The SHAPE of tRNA folding and of the 5'-end of the HIV-1 genome

Kevin A. Wilkinson

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

Chapel Hill

2007

Approved By

Kevin M. Weeks

Linda L. Spremulli

Howard M. Fried

Matthew Redinbo

Ronald Swanstrom

Abstract

Kevin A. Wilkinson

The SHAPE of tRNA folding and of the 5'-end of the HIV-1 genome

(Under the direction of Kevin M. Weeks)

Most RNAs function only once they fold to form difficult-to-predict base-paired helices and other structural elements. As an RNA forms a preferred secondary or tertiary structure, a characteristic set of nucleotides becomes constrained by base pairing and higher-order interactions, while unconconstrained positions remain flexible. Determining local nucleotide flexibility as a function of nucleotide position in a folded RNA provides important information that enables the sequence and structure of an RNA to be related to its biological function.

I have developed a technology, termed selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE), that can be used to interrogate RNA structure under diverse *in vitro* and *in vivo* conditions. SHAPE chemistry can be applied to monitor protein binding events and locate promising sites for primer annealing in arbitrary RNA. SHAPE chemistry is based on the discovery that flexible RNA nucleotides preferentially sample conformations that enhance the nucleophilic reactivity of the 2'-hydroxyl group toward electrophiles, such as N-methylisatoic anhydride. Modified sites are detected as

stops in an optimized DNA primer extension reaction, followed by sizing of the extension products. SHAPE chemistry scores local flexibility at all four ribonucleotides in a single experiment and discriminates between base-paired versus unconstrained residues with a dynamic range of 20-fold or greater.

I have applied SHAPE chemistry to observe equilibrium melting of a model tRNA at single nucleotide resolution. I observed that RNA folding is a complex process involving structural rearrangement and the formation of tertiary structure concurrent with secondary structure.

Furthermore, I have employed capillary electrophoresis and sophisticated analysis algorithms to create a high-throughput SHAPE (hSHAPE) experiment that can comprehensively interrogate the flexibility of several hundred nucleotides in a single robust experiment.

Using hSHAPE, I analyzed the structure of HIV-1 genomic RNA as a function of 4 different biologically relevant states, including infectious viral particles. Despite many thermodynamically plausible structures, the HIV-1 genome exists in a single conformation. I observed the effects of tRNA primer binding, and the effects of nucleocapsid protein on RNA flexibility. hSHAPE chemistry is a promising, scalable approach that can rapidly and accurately analyze the structure of RNA molecules under biologically relevant conditions.

I dedicate this work to the good people of the Newman Catholic Student Center Parish. You are examples of what a leaven to society means. I have grown as much in my faith at 218 Pittsboro St. as I have as a scientist in Kenan Labs.

- *Ubi carias et amour, Deus ibi est!* -

Acknowledgements

Thank you Mom, Dad, and Sara for all of the extra spinach that made me grow tall, all the homework help, and the prayers.

Thank you Weeks Lab members past and present who have: (a) helped me learn the ropes, (b) learned the ropes from me, or (c) had to put up with my occasional rants.

Thank you to Dr. Weeks for being a supportive adviser, and for keeping your door open. I deeply appreciate that you care about my success as a scholar.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

° C	degrees Celsius
μg	microgram
μL	microliter
μM	micromolar
μm	micrometer
A	adenine
$Å^2$	Angstroms Squared
AT-2	2,2'-dithioldipyridine
ATP	adenosine triphosphate
cDNA	copy DNA (from an RNA template)
CMCT	1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluensulfate
CO_2	carbon dioxide
CTP	cytosine triphosphate
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
ddATP	dideoxyadenosine triphosphate
ddCTP	dideoxycytidine triphosphate
ddNTP	dideoxynucleotide triphosphate
ddTTP	dideoxythymidine triphosphate
DEPC	diethyl pyrocarbonate
dGTP	deoxyguanosine triphosphate

dITP	deoxyinosine triphosphate
D-Loop	dihydrouridine loop
DMS	dimethyl sulfate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic Acid
DTT	dithiothreitol
dTTP	deoxythymidine triphosphate
EDTA	ethylenediaminetetraacetic acid
G	guanosine
ΔG	Gibbs free energy
g	acceleration due to gravity
GTP	guanosine triphosphate
h	hours
H_2O	water
HCl	hydrochloride, hydrochloric acid
HEPES	4-(2-hydroxyethyl)-1- piperazineethanesulfonic acid
HIV-1	human immunodeficiency Virus Type 1
hSHAPE	high throughput selective 2'-hydroxyl acylation analyzed by primer extension
kcal/mol	kilocalories per mol
KCl	potassium chloride
ln	natural lograthim
M	Molar

mg	millogram
$MgCl_2$	Magnesium (II) Chloride
min	minutes
mL	milliliter
mM	millimolar
mRNA	messenger Ribonucleic Acid
NaCl	Sodium Chloride
NaOH	sodium hydroxide
NCp7	nucleocapsid protein
nm	nanometers
NMIA	N-methylisatoic anhydride
NMR	nuclear magnetic resonance
nt	nucleotide
NTP	nucleotide triphosphate
nts	nucleotides
P	phosphorous
^{32}P	phosphorus-32
Pb(2+)	Lead (II) Ion
PCR	polymerase chain reaction
PEG	polyethylene glycol
pH	potential of hydrogen
pmol	picomoles
RISC	ribnucleic acid-induced silencing complex

RNA	Ribonucleic Acid
RNAi	Ribonucleic Acid Interference
RNAse	Ribonuclease
RRE	Rev Responsive Element
RT	Reverse Transcriptase
s	seconds
S. cerevisiae	Saccharomyces cerevisiae
SARS	Severe Acute Respiratory Syndrome
SHAPE	Selective 2'-Hydroxyl Acylation analyzed by Primer Extension
SSIII FS	Superscript III First Strand
T	Thymidine
TE	10 mM TRIS, 1 mM EDTA, pH 8
T-Loop	Ribothymidine Loop
TRIS	2-amino-2-hydroxymethyl-1,3-propanediol
tRNA	Transfer Ribonucleic Acid
tRNA(Asp)	Aspartyl Transfer Ribonucleic Acid
tRNA(Lys3)	Lysyl Transfer Ribonucleic Acid Isoacceptor 3
U	Uracil
UTP	Uridine Triphosphate
UV	ultraviolet
W	Watts
w/v	weight per volume

CHAPTER 1

RNA AND FOLDING

1.1 RNA structure

Ribonucleic acid (RNA) plays a central role in the function of all cells and many viruses. Unlike most DNA and proteins, RNA interacts with other cellular molecules and biopolymers on *both* a sequence and structural level. RNA sequences act as the central information conduit in cells between the genome and ribosome, where genes are expressed as protein.[1] Many cellular RNAs and RNA domains are not translated into proteins, but participate in events central to cellular and viral function. For example, RNA and RNA-protein conjugates play necessary roles in nearly all steps of protein production, including correct translation of gene sequences into proteins (mRNA, tRNA, and the ribosome[1]) protein localization (signal recognition particle[2,3]), and regulation of translation (riboswitches[4], RNAi[5]). RNA also serves as the genomes of several viruses, including human immunodeficiency virus (HIV)[6], hepatitis C virus[7], influenza virus,[8] and the virus that causes SARS[9]. Viral genomic RNAs represent genetic systems that contain both translated, mRNA-like regions, and domains that are not translated into protein. The noncoding domains are nonetheless critical for viral function.

To accomplish these divergent tasks, RNA forms a variety of modular, intramolecular interactions that define catalytic active sites or modulate the function of regulatory motifs.[12,13] RNA folds into structures that define binding sites for proteins,

metal ions, and small molecule ligands. Ligand binding may, in turn, cause a structural change in the RNA. For instance, RNA domains have been discovered that fold into structures that rearrange in the presence of a specific metabolite,[4] thus modulating RNA function in response to cellular state. Other regions of RNA or entire RNA molecules may include long sequences with relatively little structure or exist as an equilibrium of different structures. These flexible regions may participate in, and reduce the energy costs of binding events, such as primer binding and dimerization.[14]

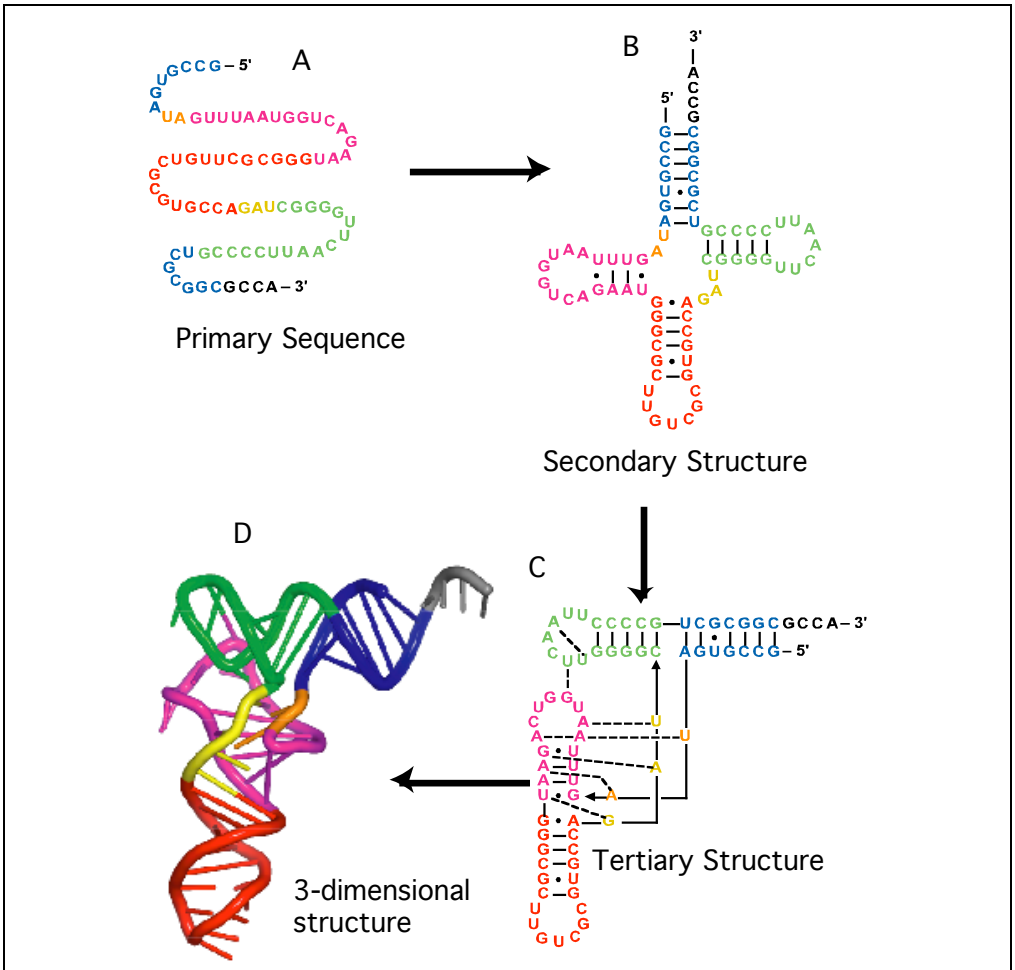


Fig. 1.1: Hierarchical folding of *S. cerevisae* tRNA^{Asp}. According to the hierarchical folding model, primary sequence (a) first forms base pairs or secondary structure (b). The secondary structure then forms tertiary interactions (c, dashed lines) to produce an active three-dimensional structure (d).[10,11] Nucleotides are colored for clarity.

Despite a few documented exceptions,[15-17] current models indicate that RNA folding is strongly hierarchical (Figure 1.1)[12] – that base pairing interactions (secondary structure) is more stable and forms more rapidly than tertiary interactions. This model indicates that formation of helices drives formation of tertiary structure by dictating which nucleotides are available for tertiary interactions.[18]

1.1.1 Examples of RNA folding. Transfer RNA represents an instructive example of the different types of structures RNA can form. Four short helices are constrained in a specific three-dimensional conformation by a network of tertiary interactions, including coaxial helical stacks and base triples (Figure 1.1 D).[10,11] Folded tRNA molecules interact with aminoacyl tRNA synthetases which recognize specific tRNAs and charge them with specific amino acids,[19] as well as with the ribosome and the mRNA sequence[20]. Folding of tRNA also plays a role in recognition by enzymes that covalently modify bases to more stringently enforce an active structure.[21,22]

Large RNAs, such as the HIV-1 genome represent a more complex example of RNA folding. Large RNAs generally fold into a superstructure consisting of several different functional domains. Although only ~15% of the HIV-1 genomic structure has been rigorously characterized, a variety of structures for functional domains that recognize cellular and viral components have been proposed.[6] For instance, the Tat and Rev viral proteins bind to specific secondary structures in the genome[6]. Tat enhances transcription of viral RNA; Rev participates in export of unspliced and singly-spliced genomes from the nucleus.[23] The Tat binding site consists of a well-defined stem loop containing a three-nucleotide bulge.[24,25] The Rev binding site (Rev-responsive element, or

RRE) also folds into an extensively paired structure, but competing secondary structural models have been proposed.[26-28]

The HIV-1 genome contains structural domains that assist in dimerization and packaging of genomes into viral particles. Both steps are required for proper selection and export of infectious particles. The ψ site, or packaging signal, is bound by the nucleocapsid domain of Gag and is targeted to the cell membrane for virion formation. This region also coincides with sequences implicated in the dimerization of the genome.[6,29] Like the RRE, various different compact structures have been proposed for this region,[29-36] but all indicate an extensive network of intramolecular interactions.

The HIV-1 genome also forms structures that facilitate interactions with cellular components. To initiate reverse transcription, a single $tRNA^{Lys3}$ binds a domain in the HIV-1 genome in as many as three different places.[37] The binding site on the genome contains an enforced flexible region in the genome, as well as a stem-loop structure.[32,33] Cellular HIV transcripts also serve as a template for protein synthesis for cellular ribosomes. The cellular ribosome also recognizes a heptanucleotide slippery sequence as well as a specific RNA folding mode to effect a -1 nucleotide frameshift at the *gag-pol* gene junction. As with other regions of the HIV-1 genome, various different structural models, including a stem loop[38] and pseudoknot[39] have been proposed for the frameshift site.

The HIV-1 genome folds into a structure with different structural domains designed to serve specific functions in the viral life cycle, including transcriptional activation, nuclear export, translation, and selection for packaging. Recent work has proposed structures for these domains, but the structures proposed are often incompatible

with one another, or miss elements other studies observe. A comprehensive view of the HIV-1 genome may be required to parse these models.

1.2 Methods of analyzing RNA structure

To explain the role of RNA structure in cellular function and disease, a number of chemical, physical, and computational methods are used to analyze the structure of RNA. Each method has yielded results that have been interpreted to develop well-supported models describing the function and structures of various important RNAs. X-ray crystallography[40] is by far the best method for developing near-atomic resolution structures of highly structured regions of an RNA molecule and has been applied to RNAs and ribonucleoprotein complexes such as the entire ribosome[41], RNAse P[42], and the hairpin[43] and GlmS[44] ribozymes. These studies have provided enormous insight to the nature of RNA folding and catalysis.

Nuclear Magnetic Resonance (NMR) has also been used to develop high-resolution RNA structures in solution phase. NMR reports the dynamic nature of RNA in solution, but atomic resolution analysis is currently limited to RNA <100 long. This limit is due to the high degeneracy of the RNA molecule. Additionally, RNA has a relatively high tumbling rate in solution, which broadens linewidths.[45,46] An important application of NMR technology is to analyze, in three dimensions, the dimerization and binding of RNA to various proteins and small molecule ligands. Crystallography and NMR provide atomic-resolution data but require comparatively large amounts of RNA and complicated data analysis.

When crystallographic or NMR approaches are not viable or unnecessary for an RNA of interest, a variety of nucleotide-resolution approaches have been developed that consume far less RNA and produce interpretable results with standard laboratory equipment. These techniques can also be used on RNAs that do not form stable, or well-defined folds.

Nucleotide-resolution probing approaches use chemicals or enzymes sensitive to local RNA structure to differentiate between constrained and flexible nucleotides. Reagents such as dimethyl sulfate (DMS), kethoxyl, bisulfite, diethyl pyrocarbonate (DEPC), and 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluensulfate (CMCT) modify or cleave different subsets of flexible nucleotides in an RNA. Several RNase enzymes, such RNaseA, T1, and V1, preferentially cleave specifically at paired or single-stranded regions in an RNA.[47,48] Structure sensitive cleavage with Pb^{2+}[49] and heavy metal complexes[50] are additional common methods to identify RNA structure. Cleaved or modified sites on an RNA may be located using a number of separation and quantitation techniques, such as gel[47] or capillary electrophoresis[51] or mass spectrometry.[52] These methods yield qualitative structural information at ~50 – 75% of nucleotides in an RNA especially when multiple reagents and enzymes are used in concert on the same RNA.

An important application of the data developed by chemical and enzymatic techniques is to constrain RNA structure prediction algorithms. Most successful structure prediction algorithms use a thermodynamic model based on experimentally and heuristically determined constraints.[53,54] However, thermodynamic models are approximate and non-thermodynamic contributions can modulate RNA structure. Therefore, *de novo* structure prediction may yield multiple structures whose accuracy

varies widely and can be quite low, especially for long sequences. Accuracy improves as constraints are incorporated into the prediction.[55] Because chemical[56] and enzymatic[33] probing methods leave gaps in structural information, some predictions remain open to misinterpretation or errors.[57]

1.3 A robust method for RNA structure analysis

With accurate RNA structures, it is possible to explain the functional bases of disease, describe basic mechanisms of life, and identify drug target sites. Crystallography, NMR, and chemical or enzymatic probing yield fair to excellent results that can be interpreted to develop structural models of entire RNAs or RNA motifs. However, the intrinsic properties of the extant technologies do not lend themselves to high-throughput approaches that can comprehensively analyze RNA structure.

In collaboration with Edward Merino, I have developed a new technology, called selective 2'-hydroxyl acylation analyzed by primer extension, or SHAPE, that rapidly analyzes the structure of RNA of arbitrary length and structural complexity.[57-60] SHAPE chemistry generates quantitative flexibility information for nearly all nucleotides in an arbitrary RNA. In this application the flexibility information is used as a guide for interpreting the state of base pairing or other structural constraints.

One important application of SHAPE chemistry is to constrain structure prediction algorithms, such as RNAstructure, to predict RNA folds.[55] SHAPE constraints applied to the RNAstructure program have been used to develop and confirm structural models or describe the structural effects of varying the state of an RNA.[14,57-59,61] The

program correctly predicts ~90% of base pairs when applied to a well characterized RNA. Errors are usually short range and involve the nucleotides at the ends of helices.[62]

Accurate structural constraints for RNA have a variety of additional applications aside from proposing RNA secondary structures. RNAi technology is a common approach used to silence gene expression in cells, and shows significant promise as a basis for viral, cancer, and genetic therapies.[63,64] RNAi requires the efficient annealing of a RISC-RNA complex to a target mRNA.[65] The location of these target sites in an RNA of unknown structure usually requires the use of time consuming screens.[66-68] SHAPE chemistry, however, is ideally suited to locate flexible regions in RNA, which may represent available sites for RISC-mediated binding of RNAi, as well as other nucleotide ligands.[69,70]

RNA often exists in cells bound to proteins as a ribonucleoprotein complex. Protein binding events on an RNA occlude and constrain various nucleotides; changes observable by SHAPE chemistry. Because SHAPE can be used in a complex solution, it is ideally suited to locate binding sites on an RNA and locate the structural changes that occur upon protein binding. Presently, SHAPE chemistry is being used to map protein binding on the telomerase RNA[71] and the *Saccharomyces cerevisiae* bI3 intron splicing complex.[72]

Structurally complex RNAs form native structure on the minute timescale or longer due to a complex energy landscape.[73-75] The recent identification of fast acting chemicals that modify the 2'-hydroxyl opens the possibility of analyzing real-time folding of native RNAs.[76] Characterization of intermediates in a folding pathway lends insight into the folding landscape of RNA and may also be used to develop the basis for more

powerful structure prediction algorithms that include kinetic as well as thermodynamic parameters.

Folding and transcription may also be coupled resulting in native structures that may represent a local energy minimum. SHAPE chemistry could characterize folding intermediates during transcription by characterizing RNA libraries of the same sequence and different length. Fast acting reagents[76] may also be used in concert with pulse-chase experiments to analyze folding intermediates during transcription.

SHAPE chemistry can be applied to develop *in vivo* sensors. RNA aptamers are small RNA sequences that have been identified that bind small molecules[77,78]. These binding events often significantly change the structure of some nucleotides in the aptamer.[79] Cells made to transcribe RNA aptamer sequences may be subjected to SHAPE chemistry. The structures determined by SHAPE can then be used to interpolate concentration of a small molecule in the cell. Significant groundwork has been completed for aptamers that bind ATP. Additionally, recent SHAPE-based data on tRNA folding indicates its potential use as an *in vivo* magnesium ion sensor.[80]

The development of SHAPE technology required the modification or replacement of several steps in the traditional chemical modification techniques. Major innovations include identification of a chemical reaction that is generic to all 4 RNA nucleotides, development of a robust primer extension reaction that is insensitive to RNA structure, separation of products on a DNA sequencer, and development of software tools for quantitative, rapid analysis of the data.

1.4 Chapter overview

In this work, I describe the development of SHAPE chemistry from chemical observation to a generic method of RNA structure analysis. Chapter 2 explains the basic principles of the SHAPE technique. In collaboration with Edward Merino, we observe that the nucleophilicity of the RNA ribose hydroxyl is exquisitely sensitive to local nucleotide flexibility. We identify an electrophile, N-methylisatoic anhydride (NMIA) that preferentially modifies the 2'-oxyanion to form a 2'-O-ester adduct. This chemical observation is coupled with primer extension to measure adduct formation at all positions in an RNA. We use SHAPE to map the structure of and to distinguish fine differences in structure for $tRNA^{Asp}$ transcripts at single nucleotide resolution. I finish Chapter 2 with a practical guide for performing SHAPE chemistry, along with a troubleshooting manual.

In chapter 3, I apply SHAPE chemistry to address the hierarchical RNA folding model to observe, for the first time, the unfolding of an RNA at single nucleotide resolution.[18] With Edward Merino, I establish that SHAPE is an RNA analogue to the protein amide hydrogen exchange experiment, and that the NMIA-RNA reaction proceeds to identical endpoints independent of temperature. I perform equilibrium folding experiments on *S. cerevisiae* $tRNA^{Asp}$ and observe a complexity for the unfolding of $tRNA^{Asp}$ transcripts that is not anticipated by current models for RNA folding. We quantify six well-defined transitions for $tRNA^{Asp}$ transcripts between 35 and >75 ˚C, including asymmetric unfolding of the two strands in a single helix, multistep loss of tertiary interactions, and a multihelix conformational shift. The three lowest temperature transitions each involve coupled interactions between the secondary and tertiary

structure. Thus, even for this small RNA, multiple nonhierarchical and complex interactions dominate the equilibrium transitions most accessible from the native state.

In Chapter 4, I describe the development of a high throughput SHAPE experiment (hSHAPE) in collaboration with Suzy Vasa, Rob Gorelick, Nicolas Guex, Alan Rein, David Mathews, and Morgan Giddings. I describe adaptation of SHAPE to a DNA sequencer. DNA sequencers can analyze SHAPE data for longer RNAs – up to 500 nucleotides in a single read, and require less time and effort than gel electrophoresis. I also describe a number of software innovations, made in collaboration with Suzy Vasa and Nicolas Guex, that allows rapid and quantitative analysis of SHAPE data from a DNA sequencer. I will also describe some techniques useful in developing structural models from hSHAPE data.

In chapter 5, I collaborate with Robert Gorelick, Suzy Vasa, Nicolas Guex, Alan Rein, David Mathews, and Morgan Giddings to use hSHAPE to comprehensively analyze the 5'-most 10% of the HIV-1 genome under four different states, including authentic virions– a structural analysis of over 8200 nucleotides. Virion-specific structures and protein binding sites are analyzed by comparison with protein-free genomic RNA and authentic RNAs in which native protein interactions are disrupted. Despite a large number of thermodynamically accessible states, the 5'-end of the HIV-1 genome exists in a single predominant conformation. I observed that untranslated regulatory domains are more highly structured than coding regions. I detected two opposing activities for the HIV nucleocapsid protein. Stable nucleocapsid-RNA interactions occur at multiple, newly defined, motifs in a regulatory domain; conversely, nucleocapsid destabilizes local RNA structure specifically at sites required to undergo conformational rearrangements,

including the region where reverse transcriptase initiates cDNA synthesis. This chapter reveals a complex coupling between the HIV-1 genome structure and function. The last two chapters also lay the groundwork for comprehensive, quantitative analysis of any viral and cellular RNA.

1.5 References

1. Lewin, B. *Genes VII* (Oxford University Press, Oxford, 2000).

2. Doudna, J. A. & Batey, R. T. Structural insights into the signal recognition particle. *Annu. Rev. Biochem.* **73**, 539-57 (2004).

3. Nagai, K. et al. Structure, function and evolution of the signal recognition particle. *EMBO J.* **22**, 3479-85 (2003).

4. Sashital, D. G. & Butcher, S. E. Flipping off the riboswitch: RNA structures that control gene expression. *ACS Chem. Biol.* **1**, 341-5 (2006).

5. Hammond, S. M. Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett.* **579**, 582209 (2005).

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

7. Gomez, J., Nadal, A., Sabariegos, R., Beguiristain N., Martell, M. & Piron, M. Three properties of the hepatitis C virus RNA genome related to antiviral strategies based on RNA-therapeutics: variability, structural conformation and tRNA mimicry. *Curr. Pharm. Des.* **10**, 3741-3756 (2004).

8. Steinhauer, D. A. & Skehel, J. J. Genetics of influenza viruses. *Annu. Rev. Genet.* **36**, 305-32 (2002).

9. Ziebuhr, J. Molecular biology of severe acute respiratory syndrome coronavirus. *Curr. Opin. Microbiol.* **7**, 412-9 (2004).

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

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

12. Gesteland, R. F., Cech, T. R. & Atkins, J. F. (eds.) *The RNA World* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1999).

13. Holbrook, S. R. RNA structure: the long and short of it. *Curr. Opin. Struct. Biol.* **15**, 302-8 (2005).

14. Badorrek, C. S. & Weeks, K. M. RNA flexibility in the dimerization domain of a gamma retrovirus. *Nat. Chem. Biol.* **1**, 104-11 (2005).

15. LeCuyer, K. A. & Crothers, D. M. The Leptomonas collosoma spliced leader RNA can switch between two alternate structural forms. *Biochemistry* **25**, 5301-11 (1993).

16. Gluick, T. C. & Draper, D. E. Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* **241**, 246-62 (1994).

17. Wu, M. & Tinoco, I. RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci. USA* **95**, 11555-60 (1998).

18. Tinoco, I. & Bustamante, C. How RNA folds. *J. Mol. Biol.* **293**, 271-281 (1999).
19. O'Donoghue, P. & Luthey-Schulten, Z. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev.* **67**, 550-73 (2003).

20. Spirin, A. S. The ribosome as an RNA-based molecular machine. *RNA Biol.* **1**, 3-9 (2004).

21. Helm, M. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucl. Acids Res.* **34**, 721-33 (2006).

22. Grosjean, H., J., E., Straby, K. B. & Giege, R. Enzymatic formation of modified nucleosides in tRNA: dependence on tRNA architecture. *J. Mol. Biol.* **255**, 67-85 (1996).

23. Emerman, M. & Malim, M. H. HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. *Science* **280**, 1880-4 (1998).

24. Berkhout, B. & Jeang, K. T. trans activation of human immunodeficiency virus type 1 is sequence specific for both the single-stranded bulge and loop of the trans-acting-responsive hairpin: a quantitative analysis. *J. Virol.* **63**, 5501-4 (1989).

25. Jakobovits, A., Smith, D. H., Jakobovits, E. B. & Capon, D. J. A discrete element 3' of human immunodeficiency virus 1 (HIV-1) and HIV-2 mRNA initiation sites mediates transcriptional activation by an HIV trans activator. *Mol. Cell Biol.* **8**, 2555-61 (1988).

26. Dayton, E. T., Konings, D. A., Powell, D. M., Shapiro B. A., Butini, L., Maizel, J. V. & Dayton, A. I. Extensive sequence-specific information throughout the CAR/RRE, the target sequence of the human immunodeficiency virus type 1 Rev protein. *J. Virol.* **66**, 1139-1151 (1992).

27. Peleg, O., Brunak, S., Trifonov, E., Nevo, E. & Bolshoy, A. RNA secondary structure and squence conservation in C1 region of human immunodeficiency virus type 1 env gene. *AIDS Res. Hum. Retroviruses*. **18**, 867-78 (2002).

28. Phuphuakrat, A. & Auewarakul, P. Heterogeneity of HIV-1 Rev response element. *AIDS Res. Hum. Retroviruses*. **19**, 569-74 (2003).

29. Clever, J. L. & Parslow, T. G. Mutant human immunodeficiency virus type 1 genomes with defects in RNA dimerization or encapsidation. *J. Virol*. **71**, 3407-14 (1997).

30. Abbink, T. E., Ooms, M., Haasnoot, P. C. & Berkhout, B. The HIV-1 leader RNA conformational switch regulates RNA dimerization but does not regulate mRNA translation. *Biochem*. **44**, 9058-66 (2005).

31. Ooms, M., Huthoff, H., Russell, R., Liang, C. & Berkhout, B. A riboswitch regulates RNA dimerization and packaging in human immunodeficiency virus type 1 virions. *J. Virol.* **78**, 10814-9 (2004).

32. Paillart, J. C., Dettenhofer, M., Yu X. F., Ehresmann C., Ehresmann, B, & Marquet, R.First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J. Biol Chem*. **279**, 48397-403 (2004).

33. Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5' region of the HIV-1 genome. *J. Mol. Biol*. **336**, 369-79 (2004).

34. Baudin, F., Marquet, R., Isel C., Darlix J. L., Ehresmann B. & Ehresmann C. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *J. Mol. Biol.* **229**, 382-397 (1993).

35. Greatorex, J., Gallego, J., Varani, G. & Lever, A. Structure and stability of wild-type and mutant RNA internal loops from the SL-1 domain of the HIV-1 packaging signal. *J. Mol. Biol*. **322**, 543-57 (2002).

36. Amarasinghe, G. K., De Guzman, R. N., Turner, R. B. & Summers, M. F. NMR structure of stem-loop SL2 of the HIV-1 psi RNA packaging signal reveals a novel A-U-A base-triple platform. *J. Mol. Biol.* **299**, 145-56 (2000).

37. Kleiman, L. tRNA(Lys3): the primer tRNA for reverse transcription in HIV-1. *IUBMB Life*. **53**, 107-14 (2002).

38. Butcher, S. E. & Staple, D. W. Solution structure and thermodynamic investigation of the HIV-1 frameshift inducing element. *J. Mol. Biol.* **349**, 1011-23 (2005).

39. Baril, M., Dulude, D., Steinberg, S. V. & Brakier-Gingras, L. The frameshift stimulatory signal of human immunodeficiency virus type 1 group O is a pseudoknot. *J. Mol. Biol.* **331**, 571-583 (2003).

40. Egli, M. Nucleic acid crystallography: current progress. *Curr. Opin. Chem. Biol.* **8**, 580-591 (2004).

41. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905-20 (2000).

42. Torres-Larios, A., Swinger, K. K., Pan, T. & Mondragon, A. Structure of ribonuclease P--a universal ribozyme. *Curr. Opin. Struct. Biol.* **16**, 327-35 (2006).

43. Rupert, P. B. & Ferre-D'Amare, A. R. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature* **410**, 780-6 (2001).

44. Klein, D. J. & Ferre-D'Amare, A. R. Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science* **313**, 1752-6 (2006).

45. Foster, M. P., McElroy, C. A. & Amero, C. D. Solution NMR of Large Molecules and Assemblies. *Biochem.* **46**, 331-40 (2007).

46. Tzakos, A. G., Grace, C. R., Lukavsky, P. J. & Reik, R. NMR techniques for very large proteins and RNAs in solution. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 319-42 (2006).

47. Ehresmann, C., Baudin, F., Mougel, M., Romby P., Ebel, J. P. & Ehresmann, B. Probing the structure of RNAs in solution. *Nucl. Acids Res.* **15**, 9109-9128 (1987).

48. Knapp, G. Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol.* **180**, 192-212 (1989).

49. Lindell, M., Romby, P. & Wagner, E. G. Lead(II) as a probe for investigating RNA structure in vivo. *RNA* **8**, 534-541 (2002).

50. Morrow, J. R. & Iranzo, O. Synthetic metallonucleases for RNA cleavage. *Curr. Opin. Chem. Biol.* **8**, 192-200 (2004).

51. Sobczak, K. & Krzyzosiak, W. J. RNA structure analysis assisted by capillary electrophoresis. *Nucl. Acids Res.* **30**, e124 (2002).

52. Yu, E. & Fabris, D. Direct probing of RNA structures and RNA-protein interactions in the HIV-1 packaging signal by chemical modification and electrospray ionization fourier transform mass spectrometry. *J. Mol. Biol.* **330**, 211-23 (2003).

53. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 71 (2004).

54. Mathews, D. H. & Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **16**, 270 (2006).

55. Mathews, D. H., Disney, M. D., Childs J. L., Schroeder S. J., Zuker, M. & Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101**, 7287-7292 (2004).

56. Felden, B., Himeno, H., Muto, A., McCutcheon, J. P., Atkins, J. F. & Gesteland, R. F. Probing the structure of the Escherichia coli 10Sa RNA (tmRNA). *RNA* **3**, 89-103 (1997).

57. Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C. & Weeks, K. M. Structures of the HIV-1 Genome. *In preparation* (2007).

58. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J. Am. Chem. Soc.* **127**, 4659-67 (2005).

59. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127**, 4223-31 (2005).

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

61. Gherghe, C. & Weeks, K. M. The SL1-SL2 (stem-loop) domain is the primary determinant for stability of the gamma retroviral genomic RNA dimer. *J. Biol Chem.* **281**, 27952-61 (2006).

62. Li, T. W., Mathews, D. H. & Weeks, K. M. *In preparation* (2007).

63. Uprichard, S. L. The therapeutic potential of RNA interference. *FEBS Lett.* **579**, 5996-6007 (2005).

64. Wall, N. R. & Shi, Y. Small RNA: can RNA interference be exploited for therapy? *Lancet* **362**, 1401-3 (2003).

65. Engles, B. M. & Hutvanger, G. Principles and effects of microRNA-mediated post-transcriptional gene regulation. *Oncogene* **25**, 6163-9 (2006).

66. Castanotto, D. & Scherer, L. Targeting cellular genes with PCR cassettes expressing short interfering RNAs. *Methods Enzymol.* **392**, 173-85 (2005).

67. Holen, T., Amarzguioui, M., Wiiger, M. T., Babaie, E. & Prydz, H. Positional effects of short interfering RNAs targeting the human coagulation trigger tissue factor. *Nucl. Acids Res.* **30**, 1757-1766 (2002).

68. Lee, N. S., Bauer, G., Li, H., Li, M. J., Ehsani, A., Salvaterra, P. & Rossi, J. Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells. *Nat. Biotechnol.* **20**, 500-505 (2002).

69. Sohail, M., Akhtar, S. & Southern, E. M. The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides. *RNA* **5**, 646-55 (1999).

70. Mir, K. U. & Southern, E. M. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol.* **17**, 788-92 (1999).

71. Legassie, J. D. & Jarstfer, M. B. *Unpublished Data*.

72. Duncan, C. & Weeks, K. M. *Unpublished Data*.

73. Thirumalai, D. & Woodson, S. A. Maximizing RNA folding rates: a balancing act. *RNA* **6**, 790-794 (2000).

74. Uhlenbeck, O. C. Keeping RNA happy. *RNA* **1**, 4-6 (1995).

75. Treiber, D. K. & Williamson, J. R. Exposing the kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* **9**, 339-45 (1999).

76. Mortimer, S. M. & Weeks, K. M. A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry. *J. Am. Chem. Soc.* **129**, 4144-5 (2007).

77. Gold, L., Polisky, B., Uhlenbeck, O. C. & Yarus, M. Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763-97 (1995).

78. Wilson, D. S. & Szostak, J. W. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611-47 (1999).

79. Merino, E. J. & Weeks, K. M. Facile conversion of aptamers into sensors using a 2'-ribose-linked fluorophore. *J. Am. Chem. Soc.* **127**, 12766-12767 (2005).

80. Wang, B. & Weeks, K. M. *In preparation* (2007).

CHAPTER 2

PRINCIPLES OF SHAPE CHEMISTRY

2.1 Introduction

SHAPE technology is divided into two overarching steps. In the first part, RNA is sparsely modified at the 2'-hydroxyl position in a structure-selective manner. Next, sites of modification are located by a carefully optimized primer extension. The extent of modification is then quantified at all positions in an RNA. Some aspects of the SHAPE strategy have similarities to classical chemical modification experiments, but advantages of SHAPE are that it is (i) rapid, (ii) quantitative, and (iii) reacts similarly with all 4 nucleotides.[1,2] In this chapter, I describe, in depth, the underlying principles of SHAPE – the nature of the structure-sensitive chemistry, and innovations in primer extension. I include a stepwise procedure for a generic SHAPE experiment with troubleshooting advice at the end of this chapter.

2.1.1 The 2'-position is a candidate for structure-sensitive chemistry. The 2'-ribose position (Figure 2.1) presents an appealing chemical target as it is both accessible for chemistry, and generic to all 4 nucleotides. In collaboration with Edward Merino, I characterized a reagent that targets the 2'-hydroxyl in a structure-selective manner and propose a model for 2'-hydroxyl reactivity in flexible and paired nucleotides. We applied this observation to map, at single nucleotide resolution, the structure of a model RNA.

2.1.2 Previous studies on the 2'-ribose position.

Previous work on the 2'-ribose position indicated that its reactivity is strongly modulated by the adjacent 3'-phosphodiester anion.[3,4] Acylation of synthetic 2'-amine substituted nucleotides to form a 2'-amide product is strongly gated by the underlying local nucleotide flexibility.[3,5-8] Flexible nucleotides in RNA are better able to reach a facilitated transition state in which the 3'-phosphodiester becomes appropriately positioned with the 2'-amine to exert a catalytic effect.[4] 2'-amine acylation thus robustly detects essentially all base paired RNA secondary structures and many tertiary interactions when the sites of 2'-adduct formation are mapped by primer extension.[3,7]

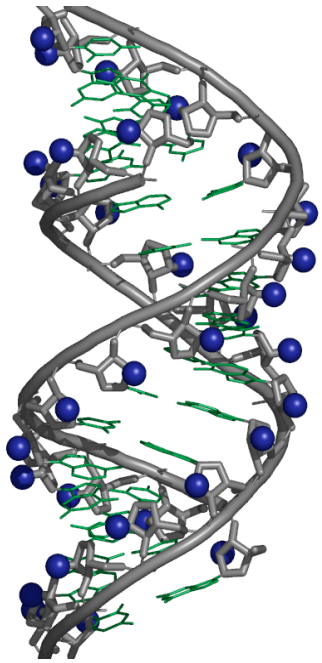


Fig. 2.1: A generic RNA helix. The 2'-hydroxyl group (blue spheres) lie on the sugar-phosphate backbone and are solvent accessible. The aromatic bases (green) are sequestered inside of the helix.

2'-amine-based chemistry is significantly more straightforward to implement than traditional chemical or enzymatic approaches for monitoring RNA secondary and tertiary structure. However, introduction of an artificial 2'-amine group makes an experiment more complex, prevents this chemistry from being used *in vivo*, and may perturb some tertiary interactions involving the 2'-ribose position.

We considered that the proximity of a 3'-phosphodiester anion might modulate the reactivity of the 2'-hydroxyl group normally present in RNA. Since every nucleotide has a 2'-hydroxyl, structural information is, in principle, obtainable for every position in an

RNA. Formation of a bulky 2'-O-adduct could then be detected as an impediment to cDNA synthesis by reverse transcriptase.[3,7]

2.1.3. Identification of a 2'-hydroxyl reactive reagent. Since prior work indicates that the nucleophilicy of a 2'-ribose substituent is modulated by the proximal 3'-group,[4] we sought to identify a small molecule reagent that modified the 2'-hydroxyl under diverse environments. The most promising reagent was N-methylisatoic anhydride (NMIA) which liberates CO_2 to form a 2-methylaminobenzoic acid nucleotide-2'-ester under mild, aqueous conditions[2,9,10] (Figure 2.2A). NMIA specifically reacts with the ribose position in nucleotides, as monitored by gel shift assays using small nucleotides.[2]

A reagent that is capable of modifying hydroxyl groups in aqueous solution will be subject to a competing hydrolysis reaction. NMIA degrades to 2-methylaminobenzoate with experimentally tractable rates over a range 25° to 75° C.[1,2] (Figure 2.2 B and C). Careful measurement of the rates of hydrolysis allows calculation of the rates for ribose 2'-O-adduct formation.[1]

2.1.4 NMIA reactivity with nucleotides is modulated by the 3'-substituent. Attack of a 2'-hydroxyl at the anhydride moiety of NMIA requires formation of the significantly more nucleophilic 2'-oxyanion prior to reaching the ester-forming transition state (see first step in Figure 2.2). Formation of the 2'-oxyanion would be destabilized if the negatively charged 3'-phosphodiester were constrained to be adjacent to the 2'-position (see red overlap symbols in Figure 2.2). Because formation of the oxyanion involves loss of a proton, the hydroxyl pK_a is a good indicator for the relative nucleophilicity of a 2'-hydroxyl. The pK_a for deprotonation of the 2'-hydroxyl of adenosine is 12.2, and this pK_a increases to 13.4 for adenosine 3'-phosphate.[11] The

relative nucleophilicity of a 2'-hydroxyl is thus strongly modulated by neighboring functional groups at the 3'-position. Formation of the oxyanion also presumably induces secondary changes in ribose conformation to yield complex relationships between nucleotide structure and the 2'-hydroxyl pK_a.

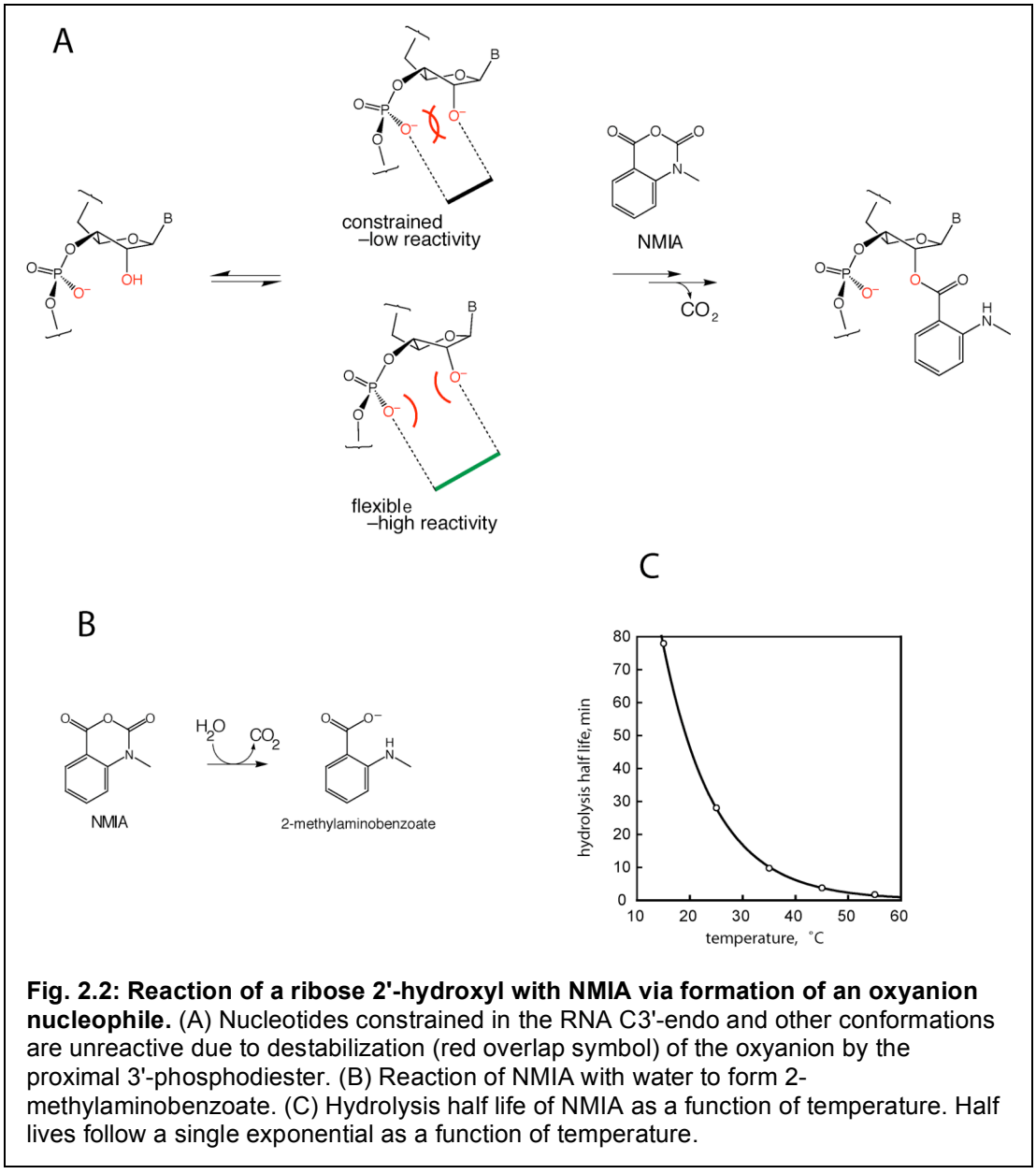


Fig. 2.2: Reaction of a ribose 2'-hydroxyl with NMIA via formation of an oxyanion nucleophile. (A) Nucleotides constrained in the RNA C3'-endo and other conformations are unreactive due to destabilization (red overlap symbol) of the oxyanion by the proximal 3'-phosphodiester. (B) Reaction of NMIA with water to form 2-methylaminobenzoate. (C) Hydrolysis half life of NMIA as a function of temperature. Half lives follow a single exponential as a function of temperature.

To test this model, a series of analyses were carried out using an instructive set of small nucleotides that contained 3'-substituents of different size and charge.[2] These studies indicated that NMIA requires a free 2'-hydroxyl group to modify RNA and that the reactivity of a ribose hydroxyl group is decreased by the proximity of *any* phosphate group. Reactivity can be recovered either by replacing a phosphate diester with an uncharged analog or by increasing the distance between the hydroxyl and phosphate groups.

2.1.5 Structure-selective 2'-hydroxyl reactivity in an oligonucleotide. These studies on the relationship between constrained local structure and 2'-hydroxyl reactivity were extended by evaluating reactivity in an 11-mer oligonucleotide.[2] This RNA was synthesized with a 5'-monophosphate radiolabel and a 3'-dideoxycytidine nucleotide so that all 10 remaining hydroxyl groups lie in internal linkages and adjacent to 3'-phosphodiester groups.

When we exposed this oligonucleotide to NMIA under mild conditions, we could observe the formation of product by gel shift assay. Rates of 2'-O-adduct formation were comparable to small nucleotide analogs of flexible RNA, supporting the interpretation that small nucleotide analog studies reproduce key features of 2'-O-adduct formation.

When the 11-mer oligonucleotide was hybridized to its complement, the rate of adduct formation decreased by 10-fold.[2] Residues at RNA helix termini can experience fraying,[12,13] making them more flexible than internal nucleotides. Therefore nucleotides imbedded within long helices most likely have a far lower reaction yield than that measured here. The 2'-hydroxyl groups in RNA helices project outward towards the solution (Figure 2.1). Therefore, the differential reactivity of single-stranded versus

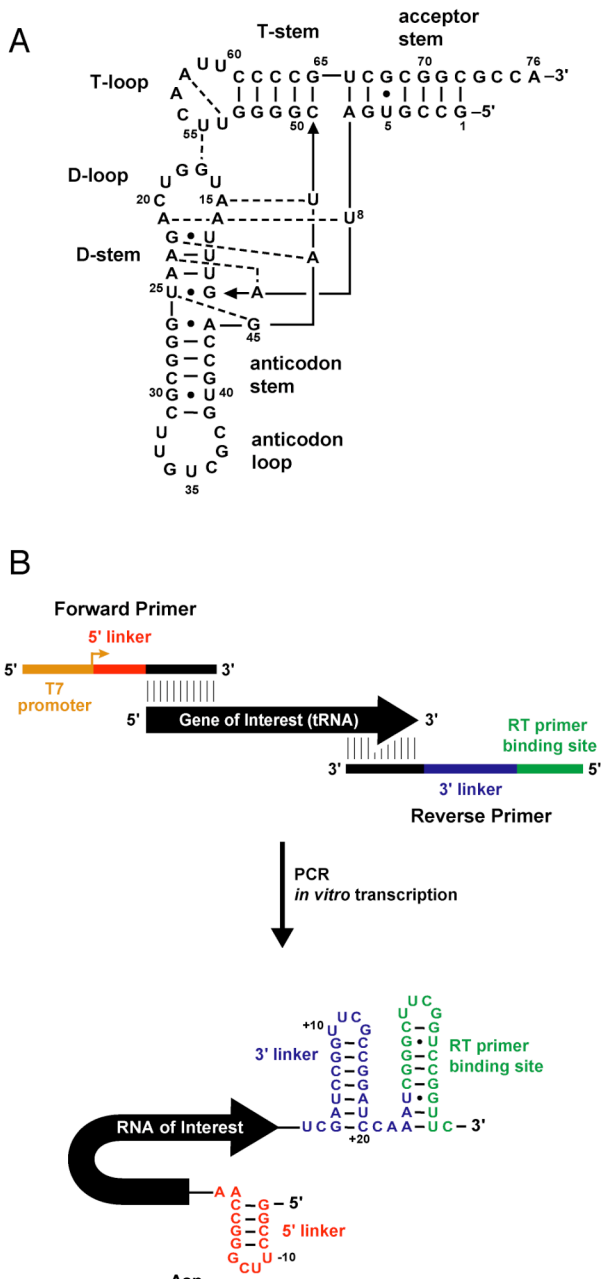


Fig 2.3. tRNA^{Asp} transcript in the context of an RNA structure cassette. tRNA^{Asp} structure is drawn to suggest orientation of helices and tertiary interactions in three-dimensional space. Tertiary interactions are indicated by dashed lines in gray. 5' and 3' linker regions facilitate analysis of sites of 2'-O-adduct formation by primer extension and have been shown to fold independently of the internal tRNA structure.[2] By convention,[14,15] nucleotides in yeast tRNA^{Asp} are numbered 1-76, with residue 47 omitted.

duplex RNA rules out solvent accessibility as the primary explanation for structure-sensitive 2'-hydroxyl reactivity.

Studies with small nucleotide analogs as well as an 11-mer model RNA indicated that NMIA is a good candidate to interrogate the structure of RNA. Three main observations support this conclusion: (i) NMIA specifically modifies the 2' position of RNA, which is generic to all 4 nucleotides, (ii) NMIA hydrolyzes under an experimentally tractable timescale, making a quench unnecessary, and (iii) NMIA scores nucleotide flexibility, not solvent accessibility.

2.2 Results

2.2.1 A structure cassette for analyzing tRNA^{Asp} conformation.

We applied selective

2'-hydroxyl acylation analyzed by primer extension (SHAPE) to explore the structure of *S. cerevisiae* $tRNA^{Asp}$ transcripts. The well studied L-shaped $tRNA^{Asp}$ structure[3,14-17] (Figure 2.3A) encompasses interactions that comprise any folded RNA including double-stranded helices, single-stranded loops and connecting regions, and a network of tertiary interactions. The secondary structure shown (Figure 2.3A) emphasizes the three-dimensional fold of the RNA in which (i) the anticodon and D-stems and (ii) the acceptor and T-stems stack. These two pairs of coaxially stacked helices are then held together in the tRNA architecture by tertiary interactions between the D- and T-loops and involve the joining regions at positions U8-A9 and G45-U48.

The $tRNA^{Asp}$ transcript was placed in the center of an RNA "structure cassette" to facilitate analysis of 2'-O-adduct formation by primer extension (Figure 2.3B). Each stem-loop structure in the cassette contains a stable UUCG tetraloop[18,19] to enforce the designed fold and eliminate alternate foldings with $tRNA^{Asp}$. The hairpins that comprise the flanking 5' and 3' cassette structures have three functions. (i) The 3'-most sequence (green in Figure 2.3B) provides the DNA primer binding site required to initiate reverse transcription. (ii) The 3'-linker (blue) between the primer binding site and the first position in the tRNA allows the reverse transcriptase enzyme to become fully processive prior to reaching the region of structural interest. This linker also prevents non-templated primer extension products from masking structural information at the 3' end of the RNA. (iii) The 5' portion of the cassette (red in Figure 2.3B) displaces the abundant full-length extension product that would otherwise obscure structural information at the 5' end of the RNA. Overall, the design of the structure cassette balances the requirements that it fold autonomously, not interact with the internal RNA, and efficiently bind a primer DNA.

2.2.2 SHAPE footprinting of $tRNA^{Asp}$. The $tRNA^{Asp}$ construct was folded under conditions that strongly favor formation of the native conformation (100 mM NaCl, 10 mM $MgCl_2$)[3,16,17] at 35 °C. NMIA was added and allowed to react with the RNA for 5 hydrolysis half lives (54 min). Sites of 2'-esterification (Figure 2.2) were identified as stops to primer extension by reverse transcriptase. The presence of a 2'-O-adduct causes the reverse transcriptase to stop exactly one nucleotide prior to the modified base.

Primer extension products were resolved in a sequencing gel and converted to intensity versus position plots (Figure 2.4A). When $tRNA^{Asp}$ was incubated under the standard reaction conditions in the absence of NMIA, primer extension yielded a strong full-length product and only minor bands of intermediate length (green trace, Figure 2.4A). When an otherwise identical SHAPE experiment was performed in the presence of NMIA, a strong and reproducible pattern of 2'-*O*-adducts was detected by primer extension [compare black (+ NMIA reaction) with green (– NMIA) traces in Figure 2.4A]. The fraction of full-length extension product was 70% of that for the no reaction control, indicating that NMIA reacts with single-hit kinetics.[20] Sites of adduct formation were identified by comparison with a dideoxy sequencing lane (blue trace, Figure 2.4A).

Regions that showed the strongest reactivity relative to background are emphasized with red bars in Figure 2.4A and all reactivities are superimposed on a secondary structure for the RNA transcript in Figure 2.4B. Comparing the most reactive nucleotide, U33 in the anticodon loop, with unreactive residues in, for example, the T- and acceptor stems indicates that the dynamic range of the SHAPE experiment spans 30-50 fold. 2'-Hydroxyl reactivity observed in the 5' and 3' linkers is exactly that expected

for proper folding of the cassette, indicating the structure cassette folds as designed (Figure 2.4B).

When superimposed on the secondary structure for $tRNA^{Asp}$ and the flanking structure cassette, absolute 2'-hydroxyl reactivities correspond closely to whether or not a given nucleotide is constrained. Nucleotides that do not participate in either base pairing or a tertiary interaction are most reactive (red and orange columns in Figure 2.4B).

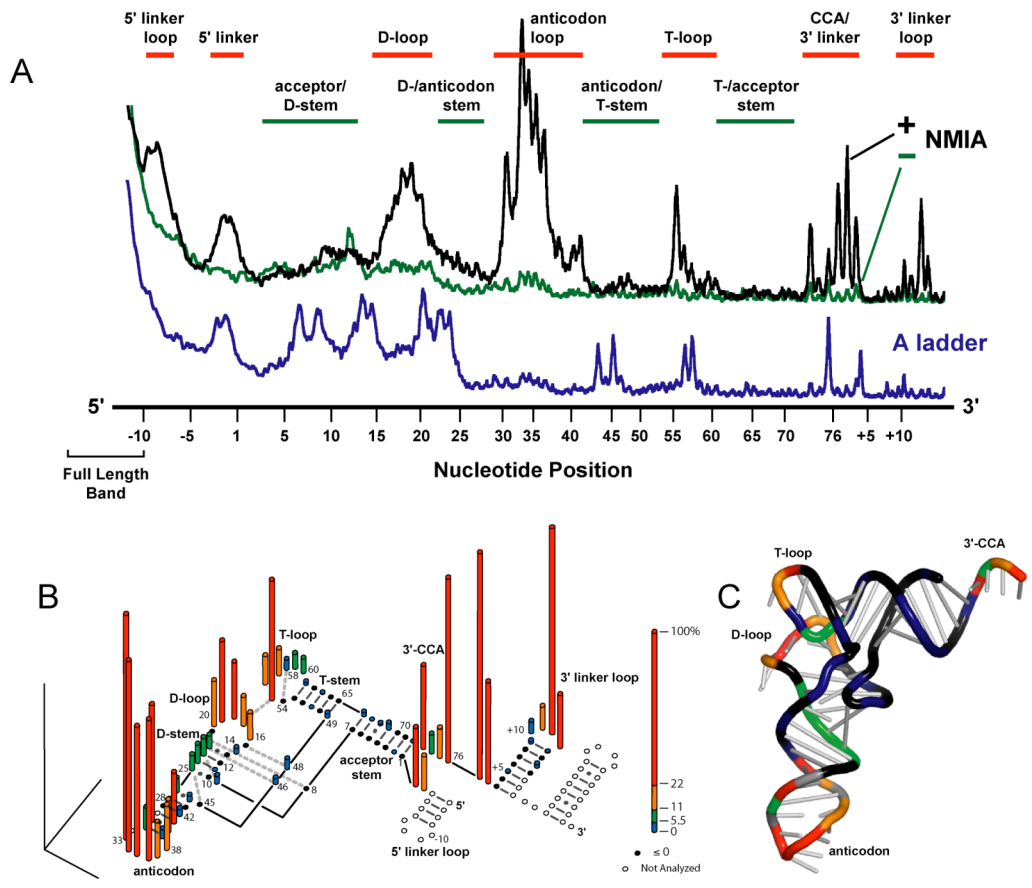


Fig. 2.4. SHAPE analysis of $tRNA^{Asp}$. (A) Profile for 2'-O-adduct formation as a function of nucleotide position, visualized by primer extension. Reactions performed in the presence and absence of NMIA are shown in black and green, respectively. Regions that react selectively with or are excluded from reaction with NMIA are emphasized with red or green bars, respectively. A dideoxy sequencing marker is shown in blue and is 1 nucleotide longer than corresponding positions in the (+) and (–) NMIA lanes. (B) Absolute 2'-hydroxyl reactivities superimposed on a secondary structure for the $tRNA^{Asp}$ construct. Column heights are proportional to the band intensities shown in part A (black trace) minus background. (C) SHAPE reactivities superimposed on the crystal structure of $tRNA^{Asp}$.[14,15] The phosphate-sugar backbone is colored according to SHAPE reactivity. The color scheme is the same as that in Figure 2.4B.

Conversely, NMIA modification is significantly reduced in each of the structures that comprise the four canonical tRNA helices (green, blue and black columns in Figure 2.4B, green, blue, and black colors, Figure 2.4C). Tertiary interactions including interaction of G45 with the G10:U25 pair, the A15:U48 Levitt pair, the A14:A21:U8 base triple, and the U54:A58 intraloop interaction are unreactive.

2'-Ribose-based NMIA chemistry is clearly generic to all RNA nucleotides. Each nucleotide (A,G,C,U) shows strong reactivity in at least one part of the $tRNA^{Asp}$ transcript and all four nucleotides show protection at other positions when constrained by base pairing or a tertiary interaction (Figure 2.4B, see Figure 2.3A for nucleotide assignments)

2.3 Discussion

2.3.1 2'-Hydroxyl acylation with NMIA scores local nucleotide flexibility.

RNA SHAPE footprinting reproduces the canonical structure of $tRNA^{Asp}$. Reactive nucleotides lie almost exclusively in the conserved T-, D- and anticodon loops and at the 3'-CCA terminus (red and orange colors in Figures 2.4B and C). Of the five non-canonical G-U or G-A pairs in $tRNA^{Asp}$, only one, in the anticodon stem, is reactive. SHAPE footprinting thus appears to be more sensitive to constraints imposed by non-canonical pairing than traditional chemistries because G-U and G-A pairs are typically reactive toward DMS and CMCT.[16,17,21]

NMIA reactivity also detects tertiary interactions. Essentially all nucleotides that form higher order interactions (dashed lines in Figure 2.4B) are unreactive at 35 °C. This

is a striking result given that the unreactive structures are constrained by diverse local tertiary interactions including Hoogsteen paring, base triples, and cross-loop interactions.

Finally, the SHAPE approach is also uniquely information-rich in the sense that the absolute reactivity amplitudes appear to be meaningful. First, nucleotides in the T- and D-loops, while generally reactive, are less so than those in the anticodon loops, consistent with formation of significant tertiary interactions in the former loops. Within the T-loop, the cross-loop U54-A58 interaction is reported by low reactivities in an otherwise reactive region. Second, nucleotides in the C-G rich T- and acceptor stems show measurably lower reactivities than many nucleotides in the D- and anticodon stems (Figure 2.4B and C). Third, inspection of the local patterns of reactivity within each of the loops in $tRNA^{Asp}$, at the 3'-CCA region, and in the 3'-cassette (the +10 to +14 loop) shows a reproducible trend in which the center-most loop nucleotides are more reactive than nucleotides that abut constrained positions in the adjacent duplexes. These data together emphasize that SHAPE mapping reveals significant fine gradations in local nucleotide flexibility.

2.3.2 NMIA modification does *not* score solvent accessibility. We compared the relative reactivities obtained by SHAPE with calculated molecular surface areas at the ribose 2'-position, assuming a solvent-sized probe and using the crystallographic coordinates [14,15] for $tRNA^{Asp}$ (Figure 2.5). Consistent with the results obtained with the 11-mer oligonucleotide,[2] 2'-O-adduct formation in $tRNA^{Asp}$ does not correlate with solvent accessibility. For example, the entire anticodon stem and loop structures, spanning nucleotides 26-44, have generally high solvent accessibility (≥ 20 $Å^2$ per ribose 2'-O atom) but feature both unreactive and highly reactive positions (compare gray bars

with black line in Figure 2.5). Analogously, both the D- and T-loops include residues with low calculated static solvent accessibility but high 2'-hydroxyl reactivity; conversely, the acceptor and T-stems (positions 49-53 and 60-72) are solvent accessible but unreactive (Figure 2.5). 2'-Hydroxyl reactivity is thus not correlated with static solvent accessibility, strongly supporting models (Figure 2.2) that emphasize local conformational changes as the primary determinant of 2'-hydroxyl reactivity.

We also compared SHAPE with chemical mapping experiments [16,17] employing the traditional base-reactive reagents, dimethyl sulfate (DMS) and diethylpyrocarbonate (DEPC). The correlation between nucleotides reactive by SHAPE and with DMS and DEPC is very strong (compare gray bars and filled circles in Figure 2.5). This is a striking conclusion considering that DMS and DEPC react largely at nucleophilic positions in purine bases [22] while NMIA reacts at the RNA backbone and strongly supports the model (Figure 2.2) that local nucleotide flexibility is the primary

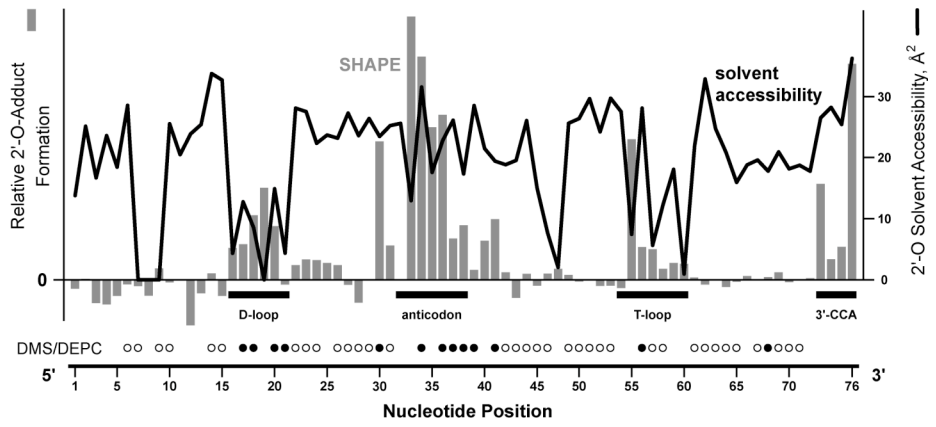


Fig. 2.5. RNA SHAPE reactivity is not correlated with solvent accessibility. Relative 2'-O-adduct formation (gray bars) were calculated from the data shown in Figure 2.4A. Solvent accessibilities (heavy lines) were computed assuming a probe radius of 1.4 Å. For comparison, results from chemical mapping studies on $tRNA^{Asp}$ transcripts using DMS and DEPC[16] are shown with spheres; reactive and unreactive positions are emphasized with filled and open spheres, respectively.

determinant for reactivity by SHAPE chemistry. The major difference occurs at position 68, where the guanosine residue in a G-U pair is reactive towards DMS, but not with the more selective SHAPE chemistry (Figure 2.5).

The direct comparison between traditional and SHAPE chemistries also emphasizes important advantages of SHAPE. First, all nucleotides are interrogated in a single reaction so there are no or fewer regions lacking structural information. Second, instead of simply classifying positions as reactive or not, SHAPE lends itself naturally to quantitative analysis (Figures 2.4B and 2.5).

2.3.3 2'-Ribose chemistry, local nucleotide flexibility, and the influence of the 3'-phosphodiester anion. Reaction of 2'-hydroxyl groups (Figure 2.6A) and of 2'-amine substitutions (Figure 2.6B) in RNA to form ester or amide linkages, respectively, are both strongly sensitive to local nucleotide flexibility (Figure 2.4B and references 1 and 4). We postulate that the physical basis for this effect in both cases stems predominantly from the influence of the proximal 3'-phosphodiester anion on reactivity at the 2'-position. The contribution of the 3'-phosphodiester potentially includes direct charge-charge electrostatic effects, differential hydrogen bonding, changes to the pK_a of the 2'-group, and generalized electrostatic and dielectric effects in the local environment.

However, the detailed reaction pathways for the 2'-ester versus the 2'-amide forming reactions are quite different. Reaction of a 2'-hydroxyl to form a 2'-ester proceeds via formation of an oxyanion in a *pre-equilibrium* step (Figure 2.2) and yields negative charge on the nucleophile in the transition state (Figure 2.6A). In contrast, 2'-amide formation involves *direct* attack of the unperturbed, 2'-NH_2 form [4,23] of the

nucleophile and a transition state in which there is a partial positive charge on the amine nucleophile (Figure 2.6B).

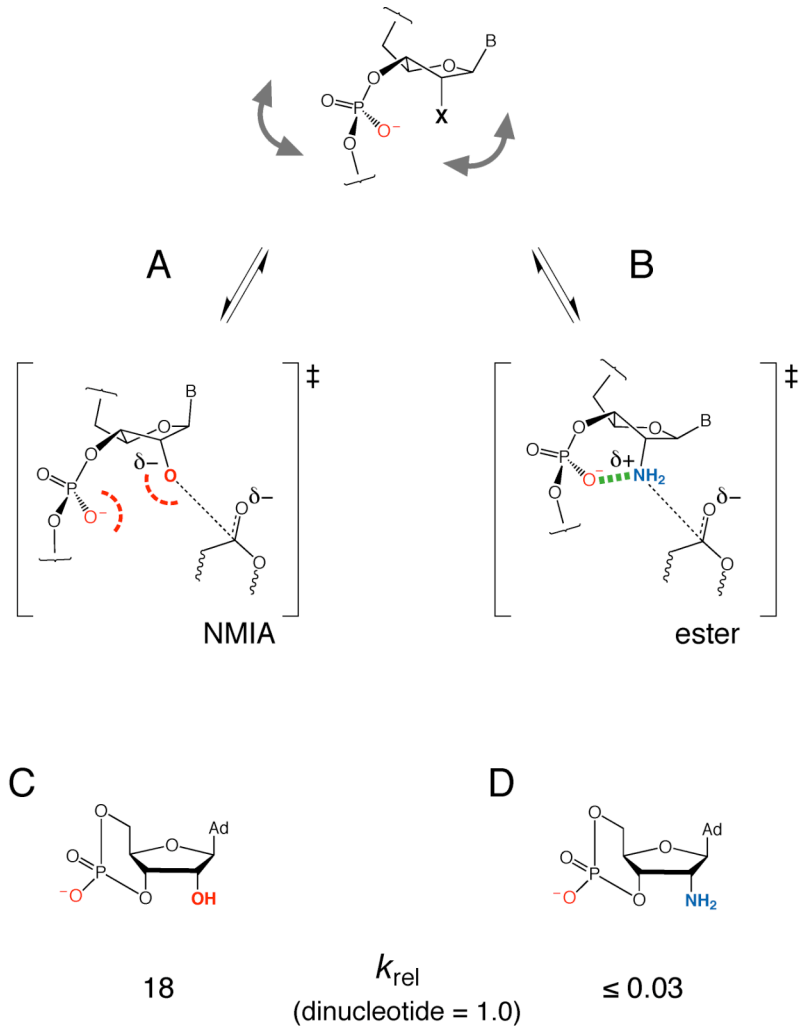


Fig. 2.6. Distinct contributions of the 3'-phosphodiester anion for reaction of a flexible (gray arrows) ribose 2'-hydroxyl versus 2'-amine. Reaction pathways for (A) 2'-hydroxyl and (B) 2'-amine acylation. (C,D) relative reactivities of cAMP analogs. Reference dinucleotides were pA$^{2'\text{-OH}}$p-ddC[2] and pU$^{2'\text{-NH2}}$pU,[2,4] respectively.

A proximal 3'-phosphodiester anion thus *destabilizes* formation of the 2'-oxyanion formed in a pre-equilibrium step (see interference symbol in Figure 2.6A), but *stabilizes* the amine nucleophile (dashed green line in the transition state in Figure 2.6B). These contrasting effects are supported directly by nucleotide analog studies[2,3] and especially by

cAMP analogs, which constrain the 3'-phosphodiester anion to be distant from a 2'-hydroxyl or 2'-amine nucleophile. Because the 3'-phosphodiester group electrostatically destabilizes the 2'-oxyanion, cAMP reacts (18-fold) faster than a reference dinucleotide (Figure 2.6C). In contrast, reaction of the 2'-amine, whose transition state is stabilized by the proximal anion, becomes undetectable (at least 30-fold slower) for 2'-amino-cAMP (Figure 2.6D). Flexible nucleotides sample a broad ensemble of conformations (gray arrows in Figure 2.6). Distinct constellations are able *both* to minimize the electrostatic influence of the 3'-phosphodiester and facilitate 2'-oxyanion formation or, alternately, to align the 3'-group and to facilitate the amide-forming transition state.

In summary, the nucleophilicity of the 2'-position in the context of an RNA nucleotide is gated by a complex energy landscape. Factors such as formation of proton resonance structures, distance between the 2'-position and the 3'-phosphate, and electrostatic effects all play a role in influencing the intrinsic reactivity. In the case of both 2'-amino RNA and unmodified RNA, constrained nucleotides do not sample conformations that enhance nucleophilicity. However, flexible nucleotides sample conformations that both enhance and reduce nucleophilicity. Therefore, on average, the reactivity of flexible nucleotides is higher than constrained nucleotides.

2.4 Perspective

SHAPE mapping takes advantage of the strong linkage between reactivity at the 2'-hydroxyl position and destabilization of nucleophilic chemistry at this site by the adjacent 3'-phosphodiester anion (Figures 2.2, 2.6 and reference 4). SHAPE footprinting affords a comprehensive view of nucleotide-resolution base pairing and tertiary

interactions in an RNA: nearly all nucleotides are interrogated in a single experiment. (Figure 2.4). SHAPE reactivity amplitudes also measure quantitatively subtle differences in the degree to which a given nucleotide is constrained. RNA structure mapping chemistries based on reaction at the 2'-hydroxyl have made possible facile nucleotide resolution analysis of RNA as a function of ion environment, temperature, ligand binding, and conformational switches.

2.5 A step-by-step guide to SHAPE chemistry

2.5.1 Requirements of SHAPE chemistry. The material requirements for a SHAPE experiment are modest.[2] A complete experiment requires 3-4 pmol of RNA. 2 pmol are required for the SHAPE chemistry itself and 1 pmol is required for each sequencing experiment used for band assignment. One or two sequencing experiments are usually sufficient. RNAs of any length are appropriate substrates for SHAPE, as long the RNA has no post-transcriptional modifications or unusually stable secondary structures that prevent its functioning as a template for primer extension. NMIA modification works well under a wide variety of solution conditions, ionic strength, and temperatures. SHAPE results have been obtained under a wide variety of chemical conditions and temperatures. Good results are obtained at 0-200 mM monovalent ion (NaCl, KCl or potassium acetate), 0-40 mM $MgCl_2$, and 20-75 °C (Figure 2.2C). In fact, titrations of magnesium and temperature using SHAPE chemistry have revealed complex structural rearrangements.[1,24]

In addition, an RNA may be modified in the presence of protein or other small and large biological ligands. Solution components that react directly with NMIA as well

as organic co-solvents, including formamide and DMSO, are well tolerated but may require that reagent concentrations be adjusted. NMIA reactivity is strongly dependent on pH; thus, the major experimental restriction is that the pH be maintained close to 8.0. In practice, pH values of 7.5–8.2 work well. The dynamic range that differentiates the most reactive (flexible) and least reactive (constrained) nucleotides typically spans a factor of 20-50.

2.5.2 RNA design. Because SHAPE reactivities are assessed in a primer extension reaction, information is lost at both the 5' end and near the primer binding site of an RNA. Typically, the 10-20 nucleotides adjacent to the primer binding site cannot be quantified due to the presence of cDNA fragments that reflect pausing by the reverse transcriptase enzyme during the initiation phase of primer extension. The 8-10 positions at the 5' end of the RNA also cannot be visualized due to the presence of the intense band corresponding to the full-length extension product.

To monitor SHAPE reactivities at the 5' and 3' ends of an sequence of interest, the RNA can either be (i) embedded within a larger fragment of the native sequence or (ii) placed between strongly folding RNA sequences that contain a unique primer binding site.[2] We have designed a structure cassette (Figure 2.3B) that contains 5' and 3' flanking sequences of 14 and 43 nucleotides and allows all positions within the RNA of interest to be evaluated in a sequencing gel. Both 5' and 3' extensions fold into stable hairpin structures that do not appear to interfere with folding of diverse internal RNAs.[2,25] The primer binding site of this cassette efficiently binds to a cDNA primer (Figure 2.3B). The sequence of any 5' and 3' structure cassette elements should be checked to ensure that they are not prone to forming stable base pairing interactions with the internal sequence.

2.5.3 RNA folding. The SHAPE experiment is most commonly performed with RNAs that have been generated by *in vitro* transcription.[2,25,26] These RNAs require purification by denaturing gel electrophoresis and then must be renatured to achieve a biologically relevant conformation. This protocol describes a simple approach that works well for renaturing many RNAs. However, any procedure that folds the RNA to the desired conformation at pH 8 can be substituted. The RNA is first heated and snap cooled in a low ionic strength buffer to eliminate multimeric forms. A folding solution is then added to allow the RNA to achieve an appropriate conformation and to prepare it for NMIA modification. The RNA is folded in a single reaction and is later separated into (+) and (-) NMIA reactions.

2.5.4 RNA modification. NMIA is added to the folded RNA to yield 2'-*O*-adducts at flexible nucleotide positions. The reaction is then incubated until essentially all of the NMIA has either reacted with the RNA or has degraded due to hydrolysis with water (see Figure 2.2). No specific quench step is required. The modified RNA is precipitated with ethanol in order to purify the RNA from reaction products and buffer components that may be detrimental to the primer extension reaction.

2.5.5 Primer extension. Although analysis of RNA adducts by primer extension is widely considered to be problematic, we find that the optimized experiment described here works well for most RNAs. Key innovations include use of an optimized primer binding site (Figure 2.3), thermostable reverse transcriptase enzyme, low [$MgCl_2$], elevated temperature, and short extension times. It is also essential to use intact, non-degraded RNA, free of reaction by-products and other small molecule contaminants. 5'-radiolabeled DNA primers are annealed to the RNA and then extended to sites of

modification in the presence of dNTPs by the activity of reverse transcriptase. The RNA component of the resulting RNA-cDNA hybrids is degraded by treatment with base. The cDNA fragments are resolved on a polyacrylamide sequencing gel.

2.5.6 Sequencing. Sequencing lanes generated by dideoxy nucleotide incorporation are used to assign bands in the (+) and (–) NMIA lanes. One or two sequencing reactions is usually sufficient to infer the entire sequence. These steps are conveniently performed concurrently with the primer extension reactions for the (+) and (–) NMIA tubes.

2.5.7 Reagents and materials. All reagents as well as reaction tubes and equipment must be maintained free of RNase contamination. For best results, all chemicals should be purchased at the highest quality available and reserved for RNA use only.

3.3× RNA folding mix (333 mM HEPES-NaOH, pH 8.0, 20 mM $MgCl_2$, 333 mM NaCl). Other conditions that are known to stabilize the functional structure of the RNA under study can be used as well. Both buffering component and ionic strength can be varied. In the modification reaction, buffer concentration should be at least twice the NMIA concentration and adjusted to pH 8.

10× NMIA in DMSO. The recommended concentration of this solution varies with RNA length. For RNA reads of 100, 200 and 300 nucleotides, 10× NMIA concentrations of 130, 65 and 30 mM, work well. Due to the solubility of the reagent, the stock concentration of NMIA should not be greater than 130 mM.

SHAPE enzyme mix (250 mM KCl, 167 mM TRIS-HCl, pH 8.3, 1.67 mM each dNTP, 17 mM DTT, 10 mM $MgCl_2$). The enzyme mix may be prepared by combining 4

parts SSIII FS Buffer , 1 part 0.1 M DTT, and 1 part 10 mM dNTP mix (10 mM in each deoxynucleotide; This solution is stable for months at –20 ˚C but is intolerant of freeze-thaw cycles. Maintaining small aliquots at –20 °C is recommended.)

5'-[^{32}P]-Labeled Primers. These are prepared by performing the following steps: (1) Mix the following well: 1 μL 60 μM DNA primer, 16 μL γ-[^{32}P]-ATP, 2 μL 10× PNK buffer, and 30 units[27] (1 μL) T4 Polynucleotide Kinase. (2) Incubate at 37 °C for 30 min. (3) Purify on 20% denaturing polyacrylamide gel (1× TBE, 7 M urea). Use autoradiography to visualize and excise the band corresponding to the radiolabeled DNA primer. (4) Passive elute overnight into water and remove small pieces of acrylamide from the RNA using a centrifugal filter device. (5) Recover radiolabeled DNA by ethanol precipitation (see step 8 of the PROCEDURE). (6) Dissolve the pellet in 100 μL 1 mM HEPES-NaOH, pH 8.0. The final primer solution concentration is ~ 0.3 μM.

2.5.8 Stepwise procedure for a SHAPE experiment.

1 Add 2 pmol RNA in 12 μL 0.5× TE (5 mM Tris-HCl, pH 8, 0.5 mM EDTA, pH 8) to a 200 μL thin-walled PCR tube.

2 Heat the RNA to 95 °C for 2 min; then, immediately place the RNA on ice for 2 min.

3 Add 6 μL folding mix and mix solutions by gentle repetitive pipetting.

4 Remove the tube from ice and incubate at the desired reaction temperature for 20 min in a programmable incubator. 37 °C is a good initial choice.

5 While the tube is incubating, remove 9 μL and place it in a second tube. One tube will be used for the (+) NMIA reaction, the other for the (–) NMIA reaction.

6 Add 1 μL of NMIA in DMSO to the (+) NMIA tube and 1 μL of neat DMSO to the (–) NMIA tube and mix well. Some initial turbidity in the tube is normal, especially when working at high NMIA concentrations.

7 Incubate the reaction for five NMIA hydrolysis half-lives. To estimate the NMIA half-life between 15 and 75 °C , use the empirical equation:

$$\text{half-life (min)} = 360 \times \exp[-0.102 \times \text{temperature (°C)}] \qquad (2.1)$$

At 37 °C, NMIA has a half-life of 8.3 min; therefore, at 37 °C, the reaction should be incubated for ~45 min.

8 After the reaction has gone to completion, transfer reactions to 1.5 mL centrifuge tubes, and recover the modified RNA by ethanol precipitation: (i) Add 90 μL water, 4 μL 5 M NaCl, 1 μL 20 mg/mL glycogen, and 350 μL absolute ethanol. (ii) Incubate at –80 °C for 30 min. (iii) Sediment the RNA by spinning at maximum speed in a microfuge at 4 °C for 30 min. Perform step 10(ii) during this centrifugation.

9 Remove ethanol supernatant; redissolve RNA in 10 μL of 0.5× TE and transfer samples to 200 μL thin-walled PCR tubes.

10 (i) Add 3 μL radiolabeled primer solution to the (+) and (–) NMIA tubes. Mix by repetitive pipetting. (ii) To sequence the RNA for assigning bands in the (+) and (-) NMIA samples, add 3 μL of primer solution to 1 pmol of RNA in 8 μL 0.5× TE.

11 Anneal the primer to the RNA by incubating tubes at 65 °C for 5 min and then at 35 °C for 5 min.

12 Add 6 μL of SHAPE enzyme mix to the (+) and (–) NMIA reactions. To each sequencing experiment, add one ddNTP solution (1-2 μL, 10 mM) as well.

13 Heat the tubes to 52 °C for 1 min.

14 Add 1 μL of Superscript III to each tube. Mix well by gentle repetitive pipetting. Immediately return the tube to the heat block.

15 Incubate tubes at 52 °C for 10 min.

16 Add 1 μL 4M NaOH. This degrades the RNA but does not damage DNA.

17 Heat the samples to 95 °C for 5 min.

18 Add 29 μL of acid stop mix.

19 Incubate at 95 °C for 5 min.

20 Load (+) NMIA, (–) NMIA, and sequencing reactions in individual lanes of a polyacrylamide sequencing gel (29:1 acrylamide:bis acrylamide, 1× TBE, 7 M urea). Load ~2 μL per lane. For extensions of 100 or fewer nucleotides, perform electrophoresis for 150 min at 70 W. To visualize RNA extension reactions spanning more than 100 nucleotides, re-load samples after 150 min in unoccupied lanes on the gel and continue electrophoresis for an additional 150 min at 70 W. The sample loaded first will have been subjected to electrophoresis for ~300 min, yielding well-resolved positions near the 5' end of the RNA.

21 Expose the gel overnight to a phosphor screen and quantify scanned bands using a phosphorimaging instrument. Quantify the intensity of every well-defined band in the gel for the (+) and (–) NMIA lanes by two-dimensional densitometry. This

step is conveniently performed using the SAFA (Semi-Automated Footprinting Analysis) program.[28]

22 Calculate absolute NMIA reactivity at each position in the RNA by subtracting (-) NMIA intensities from (+) NMIA intensities. (+) and (–) NMIA intensities should be normalized to each other by assuming the low intensity (unreactive) positions in each experiment have the same value. This calculation is equivalent to assuming that at least a few nucleotides will be unreactive in most RNAs. In assigning the SHAPE band positions, the cDNA markers generated by dideoxy sequencing are exactly 1 nucleotide longer than the corresponding (+) and (–) NMIA cDNAs.

2.5.9 Troubleshooting guide for SHAPE.

A variety of problems may be encountered in a SHAPE experiment (Table 2.1). Most problems may be diagnosed by inspection of the gel image. The nature of the problem is usually traceable to a unique step in the experiment or easily addressed problem.

Table 2.1

Troubleshooting guide for a SHAPE experiment.

Problem	**Reason**	**Solution**
Bright bands present in the (–) NMIA lane.	RNase contamination; this is, by far, the most common problem encountered for new practitioners of SHAPE technology.	Identify contaminated solution by running mock SHAPE experiments with [^{32}P]-labeled RNA.
	Structure-induced pausing by the reverse transcriptase enzyme.	Heat RNA in 0.5× TE to 95 °C for 3 min, cool on ice for 3 min before adding primer. Increase the time of extension to ~30 min. Try primer extension reactions at different temperatures (±3 °C).
Very low signal in the (+) NMIA lane but an intense full length product is observed.	Insufficient modification of RNA.	Perform modification using a 2-fold higher concentration of NMIA.
No full length product in the (+) NMIA lane or intense bands that disappear rapidly with read length.	Excessive modification of RNA.	Perform modification using a 2-fold lower concentration of NMIA.
No full length product in any lane.	Poor or incomplete primer extension.	The reverse transcriptase enzyme is very sensitive to $MgCl_2$ concentration. Make sure final solution conditions are 3 mM in $MgCl_2$. Enzyme is also sensitive to freezing. Retry the experiment carefully with fresh enzyme.
Smearing near the full length extension band.	Incomplete degradation of the RNA strand or re-extension of the 3'-end of the RNA in the RNA-DNA hybrid.	Decrease extension time; incubate at 95 °C with a higher concentration of NaOH.
No sequencing bands in ddNTP sequencing lane(s).	ddNTPs were incorporated in the wrong proportion.	Adjust the ddNTP concentrations upwards or downwards by 2-fold.
Extra bands in the ddNTP sequencing lane, not seen in (–) NMIA lane.	Pauses in extension due to misfolding.	Use a (–) NMIA reaction as the template for sequencing.
No extension in (+) or (–) NMIA lanes, but extension in sequencing lanes.	RNA lost during modification/precipitation.	Perform the ethanol precipitation step in the presence of 20 μg glycogen as a co-precipitant.
The faintest bands in the (+) NMIA lane have significantly different intensity as compared to the corresponding bands in the (-) NMIA lanes	Random error involved with volume measurement; gel loaded unevenly.	Small differences in background intensity can be treated by mathematical normalization (see step 22). If intensities are significantly different, the gel should be re-run by loading samples with similar amounts of cDNA fragments in each lane.

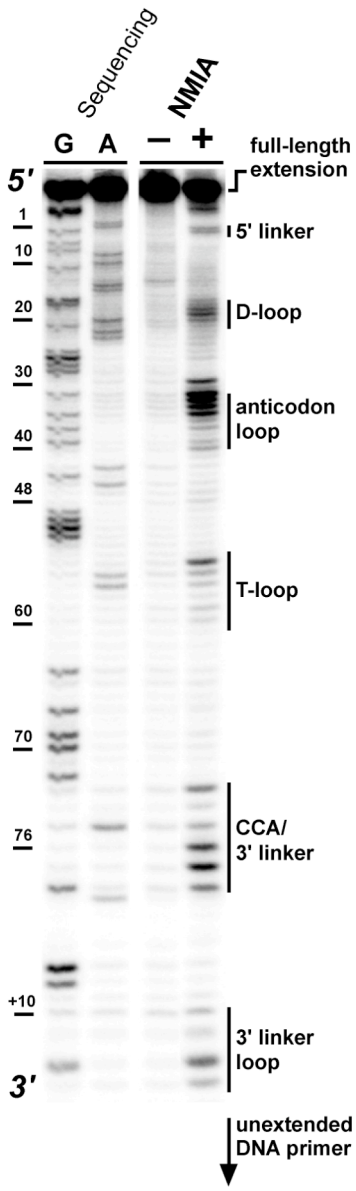


Fig. 2.7. Representative SHAPE analysis for yeast tRNA^{Asp}, embedded within a structure cassette. Typical results from a SHAPE experiment. G and A are sequencing reactions performed using cytosine and thymidine dideoxy terminating nucleotides. These marker lanes are exactly 1 nucleotide longer than the corresponding NMIA lanes. Nucleotide numbers (at left) correspond to the (–) and (+) reagent lanes. (–) and (+) NMIA lanes represent reactions differing only in the presence of the reagent. Regions expected to be unconstrained in the yeast tRNA^{Asp} construct are emphasized with bars at the right of the gel image.

2.5.10 Anticipated results.

A minimal SHAPE experiment consists of three or four lanes resolved in a sequencing gel (Figure 2.7). This representative experiment was performed using an *in vitro* transcript corresponding to yeast tRNA^{Asp} embedded within the structure cassette shown in Figure 2.3. Two sequencing lanes were used to assign the SHAPE reactivities observed in the (–) and (+) NMIA reagent lanes. The bright bands at the top of the gel correspond to the relatively abundant full-length extension product. Bands corresponding to the unextended DNA primer and to short extension products, caused by pausing of reverse transcriptase during initiation of primer extension, are too short to be observed in this gel image. Approximately 90 RNA nucleotides are sufficiently well resolved that their absolute SHAPE reactivities can be

quantified.

Positions in which SHAPE reactivity is significantly higher in the (+) NMIA reaction as compared to the no reagent (–) control are emphasized with vertical bars and correspond precisely to hairpin loops and unconstrained linker regions in the $tRNA^{Asp}$ construct (Figure 2.7). Band intensities can be quantified and absolute SHAPE reactivities at almost every position within the RNA are obtained by subtracting the (–) control intensities from the (+) NMIA intensities. Superposition of absolute band intensities on a secondary structure model for the $tRNA^{Asp}$ construct yields very precise information regarding the pattern of base pairing and the formation of non-canonical tertiary interactions in this RNA (bars, Figure 2.4B). Almost all base paired positions in $tRNA^{Asp}$ are unreactive; whereas nucleotides in the T-, D- and anticodon loops are reactive. For a more comprehensive discussion of SHAPE correlation with $tRNA^{Asp}$ structure, see Section 2.3.1

2.6 Experimental Section

2.6.1 Synthesis of $tRNA^{Asp}$ Construct. A DNA template for transcription of $tRNA^{Asp}$ in the context of the structure cassette was generated by PCR [1.5 mL; containing 20 mM Tris (pH 8.4), 50 mM KCl, 2.5 mM $MgCl_2$, 200 μM each dNTP, 500 pM each forward and reverse primer, 5 pM template and 0.025 units/μL Taq polymerase; denaturation at 94 °C, 45 s; annealing 55 °C, 30 s; and elongation 72 °C, 90 s; 34 cycles]. Primer and $tRNA^{Asp}$ coding sequences are indicated in Figure 2.3. The PCR product was recovered by ethanol precipitation and resuspended in 300 μL HE [10 mM Hepes-NaOH (pH 8.0), 1 mM EDTA, the template was the coding sequence of $tRNA^{Asp}$, and forward

and reverse primers contained the cassette sequences as well as ~18 nucleotides corresponding to the 5' and 3' ends of the coding sequence (Figure 2.3B)]. Transcription reactions (1.5 mL, 37 °C, 3 h) contained 40 mM Tris-HCl (pH 8.0), 10 mM $MgCl_2$, 10 mM DTT, 2 mM spermidine, 0.01% (v/v) Triton X-100, 4% (w/v) poly(ethylene) glycol 8000, 2 mM each NTP, 75 μL PCR-generated template, and 0.1 mg/mL T7 RNA polymerase. The RNA product was purified by denaturing (8%) polyacrylamide gel electrophoresis, excised from the gel, and recovered by electroelution and ethanol precipitation. The purified RNA (~ 2.5 μmol) was resuspended in 100 μL HE.

2.6.2 Structure-Sensitive RNA modification. RNA (20 pmol) in 6 μL sterile water was heated at 95 °C for 3 min, quickly cooled on ice, treated with 3 μL of folding buffer [333 mM NaCl, 333 mM Hepes-NaOH (pH 8.0), 33.3 mM $MgCl_2$], and incubated at 37 °C for 20 min. The RNA solution was treated with NMIA (1 μL, 130 mM in anhydrous DMSO, 35 °C), allowed to react for 54 min (equal to 5 NMIA half-lives), and placed on ice. Control reactions contained 1 μL of DMSO in place of NMIA. In prequench reactions, the RNA was added after NMIA was allowed to degrade by hydrolysis. Modified RNAs were subjected to primer extension without further purification.

2.6.3 Primer Extension. Modified RNA (2 μL, 4 pmol) was added to 5'-[^{32}P]-radiolabeled reverse transcription DNA primer (5'-[^{32}P]-GAA CCG GAC CGA AGC CCG, ~ 0.2 pmol in 11 μL H_2O), heated to 65 °C (6 min), and incubated at 35 °C (20 min). Reverse transcription buffer [6 μL; 167 mM Tris-HCl (pH 8.3), 250 mM KCl, 10 mM $MgCl_2$, 1.67 mM each dNTP, 16.7 mM DTT] was added; the RNA was heated to 52 °C; Superscript III (Invitrogen, 1 μL, 200 units) was added and mixed by gentle

pipetting; and reactions were incubated at 52 °C for 5 min. Primer extension reactions were quenched by addition of 1 μL of 4 M NaOH, heating (95 °C for 5 min), subsequent addition of neutralizing gel loading solution (29 μL; 40 mM Tris-borate, 276 mM Tris-HCl, 5 mM EDTA, 0.01% (w/v) each xylene cyanol and bromophenol blue, 73% (v/v) formamide), and heating (95 °C, for an additional 5 min). Dideoxy sequencing markers were generated using unmodified RNA and adding 1 μL 10 mM dideoxy nucleotide triphosphate after addition of reverse transcription buffer. cDNA extension products were separated by denaturing electrophoresis [90 mM Tris-borate, 2 mM EDTA, 7 M urea, 8% (29:1) acrylamide:bisacrylamide); 100 W, 5 min; 72 W, 2.5 h; 0.75 mm × 31 cm × 38.5 cm] and visualized by phosphorimaging. Two dimensional gel images were converted to plots of intensity versus position using ImageQuant. Individual band intensities for the (+) and (–) NMIA reactions were integrated using SAFA.[28]

2.7 References

1. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J. Am. Chem. Soc.* **127**, 4659-67 (2005).

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

3. Chamberlin, S. I. & Weeks, K. M. Mapping local nucleotide flexibility by selective acylation of 2'-amine substituted RNA. *J. Am. Chem. Soc.* **122**, 216-224 (2000).

4. Chamberlin, S. I., Merino, E. J. & Weeks, K. M. Catalysis of amide synthesis by RNA phosphodiester and hydroxyl groups. *Proc. Natl. Acad. Sci. USA* **99**, 14688-14693 (2002).

5. John, D. M. & Weeks, K. M. Tagging DNA mismatches by selective 2'-amine acylation. *Chem. Biol.* **7**, 405-410 (2000).

6. John, D. M. & Weeks, K. M. Chemical interrogation of mismatches in DNA-DNA and DNA-RNA duplexes under nonstringent conditions by selective 2'-amine acylation. *Biochemistry* **41**, 6866-6874 (2002).

7. Chamberlin, S. I. & Weeks, K. M. Differential helix stabilities and sites pre-organized for tertiary interactions revealed by monitoring local nucleotide flexibility in the bI5 group I intron RNA. *Biochemistry* **42**, 901-909 (2003).

8. John, D. M., Merino, E. J. & Weeks, K. M. Mechanics of DNA flexibility visualized by selective 2'-amine acylation at nucleotide bulges. *J. Mol. Biol.*, 611-619 (2004).

9. Moorman, A. R. & Abeles, R. H. New class of serine protease inactivators based on isatoic anhydride. *J. Am. Chem. Soc.* **104**, 6785-6786 (1982).

10. Hiratsuka, T. New ribose-modified fluorescent analogs of adenine and guanine nucleotides available as substrates for various enzymes. *Biochim. Biophys. Acta* **742**, 496-508 (1983).

11. Velikyan, I., Acharya, S., Trifonova, A., Foldesi, A. & Chattopadhyaya, J. The pK(a)'s of 2'-hydroxyl group in nucleosides and nucleotides. *J. Am. Chem. Soc.* **123**, 2893-4 (2001).

12. Preisler, R. S. et al. Premelting and the hydrogen-exchange open state in synthetic RNA duplexes. *Biopolymers* **23**, 2099-2125 (1984).

13. Snoussi, K. & Leroy, J.-L. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry* **40**, 8898-8904 (2001).

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

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

16. Perret, V. et al. Conformation in solution of yeast tRNA Asp transcripts deprived of modified nucleotides. *Biochimie* **72**, 735-744 (1990).

17. Romby, P., Moras, D., Dumas, P., Ebel, J. P. & Giege, R. Comparison of the tertiary structure of yeast tRNA(Asp) and tRNA(Phe) in solution. Chemical modification study of the bases. *J. Mol. Biol.* **195**, 193-204 (1987).

18. Tuerk, C. et al. CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci. USA* **85**, 1364-1368 (1988).

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

20. Brenowitz, M., Senear, D. F., Shea, M. A. & Ackers, G. K. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol.* **130**, 132-181 (1986).

21. Mathews, D. H., Disney, M. D., Childs J. L., Schroeder S. J., Zuker, M. & Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101**, 7287-7292 (2004).

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

23. Verheyden, J. P. H., Wagner, D. & Moffatt, J. G. Synthesis of some pyrimidine 2'-amino-2'-deoxynucleosides. *J. Org. Chem.* **36**, 250-254 (1971).

24. Wang, B., Wilkinson, K. A. & Weeks, K. M. in preparation. (2007).

25. Badorrek, C. S. & Weeks, K. M. RNA flexibility in the dimerization domain of a gamma retrovirus. *Nat. Chem. Biol.* **1**, 104-11 (2005).

26. Milligan, J. F. & Uhlenbeck, O. C. Synthesis of small RNAs using T7 RNA polymerase. *Meth. Enz.* **180**, 51-62 (1989).

27. One unit is defined as the amount of enzyme required to incorporate 1 nmol of radiolabeled ATP into DNA substrate in 30 min at 37°C. (USB corporation).

28. Das, R., Laederach, A., Pearlman, S., Herschlag, D. & Altman, R. B. SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA* **11**, 344-54 (2005).

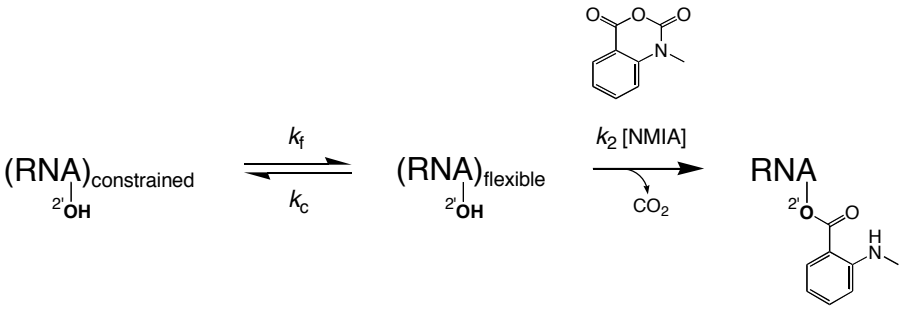
CHAPTER 3

NON-HIERARCHICAL INTERACTIONS DOMINATE EQUILIBRIUM STRUCTURAL TRANSITIONS IN $tRNA^{Asp}$ TRANSCRIPTS

3.1 Introduction

Essentially all RNA molecules have a propensity to form extensive interactions that locally constrain nucleotides in base-paired secondary structures and, often, in long-range tertiary interactions.[1-3] Nucleotides in folded RNAs thus differ significantly in the extent to which they are conformationally restrained by interactions with other parts of an RNA. Despite well documented exceptions,[4-6] current models for RNA folding and for *de novo* prediction of RNA structure emphasize that RNA folding is strongly hierarchical. Hierarchical folding means that the base-paired secondary structure is both more stable than and forms independently of the tertiary structure. The hierarchical model has been difficult to test thoroughly because it has not been possible to evaluate experimentally folding states at nucleotide resolution for large RNAs. Nucleotide resolution is essential because functionally important helices often span only a few base pairs,[2] and many important tertiary interactions in RNA involve only one or a few nucleotides.[7-9] In the protein folding world, local residue dynamics are elegantly monitored in backbone amide hydrogen exchange experiments.[10,11] A similarly comprehensive and nucleotide-resolution experiment has not been developed for analysis of RNA structure and RNA folding intermediates.

The ribose 2'-hydroxyl group reacts with N-methylisatoic anhydride (NMIA) to form ester adducts at the 2' position.[12-14] The reactivity of the 2'-hydroxyl group is gated by the underlying local structure such that flexible nucleotides react to form a 2'-O-adduct more readily than nucleotides constrained by base pairing or tertiary interactions,[12] as illustrated in Equation 3.1:



(3.1)

2'-Hydroxyl reactivity scores local nucleotide flexibility because unconstrained nucleotides are better able to adopt an open conformation in which formation of a 2'-oxyanion nucleophile is less destabilized by the adjacent 3'-phosphodiester anion.[12] Under our standard conditions, 2'-ester formation with NMIA (the pseudo-first-order k_2[NMIA] step) occurs at a rate of 0.02 min^{-1} at highly reactive 2'-hydroxyl positions. This rate is many orders of magnitude slower than the rate constants for base pair opening,[10,15] base flipping,[16] or local base destacking[17] processes, represented by the rate constants k_f and k_c in Equation 3.1. Studies with model compounds also indicate that flexible nucleotides react ~5–30-fold more slowly than nucleotide analogs specifically designed to reduce inhibition of the 2'-O-adduct forming reaction by the adjacent 3'-phosphodiester.[12] Thus, the equilibrium constant, represented by the ratio k_f/k_c, is less than 1, even for reactive, single-stranded positions. The rate constant for 2'-adduct formation, k_{adduct}, is therefore determined by the fraction of time a nucleotide exists in the unconstrained, reactive conformation, multiplied by the rate of the chemical

transformation:

$$k_{\text{adduct}} = K_{\text{f}}\, k_2[\text{NMIA}] \tag{3.2}$$

where $K_{\text{f}} = k_{\text{f}}/k_{\text{c}}$ and with the proviso that [NMIA] decreases during the reaction due to hydrolysis.[12] A significant, perhaps counterintuitive, bonus of NMIA chemistry is that this reagent degrades by competitive hydrolysis with water[12] which means that RNA structure can be monitored without the requirement that temperature-dependent reaction parameters be carefully controlled.

To use selective 2'-hydroxyl acylation to map RNA structure as a function of temperature, it is necessary to understand how the ability of 2'-*O*-adduct formation of nucleotides change with this variable. To measure ability of adduct formation as a function of temperature, the rates of hydrolysis and reaction with small nucleotide analogs were determined at different temperatures.[18] The results from these experiments indicate that NMIA degrades on an experimentally tractable timescale (Figure 2.2C), and thus RNA structure probing can be performed without an explicit reagent quenching step. It is sufficient instead simply to allow RNA modification to proceed for five hydrolysis half lives. Furthermore, the activation enthalpy is a constant over a variety of temperatures and the fraction 2'-O-adduct formed at long times for each of the four nucleotide analogues is independent of temperature. This indicates that the absolute 2'-O-adduct forming potential is a constant for an RNA nucleotide that does not undergo a change in conformation. If NMIA degradation proceeds to completion, no reactivity correction for temperature is required.

Selective 2'-hydroxyl acylation thus represents an RNA analog of the widely

employed protein amide hydrogen exchange[10,11] experiment in the bimolecular (EX2) limit. Because k_2[NMIA] is slow, NMIA-based interrogation of RNA folding states is restricted to analysis of folding intermediates at equilibrium. Each nucleotide in an RNA cycles through all possible states, including unfolded states, weighted by the Boltzmann distribution. Flexible nucleotides will populate a different distribution of conformations than base paired or constrained positions. The conformation that most facilitates 2'-O-adduct formation will dominate relative reactivity at any given position and flexible nucleotides will more often sample the most reactive conformations.[12] Because all ribonucleotides have a 2'-hydroxyl, except for a few with post-transcriptional modifications at this position,[19,20] structural information is, in principle, obtainable for every nucleotide in an RNA.

In this chapter, I use RNA SHAPE mapping to analyze unfolding profiles for yeast $tRNA^{Asp}$ transcripts with simultaneous interrogation of the local structural environment at almost every nucleotide. Even for this relatively simple and well studied RNA,[12,21-25] SHAPE analysis reveals a pathway for unfolding intermediates that is significantly richer and more detailed than previously observable. The most thermally accessible states involve strong and non-hierarchical linkages between disparate elements of the tRNA structure.

3.2 Results

3.2.1 Nucleotide-resolution analysis of RNA folding intermediates. We used SHAPE to analyze thermally accessible states for yeast $tRNA^{Asp}$ transcripts. Although this RNA does not contain the post-transcriptional modifications of the cellular RNA,

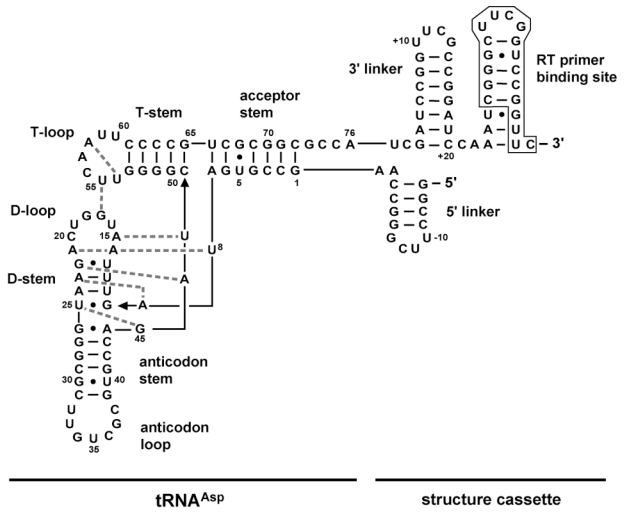


Fig 3.1. tRNA^{Asp} transcript in the context of an RNA structure cassette. tRNA^{Asp} structure is drawn to suggest orientation of helices and tertiary interactions in three-dimensional space. Tertiary interactions are indicated by dashed lines in gray. 5' and 3' linker regions facilitate analysis of sites of 2'-O-adduct formation by primer extension and have been shown to fold independently of the internal tRNA structure.[12] By convention,[21,22] nucleotides in yeast tRNA^{Asp} are numbered 1-76, with residue 47 omitted.

extensive prior work shows that the unmodified version folds into the correct L-shaped tRNA structure[23,24] and is bound and aminoacylated by its cognate synthetase indistinguishably from the wild type RNA.[26,27] tRNA^{Asp} was synthesized in the context of an RNA structure cassette (Figure 3.1). This cassette embeds the tRNA within flanking 5' and 3' linker structures that facilitate detection of sites of 2'-O-adduct formation by primer extension.[12] SHAPE analysis shows that the internal tRNA^{Asp} structure folds autonomously and that the flanking hairpins fold independently, as designed.[12]

The structure of the tRNA^{Asp} construct was analyzed by SHAPE at temperatures spanning 35 to 75 °C in intervals of ≤ 3.5 °C (Figure 3.2, left-most 18 lanes). The RNA was treated with NMIA at each temperature under single hit conditions[12] and until reagent degradation was 97% complete (5 half-lives). Sites of 2'-O-adduct formation were mapped by primer extension. Strong stops to primer extension are not observed in control experiments omitting NMIA or in which NMIA was allowed to degrade prior to addition of RNA (compare –NMIA and pre-quench lanes with +NMIA reactions, Figure 3.2).

Overall banding patterns for the 18 independent experiments show clear differences that reflect secondary and tertiary interactions and smooth transitions as a

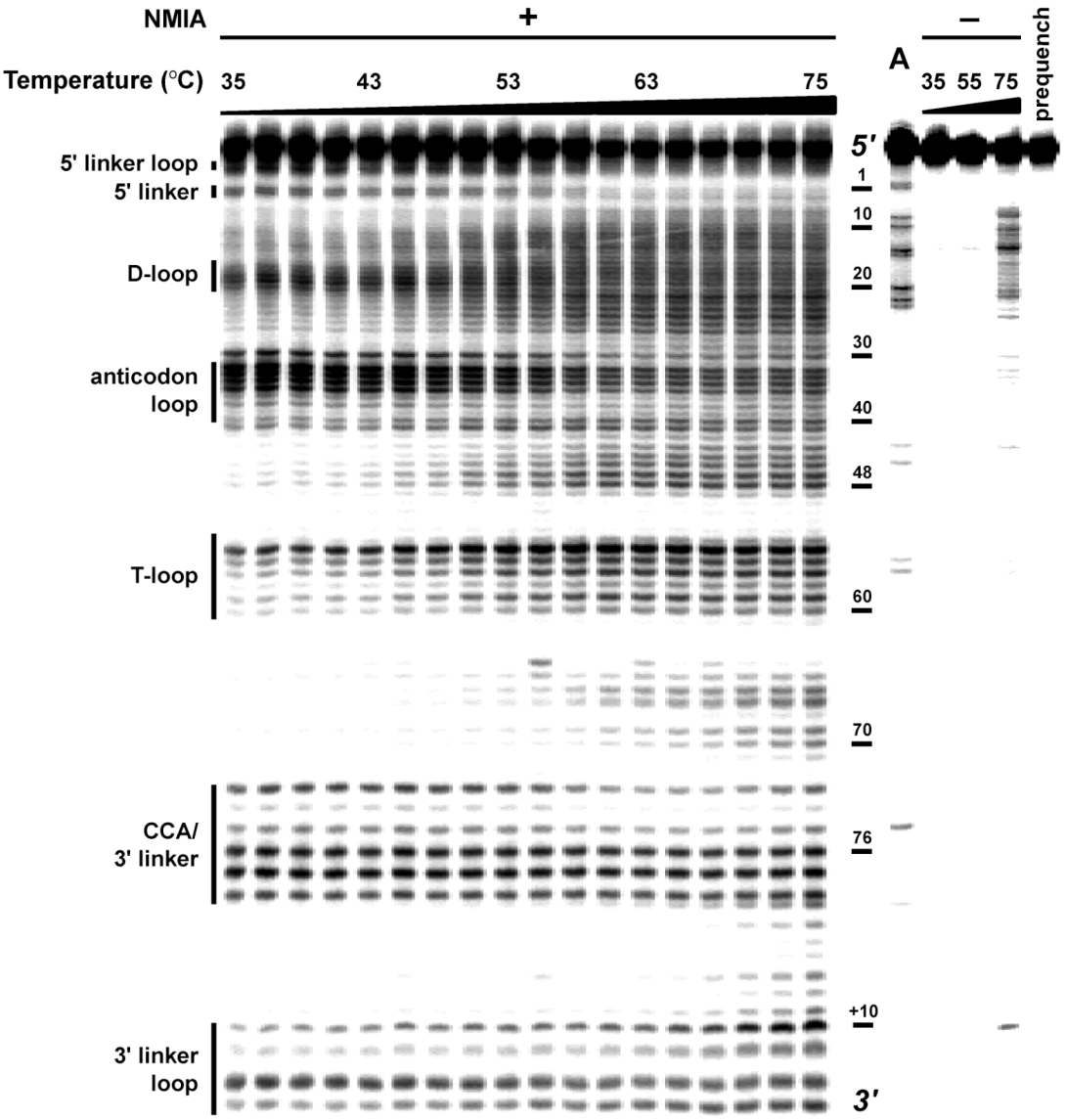


Fig. 3.2. Temperature dependent SHAPE analysis of tRNA^{Asp}. RNA structural landmarks are indicated at left. For the (+) NMIA reactions, bands correspond to stops in reverse transcriptase-mediated primer extension and report sites of 2'-O-adduct formation; (–) NMIA lanes omitted reagent. Bands in the (-) NMIA lane represent small amounts of degredation as due to heating the tRNA in the presence of magnesium. The band intensities are significantly lower than their corresponding positions in the (+) NMIA lanes. The dideoxy sequencing ladder (A) is offset exactly 1 nucleotide longer than the corresponding 2'-adduct bands; nucleotide positions are numbered relative to NMIA lanes. For the prequench lane, NMIA was added and allowed to degrade prior to addition of RNA.

function of temperature (Figure 3.2, structural landmarks are indicated at left). The SHAPE experiment accurately reports the canonical structure of $tRNA^{Asp}$ at 35 °C and 10 mM $MgCl_2$ because all highly reactive nucleotides lie in loops in $tRNA^{Asp}$ or the structure cassette. With increasing temperature, the $tRNA^{Asp}$ transcript more readily formed 2'-*O*-adducts, consistent with increased local nucleotide flexibility in the RNA (Figure 3.2).

We quantified the folding state for almost every nucleotide in the RNA as a function of temperature. Band intensities were normalized to a unit scale to facilitate comparison between different nucleotides. We observed five distinct reactivity patterns (Figure 3.3). In the first group (Figure 3.3A), 10 positions including all residues in the T-stem, as well as C2, C29, and C72, remained unreactive at all temperatures measured in this experiment (see 75 °C lane, Figure 3.2; summarized as filled stars in Figure 3.4). These residues correspond to positions that remain stably base paired below 75 °C.These nucleotides represent positions that have some of the strongest interactions in the entire RNA construct.

In the second class (Figure 3.3B), ten nucleotides in $tRNA^{Asp}$ (G18, U19, C20, U33, G34, U35, G37, C38, C75, and A76) exhibited roughly uniform high reactivity at all temperatures. These nucleotides are located in the anticodon loop, in the D-loop, and at the 3'-CCA terminus and correspond precisely to flexible, unpaired regions (summarized as red circles in Figure 3.4).

For the third and fourth classes, many positions underwent well-behaved transitions in which 2'-hydroxyl reactivity changes smoothly between 35 and 75 °C. However, the $tRNA^{Asp}$ construct does not unfold in a single transition or even in a few

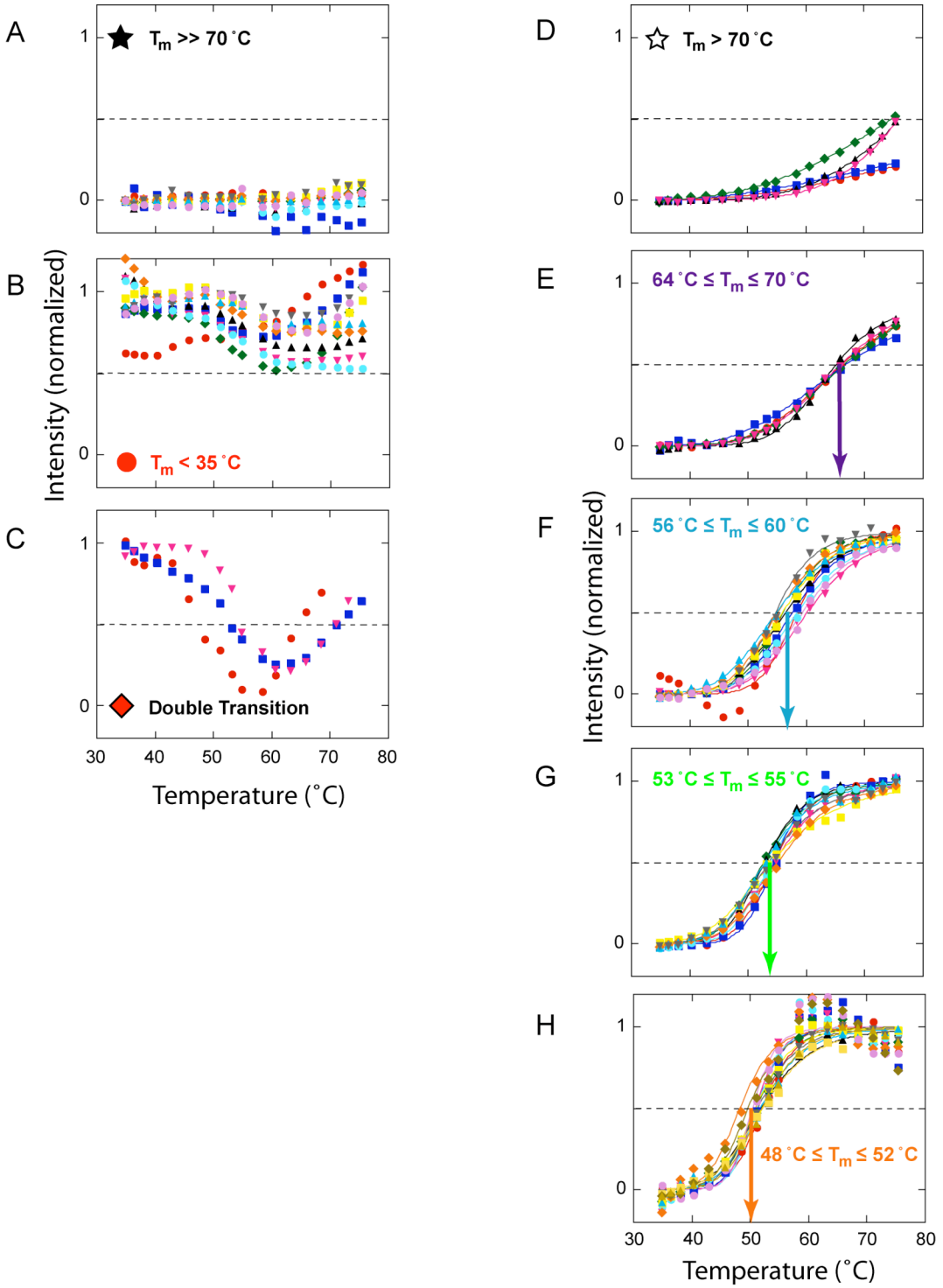


Figure 3.3. Single nucleotide unfolding profiles for $tRNA^{Asp}$. Profiles were computed by quantifying almost every band shown in Figure 3.2. (A-H) Profiles for each nucleotide binned by behavior. Data are plotted on a linear scale; dashed line indicates fraction melted = 0.5. Nucleotides with well-defined unfolding transitions (E-H) were fit to an equation assuming unimolecular denaturation and grouped according to T_m. Each group transition is indicated by a specific color (arrows, D-H) that will be used throughout this work. To determine which nucleotides are in each group transition, refer to Figure 3.4.

transitions. Instead, SHAPE quantitatively identified a minimum of six structural transitions for $tRNA^{Asp}$ spanning ~50 to >70 °C (Figures 3.3A,D-H). Transition midpoints (melting temperatures, T_m) measured at each nucleotide are summarized in Figure 3.4. We group the observed transition midpoints by color and this scheme is used consistently throughout this work.

Structures in the third class show relatively simple behavior in which reactivity increased from a low to a high value (see, for example, D-stem residues 10-13 and 22-25 in Figure 3.2) consistent with a monotonic transition from a structure constrained by base pairing or tertiary interactions to an unconstrained conformation (indicated by T_m in Figure 3.4).

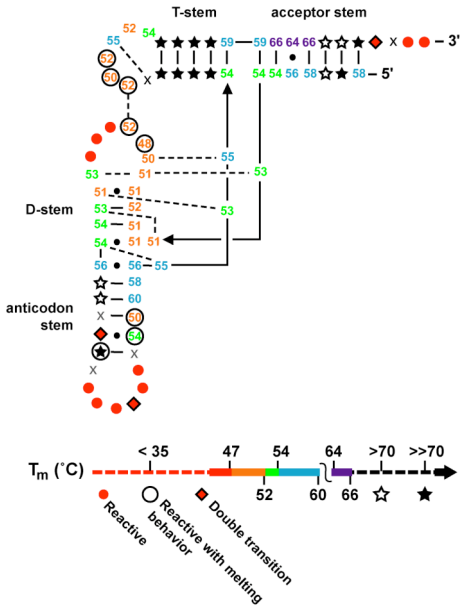


Figure 3.4. Summary of structural transitions at nucleotide resolution superimposed on a secondary structure for $tRNA^{Asp}$. Nucleotides are represented by their transition midpoint (T_m). Positions that have just begun to unfold at 70 °C or that unfold with a T_m >>70 °C are shown with open and closed stars, respectively. Colors correspond to melting temperature scheme used in previous and subsequent figures. Positions C28, U32, G39 and U54 (x) could not be analyzed, either because of electrophoretic band compression[25] or absence of a reproducible profile.

In contrast, in the fourth class, several nucleotides located in the D- and T-loops were reactive at 35 °C but additionally exhibited clear unfolding transitions with increasing temperature (see Figure 3.2; emphasized with a circled T_m in Figure 3.4). We infer that these are partially constrained positions that then experience greater conformational flexibility with increasing temperature.

Finally, three positions were initially reactive but have structural profiles that showed

complex two-phase transitions (for example, position G73 in Figure 3.2). At these nucleotides, 2'-hydroxyl reactivity first decreased and then increased (Figure 3.3C, diamonds in Figure 3.4). The net amplitude of the *reduction* in 2'-hydroxyl reactivity is comparable to that observed for the multiple, simple transitions in which reactivity *increases*. The observed, well defined, reduction and subsequent enhancement in reactivity is consistent with a model in which these nucleotides participate in new, constraining, interactions at intermediate temperatures which are then destabilized with increasing temperature.

We were also able to monitor 2'-hydroxyl reactivity in significant portions of the 5' and 3' regions of the structure cassette (Figures 3.1 and 3.2). All of the nucleotides in the 5' and 3' linker loops were reactive and all base-paired nucleotides in these structures showrd either no unfolding transition or an incomplete transition by 75 °C (comparable to the transitions shown in Figures 3.3A and 3.3D). Thus, helices in the structure cassette fold independently, as designed, and do not interfere with folding of the internal $tRNA^{Asp}$.

3.2.2 Comparison of SHAPE with absorbance-detected denaturation. We compared the nucleotide resolution information obtainable by the SHAPE approach (Figure 3.3) with a conventional absorbance-detected experiment, performed under identical solution conditions (Figure 3.5). RNA absorbance in the ultraviolet region increases with temperature due to disruption of base-stacking interactions (open symbols, Figure 3.5). Global unfolding features are more readily visualized in a first derivative plot (solid symbols, Figure 3.5), which shows two broad denaturation transitions for the $tRNA^{Asp}$ transcript spanning ~ 45-75 and 75-90 °C.

The SHAPE experiments agree well with the absorbance melting profile (transitions detected by SHAPE are emphasized in color in Figure 3.5). The transitions detected at 51 and 58 °C correspond to features (just barely) detectable in the absorbance melting experiment. The transition maximum seen in the first derivative plot of the absorbance melting experiment (~ 63 °C) does not correspond to a single RNA unfolding transition but, instead, appears to report a convolution of the SHAPE-detected transitions occurring at 58 and 66 °C (compare blue and magenta regions with solid symbols in Figure 3.5). In sum, RNA SHAPE mapping is quantitatively consistent with the bulk RNA absorbance melting experiment but additionally yields structural information at nucleotide resolution.

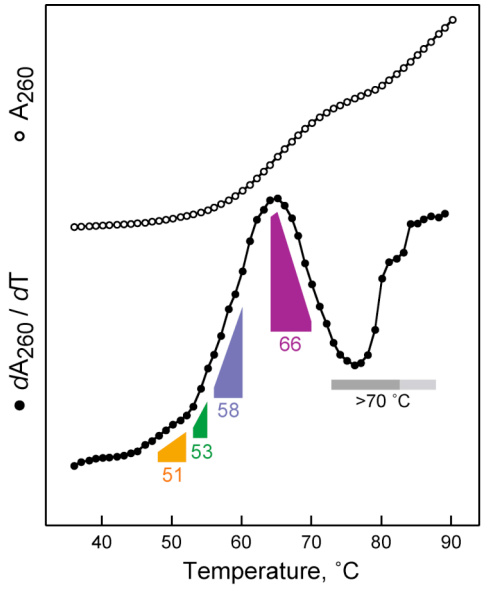


Fig. 3.5. Absorbance melting profile of the tRNA^{Asp} transcript. Absorbance, monitored at 260 nm, and the first derivative with respect to temperature are shown with open and closed symbols, respectively. Nucleotide resolution transition ranges determined by SHAPE (Figure 3.3) are illustrated in color. RNA was incubated in 10 mM $MgCl_2$ and 100 mM NaCl. Temperature was increased at a rate of 1 ˚C per minute.

3.3 Discussion

3.3.1 Model for the unfolding pathway of tRNA^{Asp} at single nucleotide resolution.

The nucleotide-resolution SHAPE approach provides model-independent information on the local nucleotide environment in tRNA^{Asp} as a function of temperature (Figures 3.2, 3.3, and 3.4). In order to interpret these data and to develop structural models for tRNA^{Asp} folding intermediates, unfolding profiles for individual nucleotides

were placed into eight groups (Figures 3.3 and 3.4). There is relatively little subjectivity in these assignments. First, calculated T_m's have uncertainties that are ± 1 °C or less at most positions. Second, nucleotides binned into groups by T_m alone are located adjacent to each other in $tRNA^{Asp}$, suggesting that the binning process reports authentic independently folding regions in the RNA (Figure 3.4).

At 35 °C, SHAPE accurately reports the canonical structure[21,22] of $tRNA^{Asp}$ (summarized in Figure 7A). Nucleotides in the T-, D- and anticodon loops and in the 3'-CCA region were reactive; whereas, nucleotides involved in base pairing and tertiary interactions were unreactive. The anticodon stem is the least constrained of the four conserved paired regions and nucleotides in or adjacent to the G-U pair in this stem were also reactive (red nucleotides in Figure 3.6A).

3.3.2 Two-phase loss of tertiary interactions for $tRNA^{Asp}$. The first unfolding transition (T_m ~ 51 °C) involves disruption of the tertiary interactions that link the D- and T-loops (in orange, Figure 3.6B). In addition, nucleotides A9–U13, that span one strand of the D-stem, became flexible in this transition. U55–A57 and U59 in the T-loop and nucleotides A14–G17 in the D loop were moderately reactive at 35 °C and become more reactive over this temperature range. We infer that these positions are moderately flexible in the native ground state and are cooperatively disrupted in this transition.

The second transition (T_m ~ 53 °C) involves complete unfolding of the D-stem, coupled with disruption of the tertiary interactions between the purine-rich strand (positions G22-U25) of this stem and nucleotides U8, A21, G45 and A46 (green positions in Figure 3.6C). In addition, nucleotides that comprise the stacked junction between the T- and acceptor stems (G6, A7 and C49) become flexible in this transition. Inspection of

the melting profiles makes clear that the 51 and 53 °C transitions are distinct (Figures 3.3G,H). The 51 and 53 °C transitions are also detected independently in the absorbance melting experiment (Figure 3.5).

An unanticipated result is thus that the purine- and pyrimidine-rich strands of the D-stem become conformationally flexible in an asymmetric way, average T_m's are 53 and 51 °C, respectively. Only the pyrimidine-rich strand (nucleotides G10–U13) becomes unconstrained in the first transition and in concert with the T- and D-loops as judged by SHAPE chemistry (illustrated by removal of base pair symbols at orange nucleotides in Figure 3.6B). The purine strand melts with most of the tertiary interactions in this region at 53 °C (green symbols, Figure 3.6C). The asymmetric stability of the purine versus pyrimidine strands of the D-stem likely reflects two local structural differences. First, purine nucleotides may be better able to maintain inter-nucleotide stacking interactions in the absence of base pairing. Second, only the purine-rich side participates in significant tertiary interactions with other parts of the tRNA (dashed lines in Figure 3.6B).

3.3.3 A conformational switch involving the anticodon and acceptor stems. In the next well isolated transition ($T_m \sim 58$ °C) both the anticodon and acceptor stems unfold *asymmetrically*. One strand of each helix unfolds ≥ 7 °C higher than the other (compare T_m's in blue with other symbols in Figure 3.4). In addition, over this same temperature range, positions G30, C36 and G73 exhibit significant *decreases* in reactivity (Figure 3.3C, diamonds in Figure 3.4) and thus structural constraints *increase* at these well defined positions in the RNA (emphasized with blue downward arrowheads in Figure 3.6D).

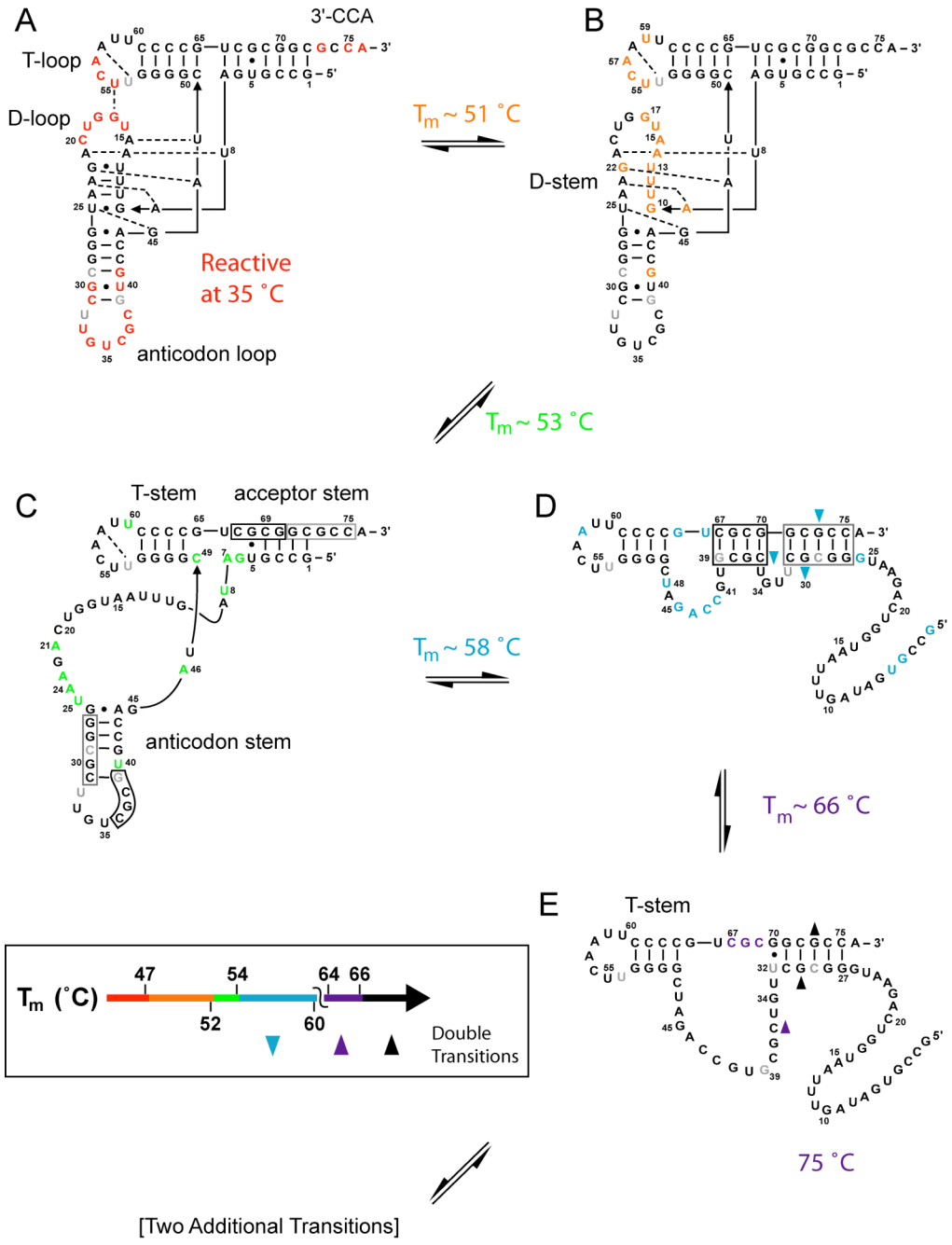


Fig. 3.6. Nucleotide-resolution model for thermal denaturation of tRNA^{Asp}. For each structure or intermediate (A-E), nucleotides that become newly reactive in a given transition are shown in color. Gray and black boxes (C,D) identify nucleotides proposed to participate in a conformational shift that maximizes G-C pairings. Up and down arrowheads emphasize positions that show two-phase behavior (see Figure 3.2C) and whose reactivity increases or decreases, respectively, in the indicated transition. Two additional transitions with T_m's ≥ 70 °C, involving disruption of the G27–U32/G70–C75 helix and the T-stem helix, are not shown.

These two, distinct, observations are well explained if nucleotides on one side of each of the anticodon and acceptor stems (emphasized with gray and black boxes, Figure 3.6C) undergo a conformational shift to form new G-C rich helices (gray and black boxes, Figure 3.6D). Thus, as one side of the acceptor and anticodon stems becomes locally flexible and reactive towards NMIA, the other side forms new constraining base pairs with complementary sequences in the RNA. This new structure (Figure 3.6D) is favored at higher temperatures because the absence of native tertiary interactions, that would otherwise enforce the canonical tRNA fold (see Figure 3.6A), changes the minimum energy structure. Alternatively, the conformational shift may reflect, in part, a temperature dependent change in ΔH for these secondary structure motifs.

Nucleotides G37, C38, C74 and C75 likely also participate in the new helices. These nucleotides were partially reactive, at ~10-50% of the rate of the fully reactive A76 position, over the entire temperature range (Figure 3.2). We interpret this conserved, intermediate reactivity to indicate that these nucleotides experience a series of stacking and base pairing constraining interactions throughout the temperature range.

3.3.4 Ultimate stages of unfolding in three incremental steps. With increasing temperature, the helices formed as a result of the conformational shift (Figure 3.6D) then unfold incrementally and in an order consistent with calculated thermodynamic parameters.[28-30] The short helix spanning G36–G39 and C67–G70 unfolds with a T_m of ~66 °C (magenta positions in Figure 3.6E). The stem comprised of G27–U32 and G70–C75 unfolds next because most of these nucleotides begin to react with NMIA by 75 °C (Figure 3.3D, illustrated with black up arrowheads in Figure 3.6E). Finally, the C:G rich T-stem unfolds with a T_m that must be significantly higher than 75 °C, consistent with the

observation that no detectable change in reactivity occurs at these positions through 75 °C (Figure 3.3A; summarized with solid stars in Figure 3.4).

3.3.5 Implications for the RNA folding problem. The single-nucleotide resolution SHAPE analysis illustrates a complexity in folding states for $tRNA^{Asp}$ transcripts that is not anticipated by current models for RNA folding. First, RNA folding is thought to be strongly hierarchical such that the base-paired secondary structure is more stable than and forms independently of tertiary interactions.[1,3] For the $tRNA^{Asp}$ construct, this model is clearly only partially predictive. It is true that the two lowest temperature transitions, at 51 and 53 °C, both involve loss of tertiary interactions. However, in both cases disruption of tertiary structure is coupled with the loss of simpler base-paired interactions (see Figures 3.6B and 3.6C). This coupling is so tight that the base paired D-stem is better viewed as an obligate component of the tRNA tertiary structure.

Second, the structural stability of an individual helix is usually approximated to be uncoupled from the structure of other helices, with the exception of end-stacking effects.[30-32] However, for $tRNA^{Asp}$, there are two distinct examples in which the two strands in an RNA helix unfold asymmetrically due to structural coupling with other parts of the molecule. In the first example, the purine-rich strand of the D-stem becomes conformationally flexible at a higher temperature than the pyrimidine-rich strand of the same helix due to the influence of the tRNA tertiary structure (see Figures 3.6B and 3.6C). Second, both the anticodon and acceptor stems melt asymmetrically, consistent with a conformational switch (Figures 3.6C and 3.6D) in which the partially denatured tRNA refolds to an alternate three-helix structure at elevated temperature.

3.3.6 Facile analysis of RNA folding pathways. SHAPE chemistry makes possible quantitative analysis of RNA base pairing and tertiary interactions at single nucleotide resolution. Because NMIA reacts at the 2'-hydroxyl position, all nucleotides are interrogated and the *absence* of reactivity can be confidently assigned to persistent stability at a given local structure. The fundamental conclusion of this work is that, even for relatively simple $tRNA^{Asp}$ transcripts, the equilibrium conformational states most structurally accessible from the native state are not well predicted by the hierarchical model for RNA folding. We are hopeful that the comprehensive and quantitative view of RNA structure afforded by the SHAPE approach will provide the experimental framework necessary for understanding the interrelationships between local and higher order folding in RNA.

3.4 Experimental Section

3.4.1 General. The $tRNA^{Asp}$ transcript was synthesized by in vitro transcription in the context of a structure cassette containing flanking 5' and 3' sequences (Figure 2) to facilitate analysis of the entire RNA by primer extension.[12] The RNA was purified by denaturing electrophoresis and stored in 10 mM Hepes-NaOH (pH 8.0), 1 mM EDTA. All experiments were performed in 100 mM Hepes (pH 8.0), 100 mM NaCl, 10 mM $MgCl_2$, 10% (v/v) dimethyl sulfoxide (DMSO), and 13 mM NMIA (Molecular Probes).

3.4.2 Temperature-dependent RNA modification. RNA (6 μL, 20 pmol; in ~2 mM Hepes, pH 7.5) was heated at 95 °C for 3 min, cooled on ice, treated with 3 μL of folding buffer [333 mM NaCl, 333 mM Hepes-NaOH (pH 8.0), 33.3 mM $MgCl_2$], and incubated at 37 °C for 20 min. The RNA solution was then equilibrated for 5 min at the

reaction temperature (35-75 °C), treated with NMIA (1 μL, 130 mM in anhydrous DMSO), allowed to react for 5 hydrolysis half-lives (54 to 1.2 min for 35 to 75 °C, respectively), and placed on ice. No NMIA controls and prequench reactions were performed by substituting DMSO alone for NMIA or by allowing the NMIA to degrade prior to addition of RNA, respectively. Temperature gradient was established using an Eppendorf Gradient Thermocycler using 45 ± 10 and 65 ± 10 °C gradient settings; temperatures used were 34.9, 36.3, 38.0, 40.3, 42.9, 45.7, 48.4, 51.0, 53.2, 54.8, 58.3, 60.6, 63.2, 65.9, 68.5, 71.0, 73.1, and 75.4 °C (see Figure 3.2). Efficiency of 2'-*O*-adduct formation was obtained from primer extension reactions, performed exactly as described in Chapter 2 and resolved on (8%) denaturing polyacrylamide gels.

3.4.3 Quantification of RNA reactivity. The gel image was quantified by phosphorimaging and band intensities integrated using SAFA.[33] Nucleotide reactivities were normalized to a uniformly reactive position (G18 for nucleotides G1–C49; A76 for G50–C75); absolute reactivities typically increased 5 to 8-fold for nucleotides that showed well defined melting transitions. Data obtained as a function of temperature at individual nucleotides were rescaled to a unit (0→1) scale and smoothed using a rolling weighting function.[34] Transition midpoints (T_m) were obtained assuming a unimolecular transition:[35]

$$I = A\frac{1}{1+\left[\exp\left[\frac{\Delta H_{vH}}{R}\left(\frac{1}{T_m}-\frac{1}{T}\right)\right]\right]^{-1}} + b \qquad (3.3)$$

where I is the band intensity at a given temperature (T), R is the gas constant, and A and b are the transition amplitude and initial intensity, respectively. The van't Hoff enthalpy (ΔH_{vH}) is returned by this equation but is characterized by large fitting errors. In contrast,

melting temperatures are generally (34 of 41 positions) reproducible to ± 1 °C or better. Error limits for nucleotides U11, U16, G65, C67, and G68 are ± 2 ˚C and for U40 and C43 are ± 3 °C.

3.4.4 Monitoring RNA denaturation by UV Absorbance. The $tRNA^{Asp}$ construct (5 μM, 600 μL, 100 mM NaCl, 10 mM $MgCl_2$) was heated from 35 to 90 ˚C in an Applied Photophysics Pistar-180 spectrometer at 1 °C/min under exactly the conditions used in the SHAPE experiments (including DMSO, but omitting NMIA). After subtracting background from a sample omitting RNA, the denaturation profile, monitored at 260 nm, was smoothed[34] and algebraically differentiated with respect to temperature.

3.5 References

1. Brion, P. & Westhof, E. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 113-137 (1997).

2. Gesteland, R. F., Cech, T. R. & Atkins, J. F. (eds.) *The RNA World* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1999).

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

4. LeCuyer, K. A. & Crothers, D. M. The Leptomonas collosoma spliced leader RNA can switch between two alternate structural forms. *Biochemistry* **25**, 5301-11 (1993).

5. Gluick, T. C. & Draper, D. E. Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* **241**, 246-62 (1994).

6. Wu, M. & Tinoco, I. RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci. USA* **95**, 11555-60 (1998).

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

8. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. & Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA* **98**, 4899-903 (2001).

9. Leontis, N. B., Stombaugh, J. & Westhof, E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucl. Acids Res.* **30**, 3497-531 (2002).

10. Englander, S. W. & Kallenbach, N. R. Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.* **16**, 521-655 (1983).

11. Englander, S. W. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 213-238 (2000).

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

13. Moorman, A. R. & Abeles, R. H. New class of serine protease inactivators based on isatoic anhydride. *J. Am. Chem. Soc.* **104**, 6785-6786 (1982).

14. Hiratsuka, T. New ribose-modified fluorescent analogs of adenine and guanine nucleotides available as substrates for various enzymes. *Biochim. Biophys. Acta* **742**, 496-508 (1983).

15. Snoussi, K. & Leroy, J.-L. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry* **40**, 8898-8904 (2001).

16. Spies, M. A. & Schowen, R. L. The trapping of a spontaneously "flipped-out" base from double helical nucleic acids by host-guest complexation with beta-cyclodextrin: the intrinsic base-flipping rate constant for DNA and RNA. *J. Am. Chem. Soc.* **124**, 14049-53 (2002).

17. Jean, J. M. & Hall, K. B. Stacking-unstacking dynamics of oligodeoxynucleotide trimers. *Biochemistry* **43**, 10277-84 (2004).

18. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J. Am. Chem. Soc.* **127**, 4659-67 (2005).

19. Decatur, W. A. & Fournier, M. J. RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.* **278**, 695-8 (2003).

20. Agris, P. F. Decoding the genome: a modified view. *Nucl. Acids Res.* **32**, 223-38 (2004).

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

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

23. Romby, P., Moras, D., Dumas, P., Ebel, J. P. & Giege, R. Comparison of the tertiary structure of yeast tRNA(Asp) and tRNA(Phe) in solution. Chemical modification study of the bases. *J. Mol. Biol.* **195**, 193-204 (1987).

24. Perret, V., Garcia, A., Puglisi, J., Grosjean, H. Ebel, J. P., Florentz, C. & Geige, R. Conformation in solution of yeast tRNA Asp transcripts deprived of modified nucleotides. *Biochimie* **72**, 735-744 (1990).

25. Chamberlin, S. I. & Weeks, K. M. Mapping local nucleotide flexibility by selective acylation of 2'-amine substituted RNA. *J. Am. Chem. Soc.* **122**, 216-224 (2000).

26. Perret, V., Garcia, A., Grosjean, H., Ebel, J. P., Florentz C. & Geige, R. Relaxation of a transfer RNA specificity by removal of modified nucleotides. *Nature* **344**, 787-9 (1990).

27. Sissler, M., Eriani, G., Martin, F., Giege, R. & Florentz, C. Mirror image alternative interaction patterns of the same tRNA with either class I arginyl-tRNA synthetase or class II aspartyl-tRNA synthetase. *Nucl. Acids Res.* **25**, 4899-906 (1997).

28. Jaeger, J. A., Turner, D. H. & Zuker, M. Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol.* **183**, 281-306 (1990).

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

30. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. & Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101**, 7287-7292 (2004).

31. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Muller, P., Mathews, D. H. & Zuker, M. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA* **91**, 9218-22 (1994).

32. Draper, D. E. & Gluick, T. C. Melting studies of RNA unfolding and RNA-ligand interactions. *Methods Enzymol.* **259**, 281-305 (1995).

33. Laederach, A., Das, R., Pearlman, S., Herschlag, D. & Altman, R. SAFA: Semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA*, **11**, 344-354 (2005).

34. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. *Graphical Methods for Data Analysis, pp. 91-104* (Wadsworth International Group, Belmont, CA, 1983).

35. John, D. M. & Weeks, K. M. van't Hoff enthalpies without baselines. *Protein Sci.* **9**, 1416-1419 (2000).

CHAPTER 4

HIGH THROUGHPUT SHAPE (hSHAPE)

4.1 Introduction

SHAPE chemistry is a useful tool to analyze the structure of any RNA. SHAPE interrogates nearly all nucleotides in an RNA, and can be performed under a wide variety of chemical and physical conditions to develop structural constraints. In a typical SHAPE experiment, RNA is sparsely modified in a structure-sensitive manner by an appropriate electrophile. The adducts, which form preferentially at flexible nucleotides, are detected by their ability to inhibit primer extension by reverse transcriptase. A control reaction omitting NMIA to assess background, as well as dideoxy sequencing extensions to assign nucleotide positions, are performed in parallel. The resulting cDNAs are quantified on a denaturing electrophoresis gel, where position and degree of modification correlates with length and amount of the extended primer.

Denaturing slab-gel electrophoresis is a powerful tool for separating nucleic acids by length. However, the production and imaging of gels is a labor-intensive task, and band resolution can be poor near the origin of separation. Software that quantifies gel electrophoresis images, such as SAFA,[1] typically cannot resolve and quantify more than 200 bands per separation at single nucleotide resolution. However, there is nothing in the other steps of SHAPE that prevents analysis of reads that are several times as long.

A more appealing method to size and quantify extended cDNA primers is to employ a capillary electrophoresis instrument of the type commonly used for DNA sequencing. These automated instruments require less hands-on effort, and can separate, at single nucleotide resolution, fluorescently labeled DNA 100-700 nucleotides long. Additionally, raw elution data can be obtained in approximately 20% of the time it takes to produce an analagous gel image.

The number of nucleotides interrogated in a single SHAPE experiment depends not only on the detection and resolution of separation technology used but also on the nature of RNA modification. Given reaction conditions, there is a length where nearly all RNA molecules have at least one modification. As primer extension reaches these lengths, the amount of extending cDNA decreases, which attenuates experimental signal. Adjusting conditions to decrease modification yield can increase readlength. However, lowering reagent yield also decreases the measured signal for each cDNA length. Given these considerations, the practical limit of a SHAPE experiment is probably around 1 kilobase of RNA.

In this chapter, I describe a high-throughput SHAPE experiment (hSHAPE). hSHAPE required the development of techniques and software to produce and quantify the cDNA products of a SHAPE experiment using a capillary electrophoresis system. To accomplish these goals, this work was completed with significant contributions from Morgan Giddings, Nicolas Guex, and Suzy Vasa.

I use data from an analysis that encompasses the 5'-most ~300 nucleotides of an HIV-1 genome transcript to illustrate the principles of an hSHAPE experiment. This

region encompasses several structural domains, such as the TAR, primer binding site, and the ψ site (described in Chapter 1).

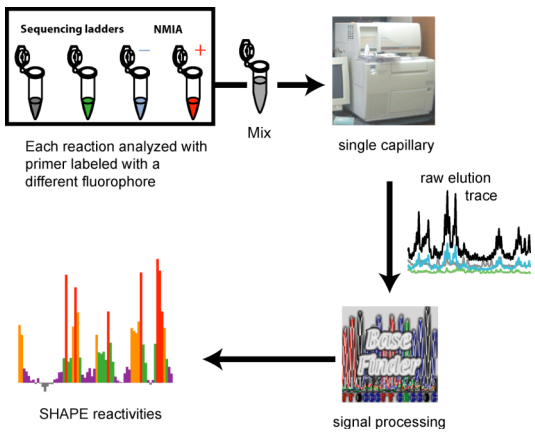


Fig. 4.1: Steps of an hSHAPE experiment. Sequencing ladders and (+) and (-) NMIA extensions are performed using different fluorophores. Resulting cDNAs are separated on a DNA sequencer. Raw elution traces are analyzed by Basefinder.[2] Correction for signal decay and normalization yield absolute SHAPE reactivity as a function of nucleotide position.

4.2 SHAPE and automated DNA sequencers

An hSHAPE experiment comprises four different reactions – a (+) NMIA, a (-) NMIA control and two dideoxy sequencing reactions (Figure 4.1). Each of these extension reactions is performed using a 5'- fluorophore labeled DNA primer (Figure 4.1). The reaction and extension conditions are almost identical to a gel-based experiment, except that primer concentration is on the order of RNA concentration to ensure readable signal (For an overview of modification and extension steps, see Chapter 2). The fluorophores employed by hSHAPE are identical to the dyes normally used for DNA sequencing. The products of the extensions are combined and purified by recovery with ethanol precipitation and resolved in a single multi-fluor run by automated capillary electrophoresis.

4.3 Analysis of hSHAPE data

4.3.1 Processing of raw elution traces. The resulting raw elution trace for the 5'-end of the HIV-1 genome resembles a DNA sequencing experiment in that it reflects the products of specific primer extension termination events (Figure 4.2A). However, in an hSHAPE experiment, the absolute peak intensities as well as the elution times of peaks are meaningful in the (+) and (–) NMIA traces. For example, missing peaks or peaks with low reactivity in the (+) NMIA trace correspond to RNA nucleotides constrained by base pairing or other interactions. Intense peaks in the (+) NMIA trace identify unstructured or flexible nucleotides in the HIV-1 RNA. Elution times indicate the position of reactive and unreactive nucleotides.

Each hSHAPE experiment contains a large peak at the low elution time that corresponds to unextended primers (Figure 4.2A). A large peak corresponding to full length RNA is observed at long elution times if the read extends to the 5'-end of an RNA. Between these two peaks, quantitative, single-nucleotide resolution RNA structure is obtained.

We employ the signal processing framework of BaseFinder[2] to analyze raw fluorescence versus elution time profiles. BaseFinder is a modular, extensible software package originally designed for DNA base calling and sequencing analysis. BaseFinder functions by applying a sequence of tools to a data trace. Each tool performs a specific analysis step, and contains adjustable parameters to account for experimental and stochastic variables, such as dye set and fluorescent baseline.

The initial processing steps of raw sequencer traces are identical to those used for DNA sequencing. Fluorescent baseline is subtracted for each channel. Next, color

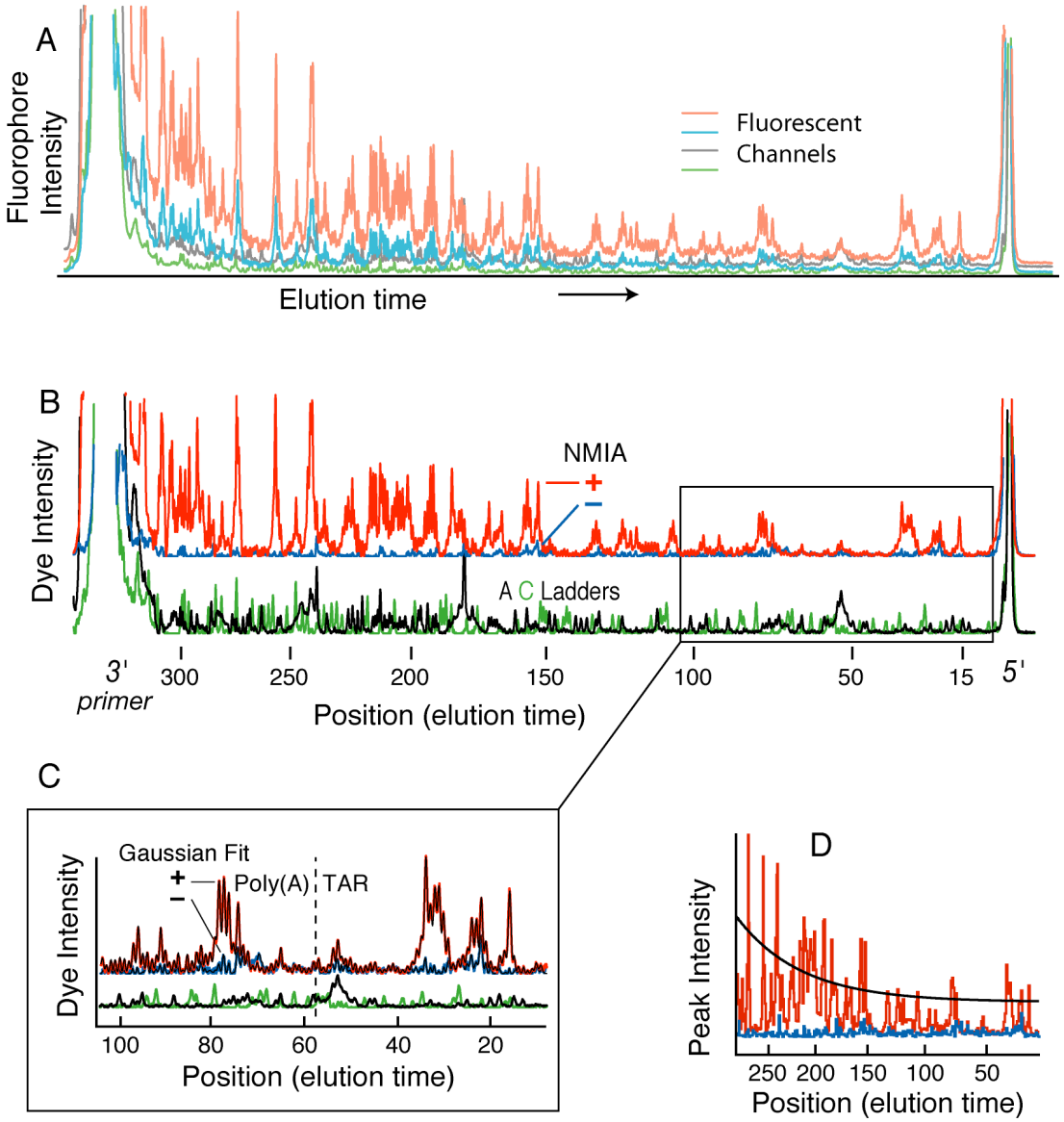


Fig. 4.2: Signal processing of HIV-1 genome structural data using Basefinder[2]. Raw elution traces are analyzed to generate structural constraints for nearly every nucleotide in an RNA. (**A**) Raw fluorescent intensity versus time plot for HIV-1 genomic transcripts. Significant spectral overlap and baseline contributions are observed in the light red and light blue traces. (**B**) Correction for baseline, spectral overlap, and mobility yields traces corresponding to each of the reactions in a SHAPE experiment. (+) and (–) NMIA reactions (red and blue) are offset from the A and C sequencing lanes (black and green) for clarity. (**C**) Whole trace peak integration of (+) and (–) NMIA traces. Peaks were fit to Gaussian curves (black dashed lines) and aligned to the sequencing channels. Peaks in the sequencing channels are exactly 1 nucleotide longer than the (+) and (–) NMIA traces. (**D**) Signal decay correction for the (+) NMIA trace assuming a constant probability for extension at each nucleotide (black line).

separation is performed to correct for spectral overlap of the multiple dyes such that each channel reports quantitative cDNA amounts (compare light blue line in Figure 4.2A to blue line in Figure 4.2B). The final analysis step common to DNA sequencing is the alignment of corresponding peaks in the four channels because each fluorophore imparts a slightly different electrophoretic mobility on cDNAs of the same length. The result of these analysis steps (Figure 4.2B) is an aligned plot of dye amount versus elution time for all the reactions in the SHAPE experiment. Each peak represents the amount of cDNA of a specific length. Corresponding peaks in all 4 traces are aligned so that they have the same elution time.

Mobility shift and color separation parameters for a specific dye set may be generated on BaseFinder by careful analysis of separate RNA sequencing experiments. To develop color separation parameters for each dye, spectral overlap in each channel is determined in the absence of other fluorophores by analysis of a single nucleotide ladder. To develop mobility parameters, each of the different fluorophores is used to generate the same nucleotide ladder from the same RNA template. The ladders are separated in the same capillary column. Mobility shifts are determined by matching corresponding sequencing peaks throughout the read. Mobility and color separation parameters are specific to a dye set, and may be used on multiple RNA reads.

4.3.2 Quantification of sequencer data. Unique analysis steps are required for *quantifying* cDNA amounts in the (+) and (–) NMIA data traces to develop RNA structural constraints. Unlike DNA sequencing, where peak position is the most important factor, both the location and intensity of peaks in the NMIA data traces are important to locate and quantify nucleotide flexibility.

We have developed a new BaseFinder tool, called Align and Integrate[3], that calculates peak area in the (+) and (–) NMIA traces versus nucleotide sequence. First, Align and Integrate detects and aligns peaks in the (+) and (–) NMIA traces with the RNA sequencing traces. Second, sequencing traces are compared and aligned with the sequence of the RNA being studied. Align and Integrate automatically accounts for the observation that cDNAs generated by sequencing are exactly 1 nucleotide longer than corresponding positions in the (+) and (–) NMIA traces.[4,5] Finally, areas under each peak are determined by performing a whole trace Gaussian-fit integration. The overall result of applying BaseFinder to raw SHAPE traces is a set of (+) and (–) NMIA trace peak areas for every nucleotide position in the read (Figures 4.1 and 4.2C and D).

Inspection of the resulting intensity data indicates signal decay associated with the (+) NMIA trace (Figure 4.2D). This signal likely reflects both the nature of NMIA reactivity as well as imperfect processivity of the reverse transcriptase enzyme. We corrected the drop by assuming that the probability of extension at each nucleotide was constant and slightly less than one:

$$D = A\,p^{(\text{elution time})} + C \qquad (4.1)$$

Where D is the signal decay adjustment factor, A and C are scaling factors that reflect the arbitrary initial and final intensities of the trace, and p is the probability of extension at each nucleotide. Typical values for p span 0.995-0.999 for elution times in units of 2×s. We apply this equation to peak intensities representing average reactive nucleotides throughout the trace. The 2% of the most highly reactive peaks as well as peaks with reactivities near zero are excluded from the calculation. Each peak intensity calculated at the same elution time was divided by D. Signal decay correction results in an unbiased

data set that does not lose overall intensity as a function of readlength (compare Figures 4.2D and 4.3A) Although uncommon, signal decay can also occur in the (–) if overall peak intensity is high in that trace. The steps to correct decay are the same as those for the (+) NMIA trace.

Because the (+) and (-) NMIA extensions are performed independently and use sequencing dyes with different quantum yields and spectral properties, the absolute scale for the (+) and (-) NMIA peak intensities are different. In order to quantitatively compare (+) and (-) NMIA peak intensities, we assumed that the peaks with the lowest ~10% of intensities throughout the (+) NMIA data accurately reflected the intensity of the corresponding (-) NMIA traces. We multiplied all peak intensities in the (-) NMIA trace by an appropriate factor that matched intensities to the unreactive nucleotides in the trace. The approach is insensitive to the dyes chosen for the (+) and (-) NMIA extensions. Indeed, interchanging the dyes used for the extensions produces nearly identical results.

4.3.3 hSHAPE authentically measures local nucleotide flexibility. Several observations strongly support use of these processing steps and a very high level of precision for the hSHAPE experiment. Highly reactive nucleotides have similar SHAPE reactivities, independent of whether they lie at the 5' or 3' end of the RNA (red bars, Figure 4.3A). Absolute SHAPE reactivities, superimposed on the well-characterized[6-8] TAR and Poly(A) stem loops (nts 1-104) of the HIV-1 genome, show that SHAPE information is exactly consistent with the consensus secondary structure for this region (Figure 4.3B). Nucleotides in loops are reactive (red and orange bars, Figure 4.3B); whereas, base paired nucleotides are unreactive (purple bars, Figure 4.3B). Notably, SHAPE reactivities also accurately report fine-scale structural differences. For example,

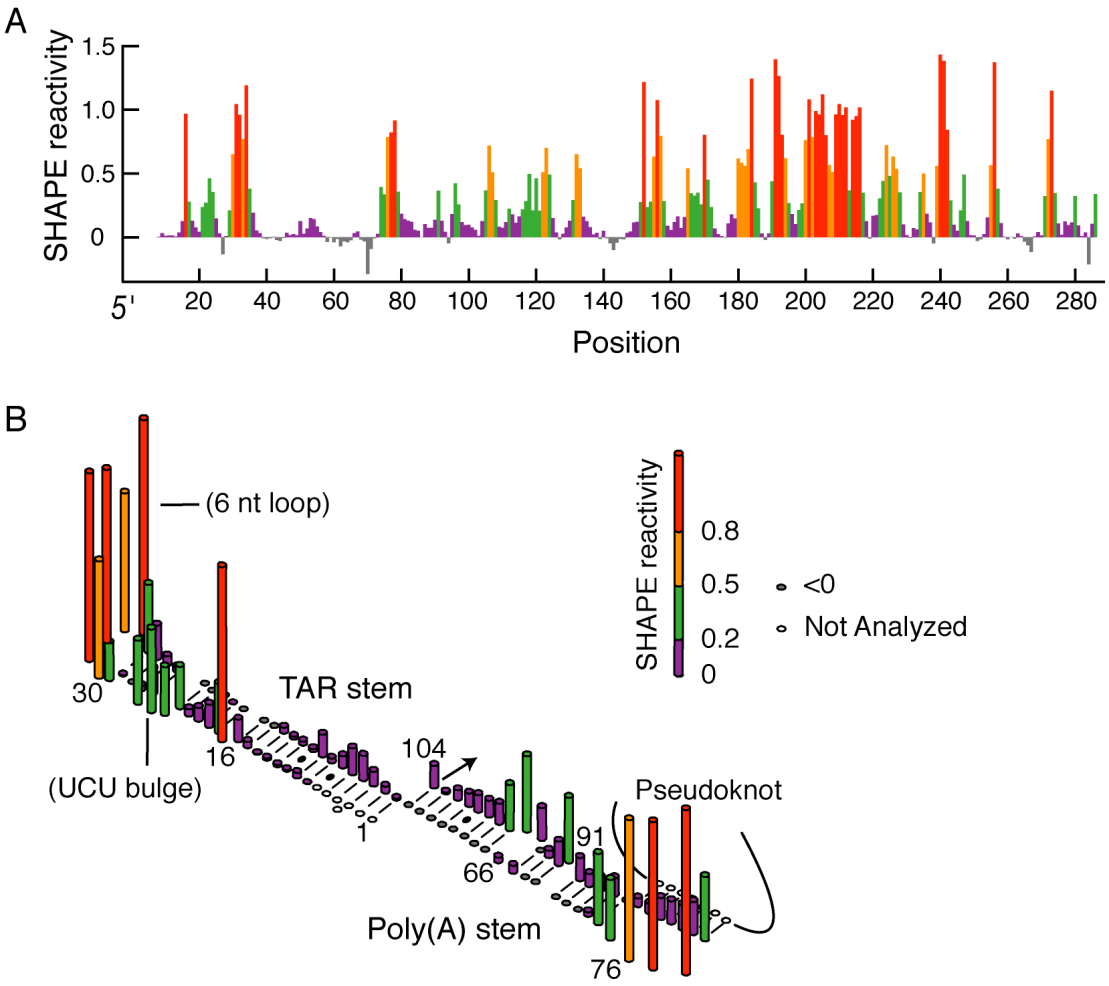


Fig. 4.3: Results of hSHAPE data processing on the HIV-1 genome transcript. (A) Processed SHAPE reactivities as a function of nucleotide position. Highly reactive nucleotides (red and orange bars) report flexible positions in the RNA. (**B**) Absolute NMIA reactivities superimposed on a secondary structure model for the TAR and Poly(A) stem-loops.

nucleotides in the UCU bulge show intermediate reactivities, consistent with NMR studies[9] that indicate that these nucleotides in the TAR stem are partially stacked.

4.4 hSHAPE on long RNAs

A single hSHAPE experiment efficiently interrogates structural constraints for RNA ~300-600 nucleotides long. For longer RNAs it is necessary to combine multiple overlapping reads of the RNA from separate primer sets. To combine structural constraints in a single data set, all experiments are normalized to the same scale. Each SHAPE data set contains a few (~2%) exceptionally reactive positions, which do not represent generically flexible nucleotides. The normalization factor for each data set is determined by first excluding the most reactive 2% of peak intensities and the calculating the average for the next 8% of reactivities. All reactivities are then divided by this average.

This simple normalization procedure generates SHAPE reactivities on a scale from 0 to ~2, where 1.0 is the reactivity of a flexible nucleotide (vertical axis, Figure 4.3A). Nucleotides with reactivities greater than ~0.8 are almost always single stranded (red bars Figure 4.3A), while positions with reactivities less than ~0.2 (black and violet bars Figure 4.3A) are almost always paired. Nucleotides with normalized SHAPE reactivities between 0.2 and 0.8 may be paired or may participate in other partially constraining interactions. The standard deviation at each nucleotide averages approximately 0.1 SHAPE unit, as determined by repeat and overlapping reads on the HIV-1 genomic RNA.[10]

4.5 Development of an RNA structure from hSHAPE constraints

SHAPE reactivities report direct and quantitative information regarding the extent of structure at each nucleotide in an RNA. An important application of SHAPE technology is to develop well-supported structural models for a given RNA. The most successful structure prediction algorithms, such as RNAstructure, use a thermodynamic model based on nearest-neighbor free energy parameters[11,12] to calculate the ΔG for potential structures for a given RNA sequence. The structure with the lowest calculated ΔG becomes the most highly predicted structure. However, the thermodynamic models used by these programs are approximate and RNA structure can be modulated by non-thermodynamic constraints. Therefore, *in silico* methods often predict different structural topologies with nearly identical energies for a given sequence. Without additional structural information, it is not possible to choose which predicted structure reflects the native conformation of an RNA sequence.

We modify the structure prediction program RNAstructure to include hSHAPE constraints in developing structural models.[11,13] We calculate an energetic penalty or credit for pairing each nucleotide according to their SHAPE reactivity. This "quasi-energetic" constraint provides a convenient and straightforward method for including SHAPE based constraints in structure prediction. Quasi-energetic constraints are an approximation of energetic penalties associated with pairing a nucleotide of a specific absolute SHAPE reactivity.

To incorporate the quantitative nature of hSHAPE constraints into structure prediction, we calculate the "quasi-energy" by:

$$\Delta G_{SHAPE} = m \ln[\text{ SHAPE reactivity} + 1.0] + b \qquad (4.2)$$

which is applied to each nucleotide in each stack of two base pairs. Therefore, the quasi-energy is added twice per nucleotide paired in the interior of a helix and once per nucleotide paired at the end of a helix. The intercept, b, is the energy bonus for formation of a base pair with zero or low SHAPE reactivity while m, the slope, drives an increasing penalty for base pairing as the SHAPE reactivity increases. The b and m parameters shown to most likely produce a correct structure[14] were –0.6 and 1.7 in units of kcal/mol, respectively (per nucleotide), but may be varied to modulate the energetic contribution of SHAPE reactivities in structure prediction.

With hSHAPE information only, RNAstructure proposes helices only in regions that exhibit low reactivity. However, there is no constraint on the distance, in nucleotides, between the paired positions. Evidence from known RNA structures suggests that pairings between nucleotides 600 positions apart or more are nearly nonexistent, and 90% of basepairs occur between positions less than 300 nucleotides distant in sequence.[15] Therefore, constraining maximum sequence distance between pairing partners can improve the predictive power of RNAstructure. We have incorporated a tool that completely forbids pairings between positions greater than an arbitrary distance apart in sequence. To develop structural models, we find using a maximum allowed distance between base pairs of 600 provides sufficient constraints for many RNAs. Reducing this value to ~300 may be helpful in locating short, poorly predicted, and transient pairings that could be explained by more probable shorter distance interactions.

Despite the aid of hSHAPE and maximum pairing distance constraints, RNAstructure often predicts multiple different structures with nearly identical energies. Additionally, certain helices in a structure may be more well-determined than other paired regions in a secondary structure. To assess the robustness of a structural prediction, we vary the thermodynamic penalty of pairing associated with hSHAPE reactivities. Predicted base pairs were assigned a "pairing persistence" based on the range of parameters in which they are observed. Helices considered to be highly persistent were observed even when the parameters in Equation 4.2 were set to values as high as $b = 0$ and $m > 4$. Increasing b and m has the effect of increasing the contribution of the SHAPE reactivity information on the secondary structure calculation. Helices with low pairing persistence are observed only at a lower SHAPE-imposed penalties.

Varying the quasi-energetic contribution of SHAPE reactivity information in structure prediction is also useful in supporting a single secondary structure model when several are predicted at a single set of constraints. We assume that predicted helices that exist under the most stringent parameters most likely also exist under less stringent parameters. By following this assumption and by incrementally decreasing the stringency of parameters, a structural model with high pairing persistence can be "built" with the assistance of SHAPE parameters.[10]

Using hSHAPE and maximum pairing distance to constrain RNA secondary structure prediction has a dramatic impact on the quality of predicted structures. For example, prediction accuracy improves from 52% to 90% for the 154 nt RNase P specificity domain[16] and from 38% to 87% for the 1542 nt *Escherichia coli* 16S rRNA.[17] SHAPE-directed predictions characteristically include overall topologies that closely

resemble the correct structure; errors tend to reflect small local structural rearrangements at the ends of helices and at multi-helix junctions.

4.6 Perspective

hSHAPE technology represents a significant improvement to the SHAPE approach. No longer limited by gel electrophoresis, structural reads as long as 600 nucleotides are accomplished in ~ 8 hours. The increased read length of hSHAPE technology decreases the amount of effort necessary to analyze long RNAs. The steps of an hSHAPE experiment may be completed in parallel, making it theoretically possible to complete dozens of analyses in a single day.

Additionally, we have here developed a set of steps to propose accurate, well-defined RNA secondary structures from raw sequencer data. Presently, several of these steps are incorporated into computer algorithms, but a significant amount of user input (~8h) is required to effect a single analysis. Continued work in developing software may make analysis of a SHAPE read possible in an hour or less.

Several RNA molecules of interest are thousands of nucleotides long, including some mRNAs, as well as viral genomes. hSHAPE makes analyzing the structure of and proposing structural models for these RNAs experimentally tractable. As an extreme example of RNA length, the SARS coronavirus RNA genome is 29,751 bases long. Assuming a readlength of 600 nucleotides and an overlap of 200 nucleotides at either end of the read, the entire SARS coronavirus may be interrogated, in duplicate, in less than 200 reads.

4.7 Experimental Section

4.7.1 HIV-1 RNA transcripts. A DNA template encoding the 5' 976 nucleotides of the HIV-1 genome and containing a promoter for T7 RNA polymerase was generated by PCR [2 mL; 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2.5 mM $MgCl_2$, 0.5 μM forward (5'-TAATA CGACT CACTA TAGGT CTCTC TGGTT AGACC) and reverse (5'-CTATC CCATT CTGCA GCTTC C) primers, ~1 μg plasmid template containing a partial sequence of the HIV-1 pNL4-3 molecular clone (Genbank AF324493, obtained from the NIH AIDS Research and Reference Reagent Program), 200 μM each dNTP, and 25 units Taq polymerase; 34 cycles]. The PCR product was recovered by ethanol precipitation and resuspended in 300 μL TE (10 mM Tris-HCl, pH 8, 1 mM EDTA). Transcription reactions [3 mL; 37 °C; 5 h; 40 mM Tris-HCl (pH 8.0), 5 mM $MgCl_2$, 10 mM DTT, 4 mM spermidine, 0.01% Triton X-100, 4% (w/v) PEG 8000, 300 μL PCR product, and 2 mM each NTP] were initiated by adding 0.1 mg/mL T7 RNA polymerase.[18] The RNA product was precipitated and purified by denaturing polyacrylamide gel electrophoresis (5% acrylamide, 7 M urea), excised from the gel, and recovered by electroelution. The purified RNA (0.6 nmol) was resuspended in 100 μL TE.

4.7.2 Modification of transcript RNA. RNA (2 pmol) in 14.4 μL of 1/2 × TE was refolded by heating at 95 °C, placing on ice, adding 3.6 μL folding buffer [250 mM Hepes-NaOH (pH 8), 1 M potassium acetate, pH 8, 25 mM $MgCl_2$], and incubating at 37 °C for 60 min. The folded RNA was divided equally between two tubes and treated with either NMIA[5,19] (1 μL, 32 mM in DMSO) or neat DMSO (1 μL) and allowed to react for

60 min at 37 °C. RNA from the (+) and (–) NMIA reagent experiments was recovered by ethanol precipitation and resuspended in 10 μL TE.

4.7.3 Detection of 2'-*O*-Adducts by Primer Extension. RNA (1 pmol, 10 μL, in 1× TE) corresponding to either the (+) or (–) NMIA reactions was heated to 95 °C for 3 min and cooled on ice for 1 min. Fluorescently labeled primer (3 μL, complimentary to positions 342-363) was added to the (+) (0.2 μM Cy5) and (–) (0.4 μM WellRED D3) NMIA reactions, respectively, and primer-template solutions were incubated at 65 °C for 5 min and 35 °C for 10 min. Primer extension was initiated by addition of enzyme mix [6 μL; 250 mM KCl; 167 mM Tris-HCl (pH 8.3); 1.67 mM each dATP, dCTP, dITP, dTTP; 10 mM $MgCl_2$; 52 °C,1 min] and Superscript III (1 μL, 200 units, Invitrogen). Extension continued at 52 °C for 15 min. Sequencing reactions used to identify peaks in the (+) and (–) reagents experiments were obtained using transcript RNA (1 pmol, in 9 μL TE), 3 μL primer (2 μM WellRED D2 or 1.2 μM LICOR IR 800), enzyme mix (6 μL), 1 μL of ddNTP solution (0.25 mM ddGTP or 10 mM other nucleotides), and Superscript III (1 μL). Depending on the quality of synthesis, primers were purified by denaturing gel electrophoresis [20% polyacrylamide, 1× TBE, 7 M urea; dimensions 0.75cm × 28.5 cm (w) × 23 cm (h); 32W; 90 min] and passively eluted into 1/2× TE overnight. The four reactions corresponding to a complete hSHAPE analysis [(+) NMIA, (–) NMIA, and two sequencing reactions] were combined, precipitated with ethanol in the presence of acetate, EDTA, and glycogen. Pellets were washed twice with 70% ethanol, dried under vacuum, and resuspended in deionized formamide. cDNA samples in 40 μL formamide were then separated on a 33 cm × 75 μm capillary using a Beckman CEQ 2000XL DNA sequencer.

4.8 References

1. Das, R., Laederach, A., Pearlman, S., Herschlag, D. & Altman, R. B. SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA* **11**, 344-54 (2005).

2. Giddings, M. C., Severin, J., Westphall, M., Wu, J. & Smith, L. M. A Software System for Data Analysis in Automated DNA Sequencing. *Genome Res.* **8**, 644-665 (1998).

3. Guex, N., Vasa, S. M., Wilkinson, K. A., Weeks, K. M. & Giddings, M. C. SHAPEfinder. *In preparation* (2007).

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

5. Merino, E. J., Wilkinson, K. A., Coughlan, J. L., & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127**, 4223-31 (2005).

6. Paillart, J. C., Dettenhofer, M., Yu, X.-F., Ehresmann, C., Ehresmann, B. & Marquet, R., First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J. Biol. Chem.* **279**, 48397-48403 (2004).

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

8. Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5' region of the HIV-1 Genome. *J. Mol. Biol.* **336**, 369-79 (2004).

9. Puglisi, J. D., Tan, R., Calnan, B. J., Frankel, A. D. & Williamson, J. R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* **257**, 76-80 (1992).

10. Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C. & Weeks, K. M. Structures of the HIV-1 Genome. *In preparation* (2007).

11. Mathews, D. H. & Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **16**, 270 (2006).

12. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 71 (2004).
13. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. & Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101**, 7287-7292 (2004).

14. Li, T. W., Mathews, D. H. & Weeks, K. M. *in preparation* (2007).

15. Mathews, D. H. *unpublished data* (2007).

16. Mortimer, S. M. & Weeks, K. M. *J. Am. Chem. Soc.* A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry. **129**, 4144-5 (2007).

17. Li, T. & Weeks, K. M. *J. Am. Chem. Soc.* **in preparation** (2007).

18. Milligan, J. F. & Uhlenbeck, O. C. Synthesis of small RNAs using T7 RNA polymerase. *Meth. Enzymol.* **180**, 51-62 (1989).

19. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J. Am. Chem. Soc.* **127**, 4659-67 (2005).

CHAPTER 5

STRUCTURES OF THE HIV-1 GENOME

5.1 Introduction

The HIV-1 RNA genome participates in multiple, pivotal, stages of the viral life cycle. It serves as mRNA for the synthesis of several viral proteins, is the substrate for a wide variety of splicing interactions, forms intermolecular dimer interactions that direct packaging and enable recombination between two RNA strands, base pairs with the $tRNA^{Lys3}$ molecule that primes proviral DNA synthesis, acts as the template for viral DNA synthesis, and binds essential regulatory and cofactor proteins.[1,2] The HIV genome represents a compelling target for antiviral therapies because it is simultaneously both the largest component of the virus and conserved interactions with proteins and other RNAs are critical for infectivity. However, our current understanding of HIV genomic RNA structure, and of the structures of virtually all long viral and cellular RNAs, has been largely limited to highly focused analyses of specific, short RNA domains and small genomic fragments. I employ, in collaboration with Robert Gorelick, Suzy Vasa, Nicolas Guex, Alan Rein, David Mathews, and Morgan Giddings, the methods described in Chapter 4 to analyze the structure of an intact HIV-1 genome in infectious virions. To characterize fully regulatory and protein-binding motifs, we also analyze the structure of genomic RNAs in which conserved RNA-protein interactions are disrupted *in situ*, and of purified viral RNAs free of other viral components.[3]

5.2 Results

5.2.1 Experimental approach. To analyze local nucleotide flexibility for long RNAs, such as the HIV-1 genome, we combined the results of multiple hSHAPE experiments performed on the same long RNA using primers that anneal 200-300 nts apart. Individual SHAPE reads were analyzed as described in Chapter 4. Combining normalized SHAPE reactivities from different reads is further supported in that overlapping regions are almost identical to each other (compare overlay of closed and open columns, Figure 5.1A).

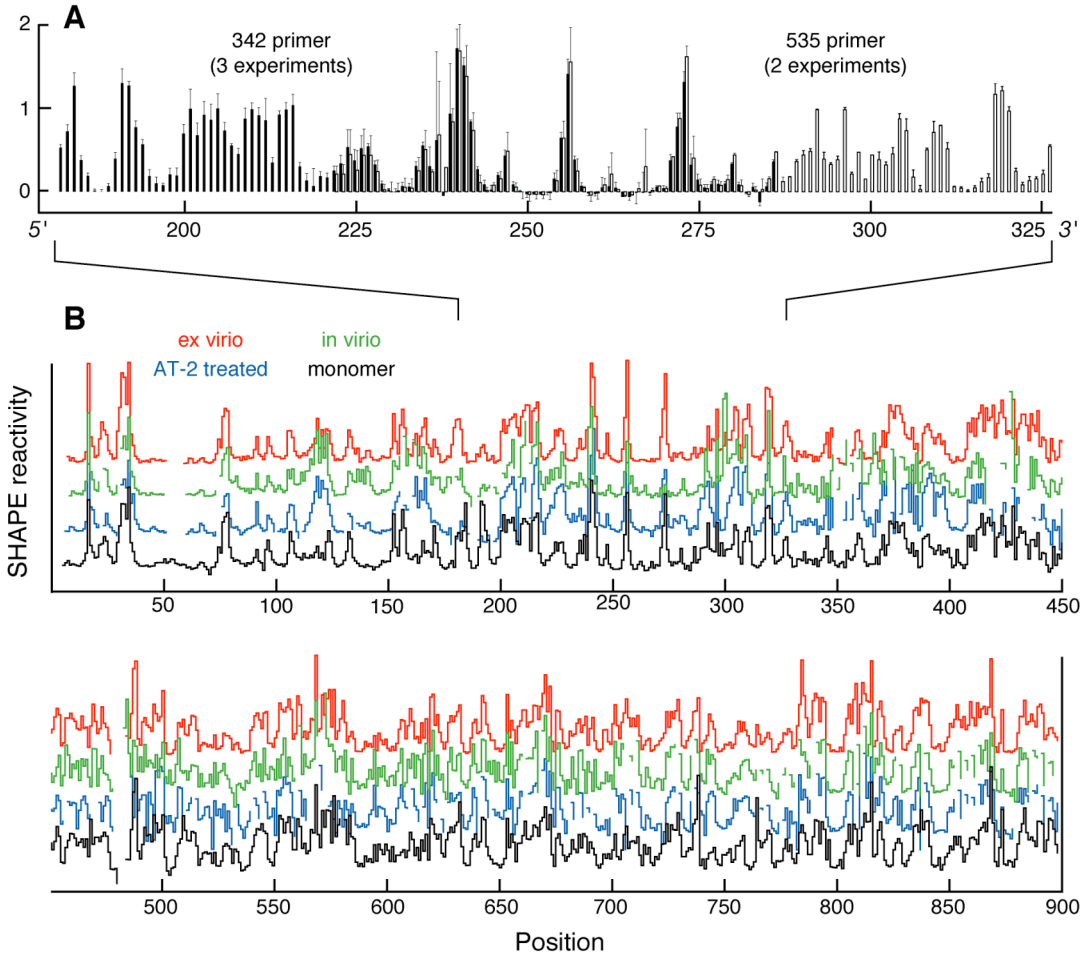


Fig. 5.1. hSHAPE is quantitative and highly reproducible. (A) SHAPE reactivities for two overlapping extension reactions using primers that anneal ~200 nucleotides apart. Solid and open bars indicate normalized SHAPE reactivities and standard deviations for primers that anneal at nts 342-362 and 535-555, respectively. (B) Complete SHAPE data for HIV-1 genomic RNA analyzed in virions and in the *ex virio*, AT-2 treated and monomer states.

5.2.2 Developing a structural model. Our goal is to develop structural models for the HIV-1 RNA genome as it exists inside infectious viral particles. To detect virion-specific RNA conformational changes and RNA-protein interactions, we used hSHAPE to analyze the structures of four states in total. In addition to (i) genomic RNA inside infectious virions (the *in virio* state), we analyzed (ii) authentic HIV-1 RNA gently deproteinized and extracted from virions (*ex virio*), (iii) genomic RNA in which select RNA-protein interactions were disrupted by treatment with Aldrithiol-2 (AT-2 treated, described in detail below), and (iv) a 976-nucleotide HIV-1 monomer generated by *in vitro* transcription (termed the monomer state). We obtained structural information for 94% of all nucleotides in these four states, in duplicate or better, for a total analysis of over 8,200 nts (Figure 5.1B). hSHAPE reactivities were then incorporated into structure prediction algorithms to propose secondary structures as described in Chapter 4.

We use the protein-free *ex virio* RNA as our reference state for the secondary structure of the 5' end of the HIV-1 genome (Figure 5.2). This structure strongly reflects the constraints imposed by SHAPE reactivities. We could therefore assess the well-determinedness of each helix in the secondary structure by varying the thermodynamic penalty imposed by the SHAPE constraints, which we term the pairing persistence. The most persistent helices (black and purple bars, Figure 5.2) were predicted even when SHAPE constraints were used to impose large pairing penalties for even slightly reactive nucleotides. Less persistent helices (blue and green bars, Figure 5.2) form only at a lower SHAPE-imposed pairing persistence.

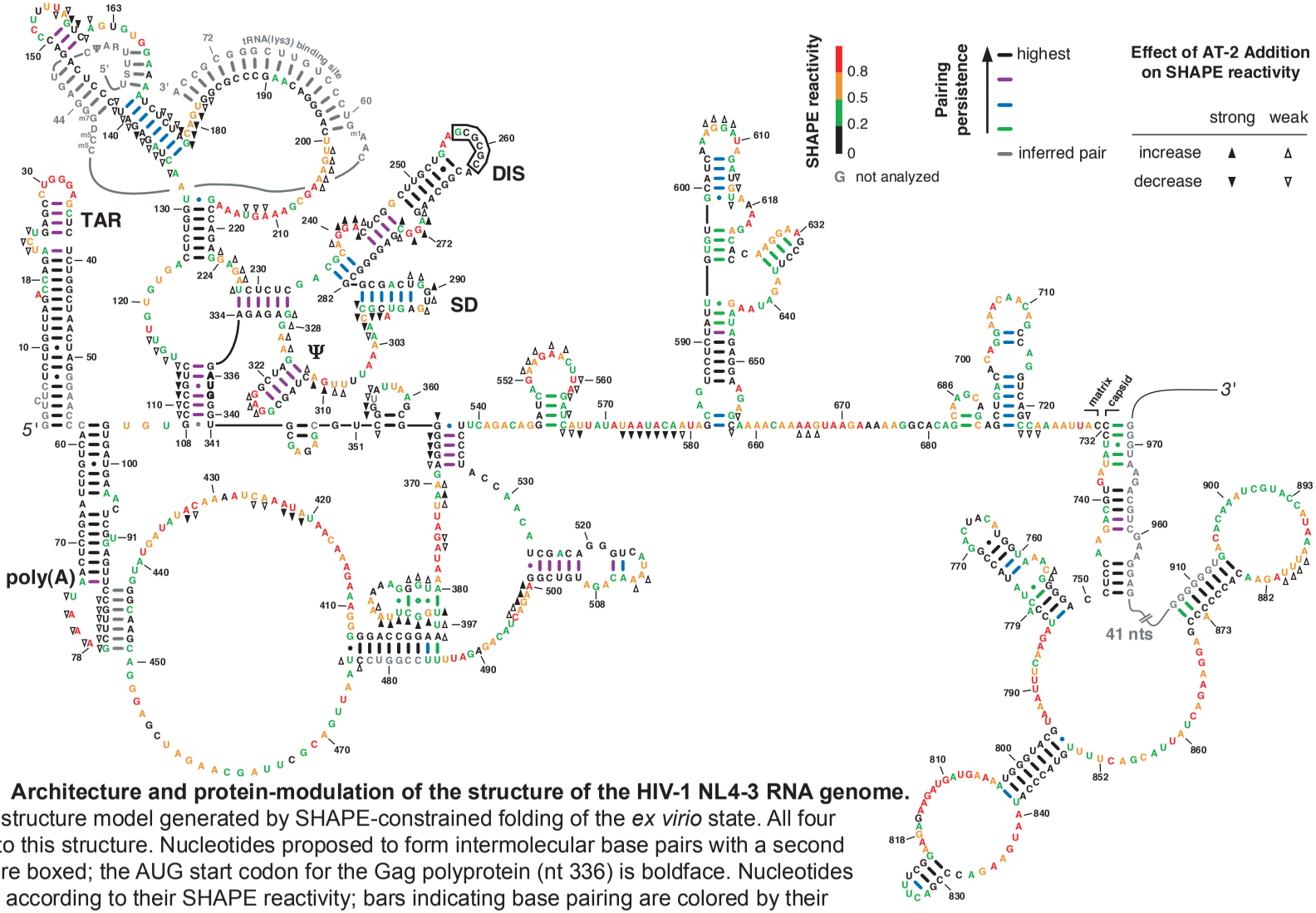


Figure 5.2. Architecture and protein-modulation of the structure of the HIV-1 NL4-3 RNA genome. Secondary structure model generated by SHAPE-constrained folding of the *ex virio* state. All four states fold to this structure. Nucleotides proposed to form intermolecular base pairs with a second monomer are boxed; the AUG start codon for the Gag polyprotein (nt 336) is boldface. Nucleotides are colored according to their SHAPE reactivity; bars indicating base pairing are colored by their hSHAPE prediction pairing persistence. Effects of pre-treatment of viral particles with AT-2 are indicated with closed and open arrowheads.Sequence boundary for the matrix and capsid components of Gag is shown with brackets.

5.3 Discussion

5.3.1 hSHAPE and previous mapping studies. hSHAPE structural constraints are consistent with previous qualitative structural mapping studies using chemical and enzyme reagents that analyzed a subset of the nucleotides analyzable by SHAPE. Given this similarity, elements of the SHAPE-constrained secondary structure (Figure 5.2) are similar to previously proposed models.[4-6] For example, there is a strong consensus regarding the structures of several stem-loop motifs including the TAR, Poly(A), DIS, SD, and Ψ elements (labels, Figure 5.2). SHAPE analysis also supports formation of a previously proposed long-range pseudoknot (nts 79-85/443-449).[7]

Our secondary structure model also contains many, substantive, differences with respect to previous models. Structural differences with respect to prior models reflect several contributions. First, the hSHAPE data set is 94% complete. Completeness is important because RNA secondary structure prediction using thermodynamic parameters alone typically yields structures with significant errors, especially as RNA length increases.[8-10] In the case of HIV-1 genomic RNA, relatively little data had been obtained for positions 110-125, 236-243, 276-282, 408-415, 432-435, and 465-477, which has led to structural proposals that are not consistent with the more complete hSHAPE data set. Second, end effects can dramatically alter structure prediction when only small pieces of a large RNA are analyzed. Structures that involve or lie inside of long-range interactions, such as the 108-114/335-341 stem must be mispredicted if the RNA sequence does not include the complete domain. Third, incorporation of SHAPE reactivity information as a pseudo-energy term makes the structure prediction calculation insensitive to errors in any single reactivity measurement. Finally, we identify a structural domain that lies 3' of position 732 (Figure

5.2), which has not previously been analyzed.

5.3.3 Structural differences in regulatory versus coding regions. The 5' end of the HIV genome spans two functional regions whose boundary lies at the AUG start codon for the Gag coding sequence (nts 336-338; in bold, Figure 5.2). Positions upstream of the AUG codon comprise a 343 nt long 5' regulatory domain; whereas, nucleotides 3' of the start codon span the Gag coding region, of which we have mapped >560 nts. It is currently not possible to distinguish coding versus non-coding regions by secondary structure prediction alone.[11,12] By two criteria, SHAPE reactivities indicate that the 5' regulatory domain is more highly structured than the 3' mRNA-like region. First, the average SHAPE reactivity, a metric for the extent of structure in the two regions, is 0.30 for the 5' regulatory domain and 0.44 for the 3' mRNA-like region. (dashed lines, Figure 5.3). Notably, the inflection point occurs almost exactly at the AUG start codon (gray bar, Figure 5.3) Second, in our secondary structure model (Figure 5.2), nucleotides in the 5' regulatory domain are 1.7 times more likely to be paired than those in the 3' coding region. Although the 3' coding region is relatively unstructured overall, several structured regions with high pairing persistence punctuate this region. The most significant region spans positions 732-972. Strikingly, this element occurs at exactly the boundary between the matrix and capsid domains of the Gag polypeptide. We speculate that RNA structure at this site modulates translation of the Gag polyprotein to facilitate independent folding of the matrix and capsid domains. hSHAPE may be broadly useful for identifying novel regulatory

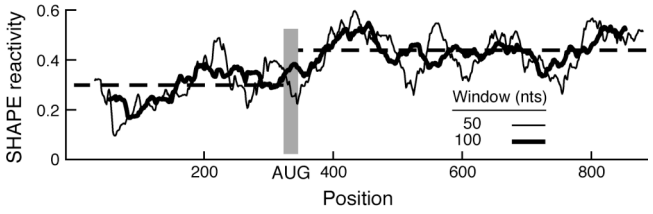


Fig. 5.3. Smoothed SHAPE reactivities (solid lines) and average reactivity (dashed lines) as a function of position 5' or 3' to the AUG Gag start codon.

motifs in cellular RNAs.

5.3.4 Structures for distinct HIV genome states. Comparison of the complete SHAPE reactivity profiles for the *ex virio* reference state with the other three states – *in virio*, AT-2 treated, and monomer – reveals that these distinct states contain extensive regions with identical structures. This is a remarkable result considering, for example, that the *in virio* RNA was maintained in its native conformation inside virions throughout the chemical modification step; whereas, the monomer state was refolded after heating to 90 °C. Thus, the functions of the HIV-1 genome appear to be governed by a single predominant conformation.

In addition, analysis of the *in virio* state and comparison with the protein-free *ex virio* RNA reveal numerous regions that are persistently accessible to SHAPE chemistry. These regions are expected to hybridize readily with complementary sequences, including antisense and RNAi-based oligomers, and represent multiple new and attractive targets for anti-HIV therapeutics (red and orange positions, Figure 5.2).

Reactivity profiles for these four states also show important structural differences, which can largely be interpreted in terms of local RNA conformation and protein-binding effects. There are three regions with significant differences between the *ex virio* reference state and the monomer RNA, which was refolded *in vitro*. The most dramatic difference is that the monomer RNA is much more reactive at positions 182 to 199 (see difference plot, Figure 5.4A). This region maps exactly to the $tRNA^{Lys3}$ primer binding site[2] and indicates that the primer remains paired to the HIV-1 RNA genome in viral particles. The *ex virio* state also has higher SHAPE reactivity at positions 161-166 and lower reactivity at positions 168-170, as compared with the monomer state. These reactivity changes are consistent with $tRNA^{Lys3}$-induced structural rearrangement at nucleotides 141-170 due to multi-site

interactions between the tRNA and genomic RNA (Figure 5.2, gray nucleotides), a subset of which have been identified previously.[13] The monomer state, which is not bound by tRNA, folds into a different local structure in these regions (compare Figures 5.2 and 5.4B).

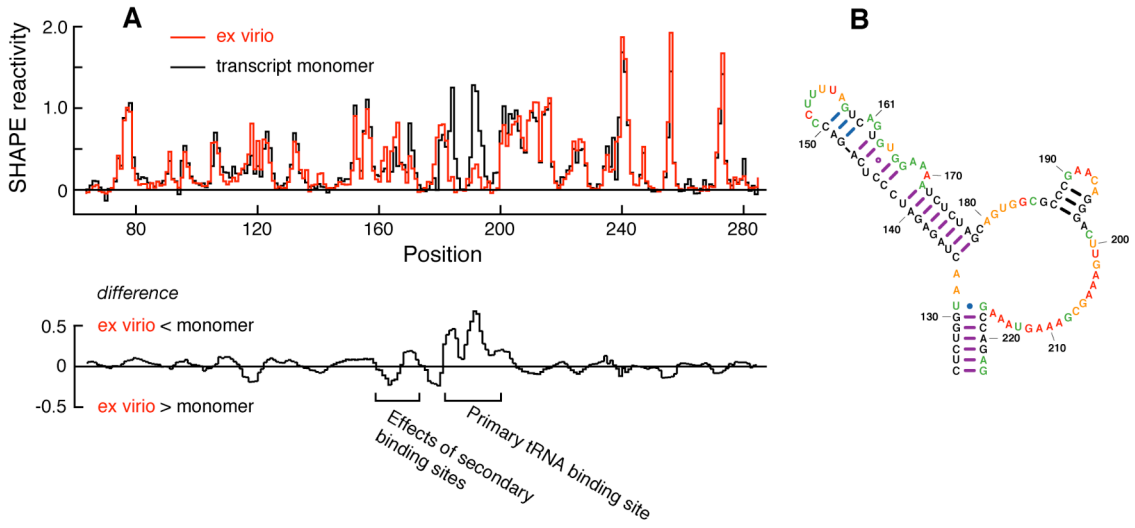


Fig. 5.4. RNA conformational changes that differentiate *ex virio* and transcript monomer states. (**A**) SHAPE reactivity as a function of nucleotide position for *ex virio* (red) and transcript (black) RNAs. Difference plot in lower panel has been smoothed over a five nucleotide window. (**B**) Secondary structure model for the primer binding site in the absence of $tRNA^{Lys3}$.

In all normal retroviral particles, the genomic RNA is in a dimeric form, with similar or identical RNA strands linked together by a limited number of base pairs and tertiary interactions. The dimeric structure is a critical element in the selective encapsidation of the genome[14,15] and template-switching between two RNA monomers during reverse transcription leads to recombination,[16,17] a major source of genetic variation for retroviruses. Dimerization is generally thought to involve an initial loop-loop interaction[18] at the self-complementary sequence $G^{257}CGCGC^{262}$. These nucleotides are unreactive in both the monomer and *ex virio* states (boxed nucleotides, Figure 5.2), which supports formation of constraining base pairing interactions at this loop. Thus, even the monomer state forms a loop-loop dimer. A similar 'early' loop-loop dimer state has been identified for the Moloney

murine sarcoma retrovirus.[19,20] We observed no reactivity differences greater than 0.1 SHAPE units between the monomer and *ex virio* RNAs in sequences flanking the 257-262 loop (compare red and black traces in Figure 5.1B and 5.4A). This result was surprising because current models postulate[18] that the stem sequences adjacent to this loop form a stable intermolecular duplex involving both genomic RNA strands. Similar SHAPE reactivities in this region do not support formation of an intermolecular duplex in mature HIV-1 viral particles, although we cannot exclude a change yielding identical local nucleotide flexibilities in the pre- and post-dimer RNAs.

5.3.5 Direct analysis of NCp7-RNA genome interactions. NMIA is a small, mildly hydrophobic reagent, which we found readily crosses the retroviral membrane. We analyzed the structure of HIV-1 genomic RNA inside infectious virions by treating viral particles with NMIA and then extracting and processing the modified RNA (the *in virio* state; Figure 5.5B). We observe numerous reproducible differences between the *ex virio* and *in virio* states (Figures 5.1B,5.5B) that must report virion-specific RNA structures and RNA-protein interactions.

The most prominent protein ligand for genomic RNA in mature HIV virions is the nucleocapsid protein.[21,22] The nucleocapsid protein (NCp7) contains two highly conserved "zinc-knuckle" motifs comprised of cysteine and histidine residues that coordinate zinc ions and bind preferentially to guanosine (Figure 5.5A). These compact motifs are flanked by positively charged residues that interact at adjacent RNA elements.[23,24] "Zinc ejecting" agents like 2,2'-dithioldipyridine (or Aldrithiol-2, AT-2; Figure 5.5C) quantitatively disrupt interactions between the zinc ion and its cysteine ligands, and thus compromise NCp7-RNA interactions, but leave the surface of the virus particle intact.[25] We therefore disrupted

nucleocapsid-RNA interactions *in situ* by treating virions with AT-2 and then analyzed the structure of the resulting genomic RNA using hSHAPE (Figures 5.5 and 5.6).

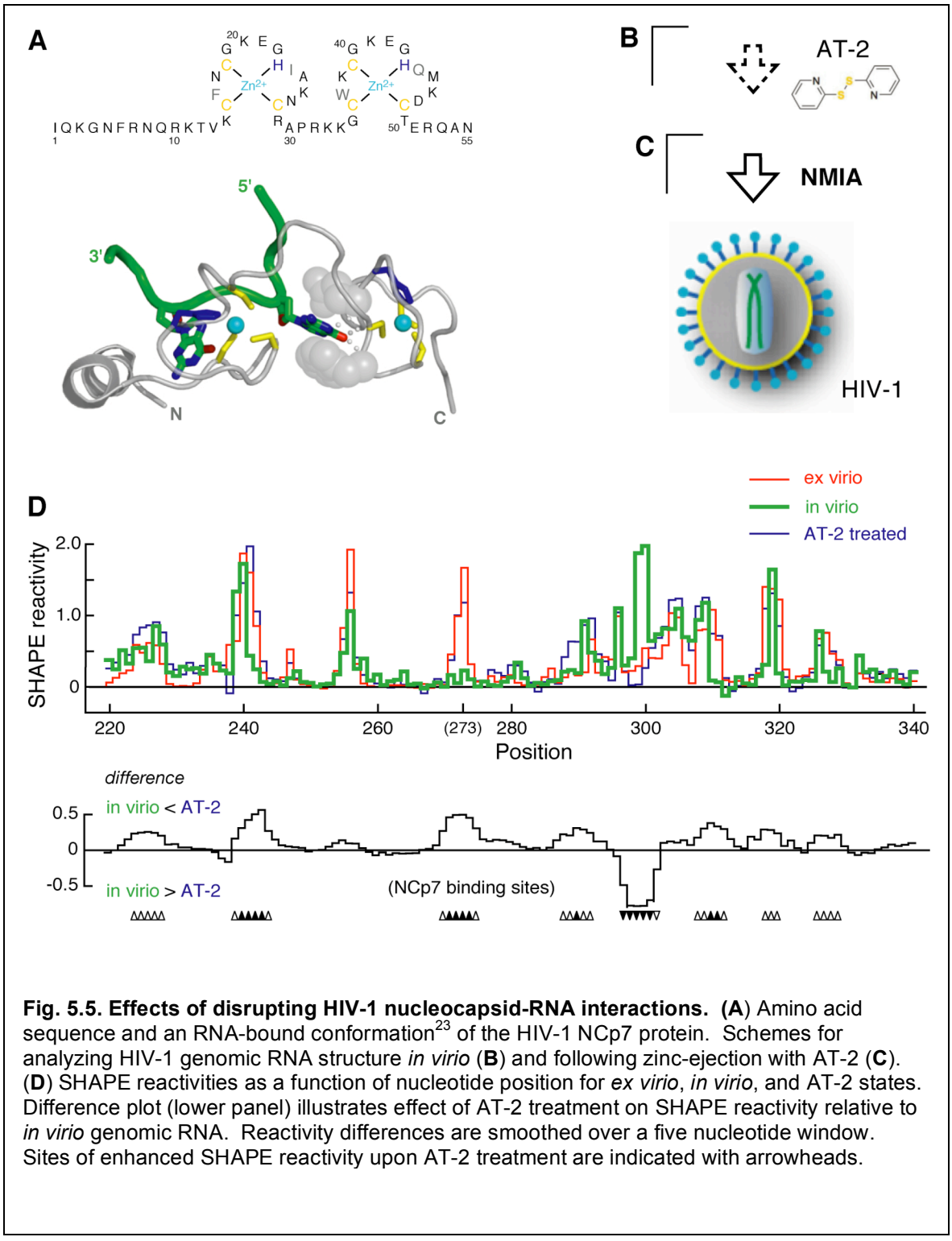


Fig. 5.5. Effects of disrupting HIV-1 nucleocapsid-RNA interactions. (**A**) Amino acid sequence and an RNA-bound conformation[23] of the HIV-1 NCp7 protein. Schemes for analyzing HIV-1 genomic RNA structure *in virio* (**B**) and following zinc-ejection with AT-2 (**C**). (**D**) SHAPE reactivities as a function of nucleotide position for *ex virio*, *in virio*, and AT-2 states. Difference plot (lower panel) illustrates effect of AT-2 treatment on SHAPE reactivity relative to *in virio* genomic RNA. Reactivity differences are smoothed over a five nucleotide window. Sites of enhanced SHAPE reactivity upon AT-2 treatment are indicated with arrowheads.

The effect of AT-2 treatment is highly specific because large regions of the genomic RNA in the intact *in virio* and AT-2 treated states show identical SHAPE reactivities. In contrast, disrupting NCp7-RNA interactions by "zinc-ejection" both increases and decreases local nucleotide flexibility in other distinct genome regions (Figures 5.5D and 5.6). Strikingly, the strongest and most systematic effects of AT-2 treatment lie in the 5' regulatory domain and are largely absent after position 580 in the 3' coding region (upward and downward pointing arrowheads, respectively; Figure 5.2).

Regions showing a strong increase in SHAPE reactivity in the AT-2 treated state almost always resemble the protein-free *ex virio* state (compare blue and red traces, Figures 5.5D and 5.6). We infer that a strong increase in reactivity in the AT-2 treated samples at these sites reflects disruption of specific NCp7-RNA interactions. The single strongest NCp7 binding site lies at positions 272-274, followed closely by positions 241-244 (upwards pointing arrowheads in Figures 5.2 and 5.5D). These sites, which have not been previously implicated in NCp7 or Gag recognition, are consistent with primary interaction motifs for the viral nucleocapsid domain at the 5' end of the HIV-1 genome.

5.3.6 Definition of a nucleocapsid interaction domain. Inspection of the strongest NCp7 binding sites (upward facing arrowheads, positions 241-244, 272-274, 288-292, and 308-312, Figure 5.2), plus several secondary sites (positions 224-227, 318-320 and 326-329) indicate that the consensus NCp7 RNA recognition motif spans 1-2 guanosine residues in a single-stranded region of ~4 nts adjacent to a helix. Most such sites lie in a single domain in our model for the HIV-1 genome (positions 224-334, Figure 5.2). This domain overlaps structures that play a major role in HIV-1 genomic RNA packaging[21] and also includes the $G^{257}CGCGC^{262}$ that likely forms intermolecular base pairs in the genomic RNA dimer (Figure

5.2). We propose that the 223-334 domain dimer interacts, potentially cooperatively, with multiple copies of the HIV-1 NCp7 protein and with the nucleocapsid motif in the Gag protein. The specific juxtaposition of high affinity NCp7/Gag binding sites in the dimer would then function as the structural motif that is specifically packaged in nascent HIV virions.

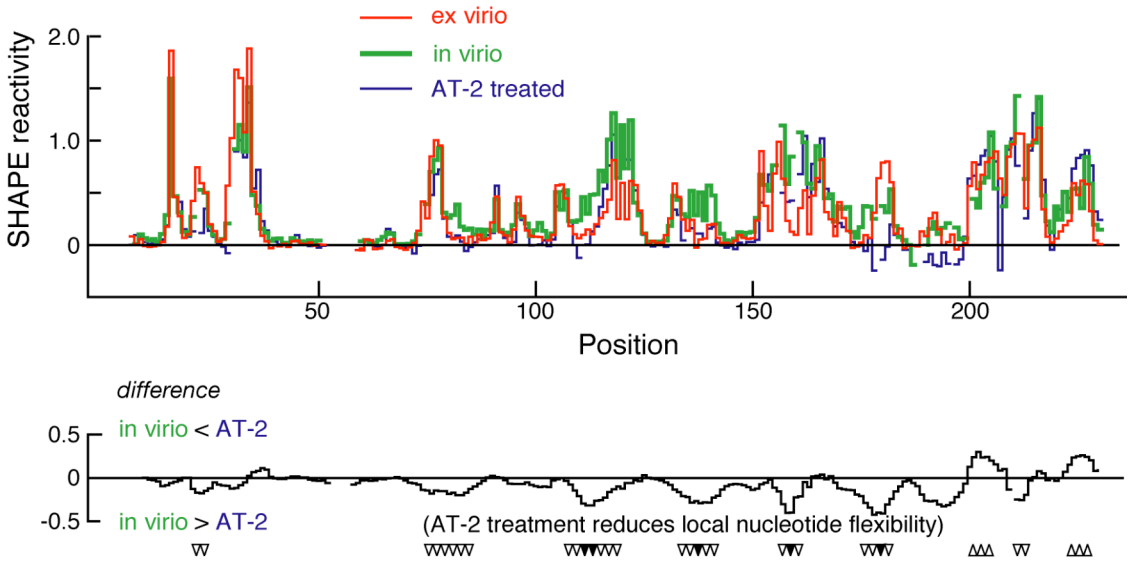


Fig. 5.6. NCp7 increases SHAPE reactivity and, by inference, RNA flexibility in regions 5' to the primer binding site. Absolute SHAPE reactivities are shown for the *ex virio* (red), *in virio* (green) and AT-2 treated (blue) states. Difference plot reports the effect of AT-2 treatment on the *in virio* state; points have been smoothed over a 5 nt window.

5.3.7 Structure destabilizing activity of the nucleocapsid domain. Whereas the nucleocapsid domain binds specifically to retroviral RNA in its packaging function, the protein also has an almost opposing function, that of binding non-specifically to nucleic acids and facilitating structural rearrangements.[22] hSHAPE analysis detects this alternate activity directly. The presence of intact NCp7, prior to AT-2 treatment, increases SHAPE reactivity and, by inference, flexibility in two regions of the genomic RNA (downward facing arrows,

Figures 5.2 and 5.6). Local nucleotide flexibility is enhanced at 5 sites 5' of the tRNA primer binding site, which likely functions to facilitate initial extension of the tRNA primer during the earliest stages of retroviral cDNA synthesis. Flexibility is also increased at ~9 sites 3' of the Gag start codon and might function to enhance either cDNA synthesis or translation by reducing RNA structure in this region.

5.4 Perspective

Using a concise set of experiments, we have obtained single nucleotide resolution structural information for 94% of the first 900 nucleotides of the HIV-1 genomic RNA inside infectious virions. Because SHAPE reactivities are quantitative and highly reproducible, we could interpret structural differences between intact genomic RNA in authentic particles with three other instructive states, representing a total analysis of over 8,200 nucleotides. These comparisons support multiple new and revised models for the intimate role of RNA genome structure in retroviral replication and infectivity. High-throughput RNA structure analysis will make possible analysis of the complete and intact RNAs that constitute a viral or cellular transcriptome, as a function of multiple biological states. We anticipate that hSHAPE will facilitate analysis of the underlying contributions of RNA structure to translational regulation, alternative splicing, higher-order packaging in compact viruses, and many other RNA-based processes.

5.5 Experimental Section

5.5.1 HIV-1 particle production. VSV-G pseudotyped HIV-1 NL4-3 viral particles were produced by cotransfecting the pNL4-3 (Genbank AF324493) and pHCMV-G (VSV-G protein expression construct)[26] plasmids at a ratio of 3:1 into 293T cells as described[27] except that TransIT293 (Mirus Bio) was used to increase transfection efficiency. In sum, 40 × 150 cm^2 culture flasks, seeded at a density of 3×10^6 293T cells were transfected. Cultures were incubated for 48 hours and supernatants harvested, clarified by centrifugation at 5000 *g* for 10 min, filtered through a 0.2 μm membrane, and stored at 4 °C overnight. Cultures were incubated for an additional 24 hours with fresh culture media, and virus-containing supernatant was again collected using the same procedure. Supernatants from both harvests were pooled at 4 °C in preparation for treatment with the AT-2 and NMIA reagents. Viral particle genomes were quantified by real-time RT-PCR;[26] the yield is typically 40 pmol HIV-1 RNA genomes/L cell culture.

5.5.2 HIV-1 particle treatment with AT-2. Aldrithiol-2 (AT-2, systematic name 2,2'-dithiodipyridine; 0.5 M in DMSO, 2.0 mL) or DMSO (2.0 mL) was added to 1.0 L virus supernatant and incubated overnight at 4 °C. Virus particles from the (+) and (–) AT-2 experiments were pelleted separately by centrifugation (113,000 g_{max}, 4 °C, 1.5 h) through a 20% (w/v) sucrose cushion in phosphate buffered saline. Pellets were resuspended in 1.0 mL NMIA reaction buffer [50 mM Hepes-NaOH (pH 8), 200 mM NaCl, 0.1 mM EDTA, and 10% fetal bovine serum].

5.5.3 NMIA modification of viral particles. Concentrated samples of either purified viral particles or particles treated with AT-2 (500 μL) in NMIA reaction buffer were treated with NMIA (50 μL, 100 mM) or neat DMSO (50 μL) for 50 min at 37 °C. The virus particle

production, AT-2 treatment, and NMIA modification steps were always performed as a single continuous process and without intermediate storage steps.

5.5.4 Extraction of HIV-1 Genomes from NMIA-modified Particles. RNA genomes subjected to reaction with NMIA *in virio* were gently extracted from viral particles as described.[15] In sum, concentrated samples of virus particles (in 550 μL NMIA buffer) were incubated at ~22 °C with 5 μL Proteinase K (20 mg/mL), 33.5 μL 1 M Tris-HCl (pH 7.5), 13.4 μL 5 M NaCl, 1.34 μL 0.5 M EDTA, 6.7 μL 1 M DTT, and 4 μL glycogen (20 mg/mL) for 30 min. RNA was purified by three consecutive extractions with phenol-chloroform, followed by precipitation with ethanol. Samples were resuspended in 1/2× TE to a concentration of 0.5 μM, based on quantitative RT-PCR analysis.

5.5.5 Extraction and SHAPE analysis of HIV-1 Genomes from native particles. For the *ex virio* state, pelleted viral particles were dissolved in 1 mL of 50 mM Hepes-NaOH (pH 8.0), 0.5 mM EDTA, 200 mM NaCl, 1% (w/v) SDS, and 100 μg/mL proteinase K and digested for 30 min at ~22 °C. The RNA was then extracted against phenol-chloroform and the resulting deproteinized genomes were then aliquoted (2 pmol) and flash frozen at –80 °C. For SHAPE analysis, the *ex virio* RNA was treated with NMIA using the same procedure as for modification of the monomer state (described above), except that the initial 90 °C heat step was omitted, and the time for incubation in folding buffer was reduced to 10 min.

5.5.6 Detection of 2'-*O*-Adducts by Primer Extension. *In vitro* transcript (prepared and modified as described in Chapter 4) or authentic genomic RNA corresponding to either the (+) or (–) NMIA reactions served as a template for primer extensions as described in Chapter 4. Sequencing experiments were also performed as described. Four sets of primers were used that were complementary to positions 342-363, 535-555, 743-762, or 956-976.

5.5.7 Data Processing. Data processing was performed as described in Chapter 4. In sum, raw fluorescence intensity versus elution time profiles were analyzed using the signal processing framework in BaseFinder[28] to (i) correct baseline, (ii) correct color separation for spectral overlap of the fluorescent dyes such that each channel reported quantitative cDNA amounts, and (iii) correct mobility to align corresponding peaks in the four channels. Areas under each peak in the (+) and (–) NMIA traces were obtained by (i) peak detection and interpolation to align peaks in each channel with the RNA sequence and (ii) performing a whole trace Gaussian-fit integration.[29] Signal decay correction and normalization were completed manually as described (Chapter 4).

5.5.8 Incorporation of hSHAPE constraints into RNAstructure. hSHAPE constraints were incorporated into RNAstructure-based prediction as a quasi-energetic constraing (Chapter 4). For the HIV-1 RNA, the b and m parameters were –0.6 and 1.7 kcal/mol, respectively (per nucleotide) and the maximum allowed distance between base pairs was restrained to be 300 nts or less. Pairing persistence of individual helices was determined and used to build the structural model (Chapter 4).

5.6 References

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

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

3. Wilkinson, K. A. et al. Structures of the HIV-1 Genome. *In preparation* (2007).

4. Paillart, J. C. et al. First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J. Biol. Chem.* **279**, 48397-48403 (2004).

5. Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5' region of the HIV-1 Genome. *J. Mol. Biol.* **336**, 369-79 (2004).

6. Baudin, F. et al. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *J. Mol. Biol.* **229**, 382-397 (1993).

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

8. Mathews, D. H. & Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **16**, 270 (2006).

9. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-free grammers for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 71 (2004).

10. Doshi, K. J., Cannone, J. J., Cobaugh, C. W. & Gutell, R. R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 105 (2004).

11. Rivas, E. & Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**, 583-605 (2000).

12. Workman, C. & Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.* **27**, 4816-4822 (1999).

13. Isel, C. et al. Structural basis for the specificity of the initiation of HIV-1 reverse transcription. *EMBO J.* **18**, 1038-1048 (1999).

14. Sakuragi, J., Iwamoto, A. & Shioda, T. Dissociation of genome dimerization from packaging functions and virion maturation of human immunodeficiency virus type 1. *J. Virol.* **76**, 959-67 (2002).

15. Fu, W., Gorelick, R. J. & Rein, A. Characterization of human immunodeficiency virus type 1 dimeric RNA from wild-type and protease-defective virions. *J. Virol.* **68**, 5013-5018 (1994).

16. Hu, W. S. & Temin, H. M. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proc. Natl Acad. Sci. USA* **87**, 1556-1560 (1990).

17. Andersen, E. S., Jeeninga, R. E., Damgaard, C. K., Berkhout, B. & Kjems, J. Dimerization and template switching in the 5' untranslated region between various subtypes of human immunodeficiency virus type 1. *J. Virol.* **77**, 3020-3030 (2003).

18. Paillart, J. C., Shehu-Xhilaga, M., Marquet, R. & Mak, J. Dimerization of retroviral RNA genomes: an inseparable pair. *Nature Rev. Microbiol.* 461-472 (2004).

19. Badorrek, C. S., Gherghe, C. M. & Weeks, K. M. Structure of an RNA switch that enforces stringent retroviral genomic RNA dimerization. *Proc. Natl Acad. Sci. USA* **103**, 13640-5 (2006).

20. Gherghe, C. & Weeks, K. M. The SL1-SL2 (stem-loop) domain is the primary determinant for stability of the gamma retroviral genomic RNA dimer. *J. Biol. Chem.* **281**, 37952-61 (2006).

21. Berkowitz, R., Fisher, J. & Goff, S. P. RNA packaging. *Curr. Top. Microbiol. Immunol.* **214**, 177-218 (1996).

22. Rein, A., Henderson, L. E. & Levin, J. G. Nucleic-acid-chaperone activity of retroviral nucleocapsid proteins: significance for viral replication. *Trends Biochem. Sci.* **23**, 297-301 (1998).

23. De Guzman, R. N. et al. Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* **279**, 384-8 (1998).

24. Amarasinghe, G. K. et al. NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the psi-RNA packaging signal. Implications for genome recognition. *J. Mol. Biol.* **301**, 491-511 (2000).

25. Rossio, J. L. et al. Inactivation of human immunodeficiency virus type 1 infectivity with preservation of conformational and functional integrity of virion surface proteins. *J. Virol.* **72**, 7992-8001 (1998).

26. Burns, J. C., Friedmann, T., Driever, W., Burrascano, M. & Yee, J. K. Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. *Proc. Natl. Acad. Sci. USA* **90**, 8033-8037 (1993).

27. Thomas, J. A. et al. Human immunodeficiency virus type 1 nucleocapsid zinc-finger mutations cause defects in reverse transcription and integration. *Virology* **353**, 41-51 (2006).

28. Giddings, M. C., Severin, J., Westphall, M., Wu, J. & Smith, L. M. A Software System for Data Analysis in Automated DNA Sequencing. *Genome Res.* **8**, 644-665 (1998).

29. Guex, N., Vasa, S. M., Wilkinson, K. A., Weeks, K. M. & Giddings, M. C. SHAPEfinder. *in preparation* (2007).