intermediates. *Proc Natl Acad Sci* **109**: 6094–6099.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**: P10008.

Borovinskaya MA, Shoji S, Holton JM, Fredrick K, Cate JHD. 2007. A steric block in translation caused by the antibiotic spectinomycin. *ACS Chem Biol* **2**: 545–552.

Buffalo V. 2011. Scythe - A bayesian adapter trimmer. *githubcom*. https://github.com/vsbuffalo/scythe (Accessed March 15, 2015).

Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106**: 97–102.

Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**: 696–700.

Dunkle JA, Wang L, Feldman MB, Pulk A, Chen VB, Kapral GJ, Noeske J, Richardson JS, Blanchard SC, Cate JHD. 2011. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332**: 981–984.

Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res* **15**: 9109–9128.

Frank J, Gonzalez RL. 2010. Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu Rev Biochem* **79**: 381–412.

Homan PJ, Favorov OV, Lavender CA, Kursun O, Ge X, Busan S, Dokholyan NV, Weeks KM. 2014. Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci* **111**: 13858–13863.

Joshi NA, Fass JN. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. *githubcom*. https://github.com/najoshi/sickle (Accessed March 15, 2015).

Kladwang W, VanLang CC, Cordero P, Das R. 2011. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem* **3**: 954–962.

McGinnis JL, Weeks KM. 2014. Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry* **53**: 3237–3247.

Mohan S, Donohue JP, Noller HF. 2014. Molecular mechanics of 30S subunit head rotation. *Proc Natl Acad Sci* **111**: 13325–13330.

Noeske J, Cate JHD. 2012. Structural basis for protein synthesis: snapshots of the ribosome in motion. *Curr Opin Struct Biol* **22**: 743–749.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Meth* **10**: 1096–1098.

Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Meth* **11**: 959–965.

Tyrrell J, McGinnis JL, Weeks KM, Pielak GJ. 2013. The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry* **52**: 8777–8785.

Zhang W, Dunkle JA, Cate JHD. 2009. Structures of the ribosome in intermediate states of ratcheting. *Science* **325**: 1014–1017.