STRUCTURAL AND THERAPEUTIC INSIGHTS FROM THE HIV-1 RNA GENOME

Justin Thomas Low

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 Biochemistry and Biophysics.

Chapel Hill
2012

Approved by:
Howard M. Fried, Ph.D
Brian Kuhlman, Ph.D
Ronald I. Swanstrom, Ph.D
Kevin M. Weeks, Ph.D
David A. Wohl, M.D.
Abstract

JUSTIN LOW: Structural and Therapeutic Insights from the HIV-1 RNA Genome
(Under the direction of Kevin M. Weeks)

Infection with HIV currently affects an estimated 30-36 million people throughout the world. Due in part to the poor replication fidelity of this RNA virus, resistance to antiretrovirals develops rapidly. Finding new ways of targeting HIV is therefore an ever urgent need. However, despite the wealth of ongoing research in HIV drug development, most new drug candidates continue to target only a few well-defined protein domains, chosen for their functional importance in HIV replication. Targeting the RNA genome itself in a structure-directed manner presents an opportunity to greatly expand the repertoire of potential target sites for anti-HIV therapeutics. We use a high-resolution SHAPE-directed secondary structure model of an entire HIV-1 RNA genome (1) to refine existing models of the Gag-Pol frameshift element, an important regulatory element and promising therapeutic target, and (2) to investigate the structural determinants for RNAi-based inhibition of HIV-1. We show that the Gag-Pol frameshift element folds into a complex structure that is distinct from currently accepted models and capable of switching between two different conformations. Additionally, we discovered that there exists a strong correlation between shRNA-mediated inhibition of HIV-1 production in a quantitative cell-based
assay and very simple thermodynamic features in the SHAPE-directed RNA genome structure model. Both of these results are highly dependent on having an accurate secondary structural model, as obtained by SHAPE data. We anticipate that these results will be broadly applicable to RNA-directed antiretroviral development efforts.
To my parents


Acknowledgements

I am grateful to many advisors, peers, friends, and family who have enriched my graduate experience. I wish to especially thank:

My advisor, Dr. Kevin Weeks, for his enthusiasm and encouragement throughout my graduate career.

Fellow Weeks lab members, for four fun years of science, friendship, and entertainment.

My thesis committee members, for their time, advice and interest in my work.

Dr. Robert Gorelick, for producing HIV RNA and for entertaining so many viral assay requests and ideas.

Drs. Steffie Knoepfel and Ben Berkhout, for their indispensable role in our shRNA collaboration.

Dr. Eugene Orringer and the staff, students, and advisors of the MD-PhD program, for their continuing support.

My family, for their love and support.
# Table of Contents

List of Tables .................................................................................................................................................. x

List of Figures .................................................................................................................................................. xi

List of Abbreviations ...................................................................................................................................... xiii

1. Introduction ................................................................................................................................................. 1
   1.1 Introduction ............................................................................................................................................. 1
   1.2 The HIV RNA as a target for antiretrovirals ....................................................................................... 2
   1.3 Structure-based targeting of HIV-1 RNA ............................................................................................. 3
   1.4 SHAPE-directed RNA secondary structure prediction ................................................................. 3
      1.4.1 The RNA secondary structure prediction problem ...................................................................... 3
      1.4.2 Dynamic programming algorithms for RNA secondary structure prediction ......................... 4
      1.4.3 Incorporating experimental data .................................................................................................. 6
      1.4.4 Towards accurate SHAPE-directed secondary structure prediction ...................................... 7
      1.4.5 SHAPE experimental procedure ................................................................................................. 9
      1.4.6 SHAPE-constrained RNA structure folding ............................................................................... 14
      1.4.7 Secondary structure model of an entire HIV-1 genome ............................................................ 19
   1.5 Research overview ............................................................................................................................... 20
      1.5.1 Structure of the HIV-1 frameshift domain .................................................................................. 20
1.5.2 Structure-based design of shRNA inhibitors of HIV-1 ................................................. 20
1.6 Acknowledgements.................................................................................................................. 21
1.7 References............................................................................................................................ 22

2. Structure of the HIV-1 frameshift element, investigated by LNA
binding and SHAPE probing ........................................................................................................ 27
2.1 Introduction ............................................................................................................................ 27
2.2 Results ................................................................................................................................... 32
  2.2.1 Both natively extracted ex *virio* and heat-denatured RNAs
       adopt similar structures in the frameshift region ............................................................ 32
  2.2.2 LNA binding experiments support the SHAPE-directed
       frameshift model ............................................................................................................... 37
  2.2.3 LNA binding can switch the SHAPE-directed alternate
       lower stem to the conventional lower stem ................................................................. 37
  2.2.4 Formation of the conventional lower stem destabilizes the
       slippery sequence helix ..................................................................................................... 42
  2.2.5 Formamide denaturation experiments reveal the relative
       stabilities of the four frameshift structure helices ....................................................... 42
  2.2.6 SHAPE probing of *in virio* genomic RNA reveals a less
       structured frameshift domain .......................................................................................... 48
2.3 Discussion .............................................................................................................................. 48
2.4 Methods................................................................................................................................ 53
  2.4.1 HIV-1 virion production ................................................................................................. 53
  2.4.2 Extraction of RNA genomes from virions ...................................................................... 53
  2.4.3 Folding of *ex virio* RNA ............................................................................................. 54
  2.4.4 Formamide denaturation .............................................................................................. 54
  2.4.5 LNA oligonucleotide design ......................................................................................... 55
2.4.6 LNA binding to genomic RNA ................................................................. 55
2.4.7 SHAPE modification of \textit{ex virio}, LNA-bound, and formamide denatured RNA ................................................................. 55
2.4.8 Primer extension and capillary electrophoresis detection of SHAPE adduct sites ................................................................. 56
2.4.9 Data processing .................................................................................. 57
2.4.10 Secondary structure modeling .......................................................... 58
2.5 Acknowledgements ................................................................................ 58
2.6 References ............................................................................................ 60

3 SHAPE-directed discovery of potent shRNA inhibitors of HIV-1 .............. 64
3.1 Introduction ............................................................................................ 64
3.2 Results ................................................................................................... 67
  3.2.1 Strategy ............................................................................................. 67
  3.2.2 Concentration dependence of shRNA inhibition ............................... 70
  3.2.3 Weak target folding energy characterizes effectively repressed sequences ................................................................. 73
  3.2.4 Strong total binding energy characterizes effectively repressed sequences ................................................................. 76
  3.2.5 Strong thermodynamic correlations are specific to the SHAPE-directed RNA structure model ............................................. 76
  3.2.6 Experimental validation of shRNA design rules ............................... 77
  3.2.7 shRNA repression and toxicity in human T cells ......................... 80
  3.2.8 Partial SHAPE information is sufficient to identify potent shRNA inhibitors ................................................................. 83
3.3 Discussion ............................................................................................. 84
3.4 Methods................................................................................................................................. 88

3.4.1 Plasmid constructs ........................................................................................................... 88

3.4.2 Cell culture ....................................................................................................................... 88

3.4.3 Transfections and HIV-1 production experiments ......................................................... 89

3.4.4 Lentiviral transduction, HIV- challenge experiments, and competitive cell growth assay. ................................................................................................................................. 90

3.4.5 Free energy calculations ................................................................................................. 91

3.4.6 Correlation calculations for other algorithms ............................................................... 91

3.4.7 Viral production datasets ............................................................................................... 92

3.5 Acknowledgements ............................................................................................................ 93

3.6 Appendix 1 ......................................................................................................................... 94

3.7 Appendix 2 ......................................................................................................................... 97

3.8 Appendix 3 ......................................................................................................................... 98

3.9 References .......................................................................................................................... 99

4 Conclusions and clinical relevance ......................................................................................... 103

4.1 Structural organization of the HIV-1 Gag-Pol frameshift element ................................ 103

4.2 Structural requirements for effective shRNA targeting of HIV-1 ................................. 103

4.3 Applications to RNA structure probing ......................................................................... 104

4.4 Clinical relevance .............................................................................................................. 104

4.5 References .......................................................................................................................... 106
List of Tables

Table 1.1  RNA secondary structure prediction accuracies for folding calculations performed without and with SHAPE constraints............................................................................................................................11

Table 2.1  LNA-containing DNA oligonucleotide sequences and target binding sites........................................................................................................................................36

Table 3.1  Correlation between HIV-1 inhibition and si/shRNA target prediction algorithms........................................................................................................................66
List of Figures

Figure 1.1 Overview of SHAPE experimental and data analysis steps.................................10

Figure 1.2 SHAPE-derived pseudo-free energy function and base pair prediction sensitivities for *E. coli* 23S rRNA.................................................................16

Figure 1.3 Summary of thermodynamic and SHAPE-derived free energy change contributions for a simple HIV-1 hairpin.........................................................18

Figure 2.1 Conventional and SHAPE-directed models of the frameshift domain.................................................................29

Figure 2.2 Frameshift domain reactivities of extracted *ex vitro* RNA and refolded heat denatured RNA using three different SHAPE reagents.........................................................33

Figure 2.3 Standard SHAPE reactivities (1M7) of extracted *ex vitro* RNA and refolded heat denatured RNA for a region spanning ~700 nts surrounding the frameshift domain........................................34

Figure 2.4 Frameshift region SHAPE reactivities and predicted secondary structure models for HIV-1 RNA bound to LNAs.........................38-40

Figure 2.5 SHAPE reactivity profiles for formamide-denatured RNA...............................43

Figure 2.6 Secondary structure models for formamide-denatured RNA...........................45

Figure 2.7 SHAPE reactivity profiles and predicted secondary structure models for *in vitro* compared with *ex vitro* RNA.................................47

Figure 2.8 Proposed model for frameshift domain unwinding during translation.................................................................51

Figure 3.1 The guide strand-target RNA interaction equilibrium.........................................68

Figure 3.2 Relative levels of virus production for five shRNAs..............................................71
| Figure 3.3 | HIV-1 genome locations of target shRNA sequences used in this study ...............................................................72 |
| Figure 3.4 | Correlation coefficients (r) between calculated target folding energies, $\Delta G_{\text{target}}$, and experimental activity values for the 84 shRNAs in the training dataset.................................................................74 |
| Figure 3.5 | Relative viral production versus target folding energies, $\Delta G_{\text{target}}$, and total binding energies, $\Delta G_{\text{total}}$, for the 84 shRNAs in the training dataset ........................................................................75 |
| Figure 3.6 | Prediction success rates of designed shRNAs ...........................................................................................................78 |
| Figure 3.7 | Inhibition of HIV-1 replication in transduced SupT1 T cells .................................................................................81 |
| Figure 3.8 | Competitive cell growth curves of representative potent shRNAs from the test set ........................................................................................................................................................................................82 |
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta G )</td>
<td>Gibbs free energy</td>
</tr>
<tr>
<td>1M6</td>
<td>1-methyl-6-nitroisatoic anhydride</td>
</tr>
<tr>
<td>1M7</td>
<td>1-methyl-7-nitroisatoic anhydride</td>
</tr>
<tr>
<td>AIDS</td>
<td>Acquired immune deficiency syndrome</td>
</tr>
<tr>
<td>CA</td>
<td>Capsid protein</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CMV</td>
<td>cytomegalovirus</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco's modified Eagle's medium</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DMSO</td>
<td>dimethyl sulfoxide</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td><em>ex vitro</em></td>
<td>Latin expression meaning extracted from virions</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>FCS</td>
<td>fetal calf serum</td>
</tr>
<tr>
<td>Gag</td>
<td>Group-specific antigen</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>HEPES</td>
<td>N-2-hydroxyethylpiperazine-N’-2-ethanesulfonic acid</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td><em>in vitro</em></td>
<td>Latin expression meaning in a test tube</td>
</tr>
<tr>
<td><em>in virio</em></td>
<td>Latin expression meaning inside virions</td>
</tr>
<tr>
<td>IRES</td>
<td>Internal ribosomal entry site</td>
</tr>
<tr>
<td>kcal</td>
<td>kilocalorie</td>
</tr>
<tr>
<td>KOAc</td>
<td>potassium acetate</td>
</tr>
<tr>
<td>Leu</td>
<td>Leucine</td>
</tr>
<tr>
<td>LNA</td>
<td>locked nucleic acid</td>
</tr>
<tr>
<td>MOI</td>
<td>multiplicity of infection</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td><em>N</em></td>
<td>number of nucleotides in an RNA</td>
</tr>
<tr>
<td>NED</td>
<td>proprietary fluorescent dye</td>
</tr>
<tr>
<td>NMIA</td>
<td>N-methylisatoic anhydride</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>NN</td>
<td>nearest neighbor</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>NTP</td>
<td>nucleotide triphosphate</td>
</tr>
<tr>
<td><em>O</em></td>
<td>Big O notation, the response of an algorithm to changes in input size</td>
</tr>
<tr>
<td>Phe</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>PI</td>
<td>protease inhibitor</td>
</tr>
<tr>
<td>Pol</td>
<td>precursor polyprotein for retroviral enzymes</td>
</tr>
<tr>
<td>PPV</td>
<td>positive predictive value</td>
</tr>
</tbody>
</table>
\( r \)  Pearson product-moment correlation coefficient

RISC  RNA-induced silencing complex

RNA  ribonucleic acid

RNAi  RNA interference

rRNA  ribosomal RNA

SHAPE  Selective 2'-Hydroxyl Acylation analyzed by Primer Extension

shRNA  short hairpin RNA

siRNA  short interfering RNA

TE  10 mM Tris (pH 7.5), 1 mM EDTA

Tris  tris(hydroxymethyl)aminomethane

tRNA  transfer RNA

VIC  proprietary fluorescent dye
CHAPTER 1

INTRODUCTION

1.1 Introduction

Infection with human immunodeficiency virus (HIV) is a major public health concern. HIV is a positive-sense RNA retrovirus that infects cells of the immune system such as T helper cells and macrophages. Ultimately, the destruction of these cells renders the host susceptible to numerous infections that a healthy immune system would suppress. These are termed opportunistic infections and are the clinical hallmarks of acquired immune deficiency syndrome (AIDS). During latent HIV infection, a copy of the viral genome is incorporated into the genomic DNA of infected immune cells. Mitigating the disease progression to AIDS therefore requires lifelong treatment with antiretroviral drugs, and there remains a need for new classes of antiretroviral medications. All current therapeutic approaches target processes crucial to the viral replication cycle.

The replication cycle begins when an HIV virion recognizes target host cells via binding of glycoproteins on the viral envelope with host cell receptors. This binding event triggers the fusion of viral and host cell plasma membranes allowing entry of viral contents into the host cell. The viral contents contain two copies of the viral genome.

Portions of this chapter have been published in Low, J.T. and Weeks, K.M. Methods. 2010. 52:150-158.
packaged with structural proteins matrix, capsid, and nucleocapsid, and viral enzymes reverse transcriptase, integrase, and protease. This package is uncoated to release the RNA genome and viral enzymes. The RNA genome is used as a template for reverse transcription into cDNA by the viral reverse transcriptase. Following transport to the nucleus, the double-stranded DNA product is integrated into the host cell genome by the viral integrase enzyme. Once integrated, full-length viral RNA can be generated by the host transcriptional machinery. This RNA can function as genomic RNA, mRNA for Gag and Gag-Pol, or spliced to form subgenomic mRNA. These viral mRNAs are translated by host ribosomes and the resulting precursor proteins are assembled together with two copies of the RNA genome. This assembly buds out of the host cell, taking part of the cell membrane as its envelope. Precursor proteins within this immature viral particle then undergo proteolytic cleavage via the action of the viral protease enzyme to generate the mature, infectious particle.

1.2 The HIV RNA as a target for antiretrovirals

In principle, each step of this replication cycle could be a target for antiretroviral therapeutics. Currently available drugs target critical proteins involved in four of these stages. Three of these classes of drugs target the three viral enzymes (reverse transcriptase, protease, and integrase), while the fourth targets proteins involved in fusion of virus and host cell membranes. However, no existing drugs target the HIV RNA, which is a central player in many stages of the replication cycle. The high mutation rate of HIV poses special challenges for the development of antiretroviral medications. Nevertheless, the relatively short length of the HIV genome (~9200
nucleotides) means that the essential, evolutionarily stable, replication information encoded within the genome must be relatively densely packed.

1.3 Structure-based targeting of HIV-1 RNA

In principle, the regular base-pairing properties of RNA make designing oligonucleotide inhibitors conceptually straightforward, as one simply uses the reverse complement of target sequences. However, RNA folds into secondary and higher order structures that are critical for carrying out their functions. These structures can reduce the accessibility of component sequences towards oligonucleotide binding. Additionally, three-dimensional structural motifs represent an additional avenue for therapeutic recognition beyond the linear primary sequence. Detailed knowledge of RNA structure is consequently critical for developing RNA-directed inhibitors of HIV-1.

1.4 SHAPE-directed RNA secondary structure prediction

1.4.1 The RNA secondary structure prediction problem

Determining the complete three-dimensional (termed the tertiary) structure is the ultimate goal for many RNAs. However, only limited sets of RNAs are candidates for current high-resolution crystallography and NMR approaches. A simpler problem is to determine the base-pairing pattern (termed the secondary structure) of an RNA. Secondary structure determination, independent of higher order structural information, is possible because the hydrogen bonding and stacking interactions that collectively form secondary structure are usually stronger than tertiary interactions (1-4), and because RNA folding is often hierarchical (5,6), with many secondary structural motifs forming prior to tertiary contacts. Additionally, knowledge of the secondary
structure greatly restricts possible three-dimensional conformations and facilitates tertiary structure prediction (7-9). Moreover, a subset of RNA functions may depend more directly on secondary structural motifs than on global folds.

Insight into the secondary structure can be gleaned using computer-based predictions performed using the sequence alone, or in combination with sequence alignment information or experimental data. Sequence-based folding generally includes two main elements: an energy function based on experimentally derived thermodynamic parameters, and an algorithm that explores the conformational space available to the RNA and ranks computed structures. Most energy functions use the Turner et al. (10,11) set of nearest neighbor parameters, derived from optical melting experiments. A summary of these parameters is available at the Nearest-Neighbor Database (12). Exploring conformational space is challenging because of the vast number of possible secondary structures, which is estimated to scale exponentially as \(~1.8^N\), where \(N\) is the number of nucleotides in the RNA (13). This means that a “brute force” approach that samples every possible conformation is impossible both from a computational standpoint and from the perspective of efficient RNA folding in vivo. Consequently, the intrinsic thermodynamics and kinetics of RNA folding must conspire to restrict the folding pathway to a narrow subset of these structures, only one (or perhaps a few) of which is likely to dominate the equilibrium ensemble. Especially for short RNAs, thermodynamic considerations are likely paramount and thus the structure with the lowest free energy is the biologically active one.

1.4.2 Dynamic programming algorithms for RNA secondary structure prediction

Programs based on the Zuker dynamic programming algorithm (14,15) are
widely used to search for the minimum free energy structure (16-22). These
algorithms are deterministic, meaning that given a defined set of energy rules, they
always find the lowest free energy structure. The Zuker algorithm scales as $O(N^3)$ in
time, where $N$ is the number of nucleotides in the sequence. This means that doubling
the sequence length requires eight times as much time to predict the structure.
Nevertheless, on modern computers, the time to make a prediction is reasonably fast.
The guarantee that the optimal structure can be computed and the relative
computational efficiency are made possible, first, by incorporating simplifying
assumptions into the energy function, and second, by limiting the types of allowed RNA
folds.

The total energy is assumed to be a simple sum over all energetic components
that characterize local structural elements. Two features primarily contribute to the
total energy: negative (favorable) free energies arising from stabilizing base stacking
and hydrogen bonding interactions in and adjacent to helices, and positive
(unfavorable) free energies arising from the entropic cost of restricting conformational
freedom in loops. Helix energy terms are sequence-dependent, reflect the energetic
bonus of adding a base pair to a helix, and implicitly include both canonical hydrogen
bonding and base stacking. These terms depend solely on interactions involving
adjacent base pairs or interactions at the ends of helices. This local interaction model is
termed the nearest-neighbor approximation (23).

The dynamic programming algorithm calculates the energy of the lowest free
energy structure (but does not compute the complete structure itself) for all possible
subsequences of an RNA. This approach is efficient because the solution for each
subsequence is computed from solutions for pre-computed smaller subsequence,
allowing the energies for each structural element to be computed only once. The
results are stored in triangular $N \times N$ arrays whose elements $ij$ represent the optimal
folding energy for an RNA subsequence from nucleotide $i$ to nucleotide $j$. The structure
for the entire RNA sequence is obtained by tracing a structure through an optimal
combination of component subsequences in the array (24).

Thermodynamics-based dynamic programming algorithms have several
limitations. First, computing the minimum free energy structure in a relatively efficient
$O(N^3)$ manner excludes consideration of non-nested topologies. These include the
biologically important case of pseudoknots, in which a loop in one helix forms the stem
of another helix. Second, the assumption that the minimum free energy structure is the
biologically active one may not always hold for larger RNAs, where folding kinetics may
play a prominent role. Third, the biologically relevant ensemble may be dominated by
several interconverting states, making a single structural model inadequate. Finally,
incomplete thermodynamic rules and the simplifications inherent in the nearest
neighbor model introduce uncertainty to the energy calculations.

The net effect of these limitations is that the current best-performing algorithms
achieve prediction accuracies of 50-70% (11,25-29). Accuracies tend to be especially
poor for larger RNAs. For example, for *Escherichia coli* 16S rRNA, which is probably the
most thoroughly studied large RNA, the prediction accuracy based on sequence alone is
less than 50% (26,30).

1.4.3 **Incorporating experimental data**

Significant improvements to RNA secondary structure prediction can be
achieved when computer predictions are constrained by experimental data derived from structure-sensitive enzymatic cleavage and chemical probing reagents \(11,31,32\). However, the net improvement gained from using traditional reagents is often modest. First, traditional reagents tend to react with only a subset of nucleotides, so the absence of reactivity cannot usually be taken as evidence for likely base pairing. Second, different reagents are required to react with all four RNA nucleotides and some of the more useful reagents, like dimethyl sulfate (DMS), react at different base functional groups depending on the nucleotide. Third, the dynamic range for many reagents is low, making it difficult to distinguish levels of reactivity beyond a qualitative “low,” “medium,” and “high” scale. Finally, while alternative chemistries such as in-line probing \(33\) and hydroxyl radical footprinting \(34\) provide valuable insight into higher order structures and react broadly with all four RNA nucleotides, they less directly report the intrinsic nucleotide flexibilities that largely characterize secondary structure. Thus, it is challenging to create quantitative relationships between reagent reactivity and RNA secondary structure.

1.4.4 Towards accurate SHAPE-directed secondary structure prediction

Our lab has developed the Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) \(35,36\) chemical probing technology that largely addresses these challenges. SHAPE yields quantitative reactivity information for nearly every nucleotide in an RNA. Advantageously, SHAPE is not limited by RNA size and is remarkably insensitive to solvent accessibility \(35,37,38\). Additionally, SHAPE can be applied to both \textit{in vitro} transcripts and also to RNAs from native-like cellular and viral environments. Combining SHAPE information with a thermodynamics-based dynamic
programming algorithm, as implemented in RNAstructure (11), results in highly accurate secondary structure models (30). This approach has been benchmarked and shown to yield secondary structures for diverse RNAs, including the *E. coli* 16S rRNA (1542 nucleotides), with >95% accuracy as judged by sensitivity (percentage of known base pairs predicted correctly) and positive predictive value (PPV, percentage of predicted base pairs in the known structure) (30) (Table 1.1). SHAPE has been used to propose experimentally-informed secondary structural models for many RNAs, including an entire HIV-1 genome (38).

SHAPE technology involves covalently modifying RNA in a structure-dependent manner (selective 2'-hydroxyl acylation), followed by detecting the sites of modification by primer extension (original protocols described in (39,40)). The RNA modification involves the nucleophilic attack of the 2'-hydroxyl group of the RNA ribose moiety on an electrophilic SHAPE reagent to form a 2'-O-adduct (Fig. 1A) (35). This reaction occurs more readily with conformationally unconstrained or flexible nucleotides such as those in single stranded regions, loops, or bulges (spheres, Fig. 1B). Flexible nucleotides react preferentially because they more readily sample conformations conducive to nucleophilic attack. In contrast, nucleotides in highly structured regions are conformationally constrained and less frequently achieve an optimal geometry, making them less reactive towards SHAPE reagents. In general, solvent inaccessible, but unconstrained, nucleotides are still reactive by SHAPE.

Following modification of the RNA, modified positions are detected by primer extension using end-labeled, target-specific primers and a thermostable reverse transcriptase (Fig. 1C). Since the reverse transcriptase enzyme cannot proceed past 2'-
O-modified sites in RNA, the lengths of the resulting cDNA products correspond to the distance between the primer binding and 2’-O-adduct sites. Due to differential modification of structured versus unstructured nucleotides, the frequency of producing a given cDNA product reflects the underlying RNA structure. Comparison with dideoxy nucleotide sequencing ladders allows each SHAPE reagent-dependent peak to be matched with the corresponding nucleotide position (Fig. 1D).

SHAPE technology can be implemented in an efficient and high-throughput way by automated capillary electrophoresis using DNA sequencing instruments (Fig. 1E). The capillary electrophoresis data are analyzed using the software program QuShape (41) to yield normalized SHAPE reactivity values (Fig. 1F). These reactivities can be converted to $\Delta G_{SHAPE}$ pseudo-free energy terms and used with the energy function in the RNAstructure program to yield, generally highly accurate, secondary structure models for RNA (Table 1.1) (11,30).

1.4.5 SHAPE experimental procedure

The experimental component of a SHAPE analysis has been recently reviewed in detail (40,42). Briefly, RNA is modified in a structure-selective way using an electrophilic SHAPE reagent. While SHAPE has been most commonly performed on in vitro RNA transcripts or RNAs extracted from biological environments, SHAPE reagents readily cross biological membranes and, for example, react with RNAs inside authentic HIV-1 particles (36).

Approximately 2 pmol of RNA is needed in each primer extension reaction to obtain adequate signal intensity in the capillary electrophoresis detection step, using commercially available instruments. We routinely achieve read lengths of 300 – 650
**Figure 1.1** Overview of SHAPE experimental and data analysis steps. Adapted from ref. (44)
<table>
<thead>
<tr>
<th>RNA</th>
<th>Size (nts)</th>
<th>No constraints</th>
<th>With SHAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>PPV</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Yeast tRNA $^{\text{Asp}}$</td>
<td>75</td>
<td>95 95</td>
<td>100 100</td>
</tr>
<tr>
<td>HCV IRES domain</td>
<td>95</td>
<td>57 59</td>
<td>96 100</td>
</tr>
<tr>
<td>$\text{Bacillus subtilis}$</td>
<td>154</td>
<td>53 51</td>
<td>75 83</td>
</tr>
<tr>
<td>RNase P, specificity domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bl3 group I intron, P546</td>
<td>155</td>
<td>43 44</td>
<td>96 98</td>
</tr>
<tr>
<td>domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{E. coli}$ 16S rRNA</td>
<td>1542</td>
<td>50 46</td>
<td>97 95</td>
</tr>
</tbody>
</table>

**Table 1.1** RNA secondary structure prediction accuracies for folding calculations performed without and with SHAPE constraints.
nucleotides in each primer extension reaction (43,44). For longer RNAs, information obtained from multiple primers, with overlapping read windows, can be combined to create datasets spanning arbitrarily long lengths (30,36,38).

To maintain a native-like conformation, the RNA must be renatured (in vitro transcripts) or maintained (RNAs from cellular or viral sources) in a physiological-like folding buffer. We typically use a simple standard solution (50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM MgCl₂), and incubate at 37 °C for 10 – 30 min prior to modification. SHAPE works well under a wide variety of conditions, including in the presence of biological amines and carbohydrates and proteins that bind RNA. The main requirement for SHAPE is that the pH be maintained in the 7.6 – 8.3 range (35).

RNA structure is interrogated by adding a SHAPE reagent. Initial work in our laboratory used the commercially available NMIA reagent (39); more recent work has utilized the faster-reacting 1M7 reagent, whose synthesis is described in (45). A variety of other reagents with subtly different reactivity preferences have also been developed to identify nucleotides engage in slower (46) or faster (47) conformational dynamics. Additionally, a SHAPE reagent that reacts preferentially with nucleotides that present an unoccupied nucleobase for π-π stacking is also under development (48). The SHAPE reagent is dissolved in DMSO and added to the RNA solution to a final concentration of about 5 mM. The optimal reagent concentration varies and can be system-specific: too high a concentration of SHAPE reagent results in significant signal decay and reduced read lengths, while too low a concentration yields data with a poor signal. Background signals in the primer extension reaction are measured by performing a no-reagent
control in which DMSO is added in place of the SHAPE reagent, in an otherwise identical reaction. Both reactions should be incubated at 37 °C for either 35 min if using NMIA or 70 sec if using 1M7. Both reagents self-quench by reacting with water in the aqueous solution.

Following an ethanol precipitation step, fluorescently-labeled primers are annealed to the (+) and (–) reagent-treated RNA and to untreated RNAs or DNA plasmids (the latter are used for sequencing). A thermostable reverse transcriptase enzyme is used for the primer extension reactions to convert the structural information into cDNA libraries. We perform the separation step in two capillaries: the (+) reagent reaction and one sequencing ladder in one capillary and the (–) reagent reaction and an identical sequencing ladder in a second capillary (41). Since each capillary contains two cDNA libraries (one (+) or (-) reagent library and one sequencing ladder), we use two fluorescent dyes, chosen to have similar electrophoretic mobilities to simplify the alignment of the electropherograms during the data processing steps. Reagent traces from both capillaries are aligned to each other and to their primary nucleotide sequences using their identical sequencing ladders. The cDNA products are recovered by ethanol precipitation, resuspended in formamide, and resolved on a commercial capillary electrophoresis DNA sequencing instrument. Electrophoresis data are processed using custom software (41) to created SHAPE reactivities that are normalized on a scale where a normalized reactivity of 1.0 is defined as the average intensity of the top 10% most reactive peaks, excluding a few highly reactive nucleotides taken to be outliers. Reactivities typically span a scale from 0 to ~1.5, where 0 indicates no reactivity (and a highly constrained nucleotide) and reactivities
>0.7 typically indicate highly flexible nucleotides.

### 1.4.6 SHAPE-constrained RNAstructure folding

A major challenging endeavor in RNA biology is to consistently and efficiently develop correct secondary structure models for RNAs of arbitrary length and complexity. The thermodynamics-based computational methods outlined above (Section 1.1) are highly useful for rapid computation of candidate structural models. However, prediction accuracies are inconsistent for many RNAs and tend to be particularly poor for large RNAs. These limitations can be broadly attributed to simplifications inherent in the nearest-neighbor model and incomplete knowledge of RNA energetics. However, for many RNAs, it is possible to obtain robust secondary structure predictions by incorporating SHAPE reactivities into the energy function used in a nearest neighbor dynamic programming algorithm. This approach has been implemented in the RNAstructure program.

The RNAstructure energy function is modified by adding pseudo-free energy change terms derived from SHAPE reactivities. This approach is grounded in the observation that SHAPE reactivities correlate strongly with local nucleotide flexibility (35,37) and, thus, also with the probability that a nucleotide is single stranded. The NMIA and 1M7 SHAPE reagents react with all four RNA nucleotides with limited base-dependent preferences (49). It is therefore possible to create a softer, continuous, and more physically grounded restraint function than is typically used with conventional chemical mapping reagents that exhibit strong idiosyncratic and nucleotide-specific reactivities. In essence, these additional energetic terms provide a knowledge-based correction to the nearest-neighbor energy function.
We derive a pseudo-free energy change term for each base-paired residue $i$ from its SHAPE reactivity:

$$
\Delta G_{\text{SHAPE}} \ (i) = m \ln(\text{SHAPE reactivity}(i) + 1) + b \tag{1.1}
$$

The empirical parameters $m$ and $b$ serve to scale the strength of the experimental contribution to the energy function (Fig. 1.2 A). The intercept $b$ represents the pseudo-free energy contribution of a base-paired nucleotide whose SHAPE reactivity is zero. The sign of $b$ is negative to reflect an energetic bonus for base pairing by constrained nucleotides. In contrast, the slope $m$ represents the strength of the energetic penalty assigned for pairing nucleotides with high SHAPE reactivities and consequently has a positive sign.

Optimal values for $m$ and $b$ were determined by assessing the prediction accuracy for *E. coli* 23S rRNA over a range of slope and intercept values (30). This work identified $m = 2.6 \text{ kcal/mol}$ and $b = -0.8 \text{ kcal/mol}$ as optimal values for folding large ribosomal RNAs and, importantly, also established these values as being located at the center of a “sweet spot” of a broad set of $m$ and $b$ values that yields accurate SHAPE-directed structure predictions (30) (emphasized in red, Fig. 1.2B). Given the large size (2,904 nts) of the *E. coli* 23S rRNA and the diversity of structural motifs it contains, these parameter values are also likely to work well for other RNAs. We empirically find this to be the case, although slightly different parameter values, still in the sweet spot (Fig. 1.2B), can be chosen heuristically to refine predictions for some RNAs (38). The logarithmic relationship between SHAPE reactivities and the derived $\Delta G_{\text{SHAPE}}$ term has the effect of forgiving differences among the most highly reactive nucleotides. The usefulness of this behavior reflects the observation that highly reactive nucleotides are the most sensitive
Figure 1.2 (A) SHAPE-derived pseudo-free energy function and (B) base pair prediction sensitivities for E. coli 23S rRNA for a range of slope (m) and intercept (b) values (Eqn. (1.1)). Optimal values of m = 2.6 kcal/mol and b = -0.8 kcal/mol are depicted by the white box. Adapted from Ref. (30).
to signal processing artifacts and have the highest variance. Furthermore, the logarithmic relationship between SHAPE reactivity and pseudo-free energy change loosely reflects a statistical mechanical interpretation of SHAPE reactivity, which indirectly measures the number of conformational states accessible to each nucleotide.

We illustrate the combined nearest-neighbor and SHAPE energy function, as implemented in RNAstructure, for a short fragment of an HIV-1 RNA sequence (Fig. 1.3). Nucleotides are color-coded by their SHAPE reactivities as reported in (38). The energy function (12) includes favorable nearest-neighbor energy terms for helix stacking (in green, Fig. 1.3) and entropic penalties for anchoring loops (in red, Fig. 1.3). Stacking terms are added for all helical interactions, including terminal mismatches and dangling ends at helix termini, as well as for coaxial stacking between adjacent helices (25,50). Stacking terms depend on the sequence identity of all nucleotides participating in the stack (the nearest-neighbors), while loop entropy terms depend primarily on the number of nucleotides in the loop.

In contrast to the nearest-neighbor thermodynamics-based energy parameters, pseudo-free energy terms ($\Delta G_{\text{SHAPE}}$) are calculated for each nucleotide individually (Fig. 1.3, black and grey numbers). Nucleotides with high SHAPE reactivities have positive pseudo-free energies and those with low SHAPE reactivities have negative pseudo-free energies (Eqn. 1.1). $\Delta G_{\text{SHAPE}}$ terms are only added to the free energy calculation for base paired nucleotides (Fig. 1.3, black numbers). $\Delta G_{\text{SHAPE}}$ terms for nucleotides at the ends of helices are counted once and those in the interior of helices are counted twice since they contribute to two stacks (Fig. 1.3, blue $\times 1$ and $\times 2$ symbols, respectively). Base paired nucleotides with high SHAPE reactivities contribute large
**Figure 1.3** Summary of thermodynamic and SHAPE-derived free energy change contributions for a simple HIV-1 hairpin (NL4-3 nucleotides 594–626) (38). Favorable nearest-neighbor stacking and unfavorable loop thermodynamic terms are shown in green and red, respectively. The total nearest-neighbor free energy change $\Delta G_{NN}$ is the sum over all these contributions. $\Delta G_{SHAPE}$ pseudo-free energy change terms are shown for base-paired (black) and non-base-paired (grey) nucleotides; only base-paired values are included in the net free energy change. The $\Delta G_{SHAPE}$ term is added once for each nucleotide at the ends of helices and twice for interior nucleotides (blue symbols). The $\Delta G_{SHAPE}$ calculations used $m = 3.0$ kcal/mol and $b = -0.6$ kcal/mol. The total folding free energy change, $\Delta G_{total}$, is the sum of nearest-neighbor and SHAPE-derived contributions.
positive pseudo-free energies (for example, see the red G in Fig. 1.3). Such nucleotides are more likely to be allowed at the end, as opposed to the interior, of a helix because they are added to the total free energy only once. This is consistent with the observation that nucleotides at the ends of helices are more dynamic, and experience greater fraying, than interior nucleotides. On the other hand, unpaired nucleotides with low SHAPE reactivities represent an incomplete model and could suggest non-canonical interactions that are not currently predicted by the algorithm (for example, see the tandem black G residues in the apical loop of Fig. 1.3). The total folding energy ($\Delta G_{\text{total}}$) is simply the sum of all nearest neighbor thermodynamic terms ($\Delta G_{\text{NN}}$) and pseudo-free energy ($\Delta G_{\text{SHAPE}}$) contributions (Fig. 1.3). This sum is used to rank RNA structures and should not be interpreted as a physical energy because it includes both thermodynamic terms and SHAPE-derived pseudo-free energy change terms.

1.4.7 Secondary structure model of an entire HIV-1 genome

This SHAPE-directed RNAstructure folding approach was used to create a nucleotide-resolution secondary structure model of an entire HIV-1 genome (38). Structural models proposed by this work largely agreed with prior studies of important regulatory motifs. However, SHAPE-directed modeling suggested an alternative model for an important regulatory domain, the Gag-Pol frameshift element. Additionally, prior studies collectively span only about 15% of the HIV-1 genome, and the SHAPE-directed model revealed many more structured elements in regions that had not been investigated in prior work.
1.5 Research overview

1.5.1 Structure of the HIV-1 frameshift domain

In Chapter 2, we consider the structure of the Gag-Pol frameshift domain, given both its functional importance for viral replication and the disagreement between the SHAPE-directed and conventionally accepted models. This important regulatory element uses a highly structured motif to regulate translation of all viral-encoded enzymes and is therefore a promising potential therapeutic target in the HIV-1 RNA genome. Using high-affinity locked nucleic acid oligonucleotides to disrupt three helices unique to this new model, we obtain strong support for the SHAPE-directed frameshift domain. We also use SHAPE to examine the thermodynamic stability of helices in the frameshift domain by probing HIV-1 RNA in the presence of the denaturant formamide. We discover that the frameshift domain is anchored by two very stable helices, and that less stable helices within the domain have the propensity to both unfold and switch from the SHAPE-directed to the conventional conformation. We hypothesize that this conformational switch occurs during the frameshifting process, raising the possibility of a functional role for both the SHAPE-directed and conventional models. We furthermore find that formation of the SHAPE-directed model is dependent upon the presence of the complete 140-nucleotide domain, underscoring the importance of global sequence context for RNA folding.

1.5.2 Structure-based design of shRNA inhibitors of HIV-1

In Chapter 3, we use whole-genome SHAPE information to investigate the structural basis for targeting HIV-1 RNA via the RNA interference pathway. The cellular
RNA interference (RNAi) pathway can be exploited using short hairpin RNAs (shRNAs) to durably inactivate pathogenic genes. Prediction of optimal target sites is notoriously inaccurate and current approaches applied to HIV-1 show weak correlations with virus inhibition. In contrast, when using the SHAPE-directed model of an entire HIV-1 genome, we discovered strong correlations between inhibition of HIV-1 production in a quantitative cell-based assay and very simple thermodynamic features in the target RNA. Strongest inhibition occurs at RNA target sites that both have an accessible “seed region” and, unexpectedly, are structurally accessible in a newly identified downstream flanking sequence. We then used these simple rules to create a new set of shRNAs and achieved inhibition of HIV-1 production of 90% or greater for up to 82% of designed shRNAs. These shRNAs inhibit HIV-1 replication in therapy-relevant T-cells and show no or low cytotoxicity. The remarkable success of this straightforward SHAPE-based approach emphasizes that RNAi is governed, in significant part, by very simple, predictable rules reflecting the underlying RNA structure and illustrates principles likely to prove broadly useful in understanding transcriptome-scale biological recognition and therapeutics involving RNA.

1.6 Acknowledgements

Work described in this chapter was supported by National Institutes of Health grant AI068462 (to K.M.W.), National Research Service Award F30DA027364 (to J.T.L.), and Medical Scientist Training Program T32GM008719. Work in our laboratory on experimentally-directed RNA secondary structure prediction benefits from a close and lively collaboration with David Mathews (University of Rochester). We thank David Mauger and David Mathews for critically reviewing this manuscript.
1.7 References


dependence of thermodynamic parameters improves prediction of RNA

suitability of free-energy minimization using nearest-neighbor energy

counter-free grammars for RNA secondary structure prediction. *BMC
Bioinformatics*, **5**, 71.


31. Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.P. and Ehresmann, B.
9128.


analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and

structures in HIV-1 genomic RNA strongly conserved across distinct biological

Strong correlation between SHAPE chemistry and the generalized NMR order


CHAPTER 2

STRUCTURE OF THE HIV-1 FRAMESHIFT ELEMENT, INVESTIGATED BY LNA BINDING AND SHAPE PROBING

2.2 Introduction

HIV-1 viral enzymes, including protease, reverse transcriptase, and integrase, are generated by cleavage of the precursor polyprotein Gag-Pol. The pol gene cannot be expressed independently as it lacks a translation start site such as a start codon or an internal ribosomal entry site (IRES). Instead, Pol is translated as a fusion product with the Gag polyprotein that lies immediately upstream in the full length HIV-1 mRNA. While these two polyproteins are translated together, the pol gene is encoded in a reading frame that is offset from the upstream gag reading frame by one nucleotide in the 5′ direction. Consequently, translation of the Gag-Pol fusion protein relies on a process termed programmed ribosomal frameshifting that involves standard translation of the upstream Gag polyprotein followed by a recoding event that shifts the ribosome from the gag to the pol reading frame (1).

Frameshifting occurs at a specific site at the gag-pol junction that requires two key elements: (i) a UUUUUUA sequence that is strongly conserved across HIV strains (2) termed the slippery sequence at which the switch in reading frame occurs, and (ii) a downstream structural element termed the frameshift stimulatory stem. This downstream structure is thought to function by pausing the ribosome while the A and P
sites are occupied by the slippery sequence. While the precise mechanism remains incompletely understood, this arrangement allows ‘slipping’ of the ribosome by one nucleotide in the 5’ direction. Frameshifting occurs with a frequency of approximately 5-10% in cultured HIV-transfected human cells and the Gag to Gag-Pol ratio appears to be important for viral fitness (3). The frameshifting process has consequently attracted interest as a target for potential therapeutic development (4,5).

The frameshift stimulatory stem was originally proposed to comprise a single stem-loop (6). However, a number of refinements and extensions of this model have subsequently been proposed, including pseudoknots (7,8), a triplex (9), and a two-helix model (10). NMR studies performed on 41 and 45 nt transcripts have subsequently supported the formation of the two-helix model and this model is therefore now generally accepted as correct (11,12). The two-helix model includes the originally proposed stem, now termed the upper stem, and adds an additional lower stem that is separated from the upper stem by a three-purine bulge (Fig. 2.1). The functional importance of this lower stem is further supported by experiments demonstrating decreased frameshifting when the lower stem is destabilized by mutation (10) or when a truncated construct containing only the classical upper stem is used (13).

An alternative, more complex model was recently suggested based on SHAPE chemical probing experiments performed on an entire HIV-1 genome (14). SHAPE reagents react with RNA in a structure-selective manner. Flexible, single-stranded nucleotides tend to be reactive towards SHAPE reagents, whereas nucleotides in base-paired or otherwise constrained conformations tend to be unreactive. This structure-selective reactivity pattern provides nucleotide-resolution information about RNA
Figure 2.1 Conventional (10) and SHAPE-directed (14) models of the frameshift domain. Nucleotides are colored by *ex vivo* SHAPE reactivities and are numbered according to the NL4-3 genome.
secondary structure and is not limited by the size of the RNA under study (15). SHAPE reactivities can be incorporated into the thermodynamics-based computer folding algorithm RNAstructure (16) to obtain highly accurate RNA secondary structure models (17).

The SHAPE-directed model of the frameshift domain includes a total of four main helices (Fig. 2.1). One of these helices is equivalent to the upper stem of the conventional two-helix model. An additional helix involves refolding of the lower stem of the conventional model. We term these alternative base-pairings the alternate lower stem. SHAPE data also support the formation of two additional helices that fall outside the domain traditionally identified as the frameshift element. These include a helix that sequesters most of the slippery sequence in base-pairing interactions, and a 10-nt anchoring helix that completes the 140 nt domain (Fig. 2.1).

Given the critical role RNA structure plays in regulating frameshifting, we sought to confirm the existence of helices proposed by SHAPE and to ask whether the SHAPE-directed model could be reconciled with the conventional two-helix model. Our strategy takes advantage of the high binding affinities of locked nucleic acid (LNA)-containing oligonucleotides to RNA (18,19), in order to selectively bind and thereby disrupt specific helices in the SHAPE-directed frameshift model. We monitored the resulting structural changes by SHAPE. If the targeted helices were present, we expected that binding of LNAs to one strand of the helix would displace nucleotides on the partner strand. Released from their native base-pairing interactions, we anticipated that the resulting SHAPE reactivities of the displaced partner strand would increase.
Throughout this study we used full length genomic RNA extracted from authentic HIV-1 virions, allowing us to probe frameshift domain structures in their complete sequence context. Our results demonstrated SHAPE reactivity changes induced by LNA binding that strongly support the SHAPE-directed frameshift model. Unexpectedly, these experiments also revealed the ability of this domain to adopt the conventional two-helix model upon disruption of specific helices in the domain. Interpreting LNA binding and subsequent disruption as a rough approximation of ribosomal unwinding of the frameshift domain, we speculate that the different structural states induced by helix unwinding could have biological functions during the frameshift process. This interpretation would preserve a role for the conventional two-helix model as the frameshift domain is unwound during translation.

We also investigated the relative stability and conservation of the frameshift domain helices between two distinct biological states by performing SHAPE in the presence of formamide denaturant and in the context of the packaged RNA inside virion particles. We found that the frameshift domain is organized into highly structured, invariant helices with intervening sequences capable of forming varying degrees of structure in different environments. These results provide an experimentally-supported framework for further investigation of the structural dynamics and function of this important regulatory domain.
2.2 Results

2.2.1 Both natively extracted ex virio and heat-denatured RNAs adopt similar structures in the frameshift region.

We initially attempted to bind LNAs to sequences in folded full-length genomic RNA extracted from HIV-1 virions. However, addition of LNAs directed against the anchoring helix failed to produce detectable SHAPE reactivity increases at the partner strand. Two possible explanations for this initial result are: (i) the preformed, targeted helix was too stable to be disrupted by LNA binding, or (ii) inaccuracies in our model meant that the proposed helix was actually not present. To determine if a highly stable helix was the cause, we introduced a heating step where we first heated the HIV-1 RNA at 95°C for 5 min in the presence of LNA, snap cooled on ice, and incubated in a folding buffer. The goal of this heating step was to denature helices and allow the LNA to out-compete native base pair partners. We found that SHAPE reactivities on the partner strand did increase upon addition of LNA using this procedure. This demonstrates that anchoring helix stability, rather than an incorrect model, was the source of our initial inability to detect SHAPE reactivity changes (Section 2.2.2).

To properly interpret SHAPE reactivity changes following heat denaturation and LNA binding in the frameshift element region we must assume that RNA folding following heat denaturation results in a native-like conformation. To test this assumption, we performed SHAPE on both extracted, natively folded ex virio RNA and this same RNA following heat denaturation and refolding. We found that SHAPE reactivities for the genome region near the frameshift element are highly similar in both ex virio and refolded RNAs (Fig. 2.2A). A comparison of reactivities from a larger ~700
Figure 2.2 Frameshift domain reactivities of extracted *ex vivo* RNA and refolded heat denatured RNA using three different SHAPE reagents.
Figure 2.3 Standard SHAPE reactivities (1M7) of extracted *ex vivo* RNA and refolded heat denatured RNA for a region spanning ~700 nts surrounding the frameshift domain.
nt region encompassing the frameshift element reveals more significant reactivity differences between *ex vivo* and refolded RNA (*r* = 0.75), but the overall profile is largely the same (Fig. 2.3).

To further probe the differences between *ex vivo* and refolded RNA, we performed SHAPE reactions on each RNA state using two alternative SHAPE reagents, 1M6 and NMIA. While these electrophiles are similar to the currently used standard SHAPE reagent 1M7 (20), these reagents exhibit subtle reactivity differences that are sensitive to different local structural states and timescales. In particular, the 1M6 reagent has an enhanced ability to engage in π-π stacking interactions and displays increased reactivity towards nucleotides where one side of the nucleobase stacks with a neighboring nucleotide and the other side is available for stacking with the reagent (21). The NMIA reagent reacts on a slower time scale than 1M6 and 1M7 and can therefore detect nucleotides undergoing slow conformational dynamics (22). These reactivity differences can be used to provide a structural fingerprint of an RNA, and we can compare the two fingerprints to gauge their similarity. Reactivity profiles for all three reagents are similar at the frameshift domain (Fig. 2.2), although at the nucleotide level, some variability is apparent. Pairwise linear correlation coefficients for the 140 nt SHAPE-directed *ex vivo* frameshift domain are 0.80 between 1M7 and 1M6, 0.91 between 1M7 and NMIA, and 0.73 between 1M6 and NMIA. In contrast, reactivity profiles between *ex vivo* and refolded RNAs for each reagent are much more similar, with linear correlation coefficients of 0.94, 0.94, and 0.96 for 1M7, 1M6, and NMIA, respectively. This high similarity between SHAPE reactivities of *ex vivo* and refolded RNA as reported by three different reagents strongly suggests that the structures of
<table>
<thead>
<tr>
<th>LNA</th>
<th>LNA sequence</th>
<th>Target sequence</th>
<th>Region bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAATCTTTTC</td>
<td>GAAAGAUUG</td>
<td>1604-1612</td>
</tr>
<tr>
<td>2</td>
<td>CTAAAAAATT</td>
<td>AAUUUUUUUG</td>
<td>1629-1638</td>
</tr>
<tr>
<td>3</td>
<td>GTCTTTCTCT</td>
<td>AGGAAGGAC</td>
<td>1588-1596</td>
</tr>
<tr>
<td>4</td>
<td>TCTGAAGAA</td>
<td>UUCUUCAGA</td>
<td>1678-1686</td>
</tr>
<tr>
<td>5</td>
<td>TCCACACGC</td>
<td>GCUGUUGGA</td>
<td>1570-1578</td>
</tr>
<tr>
<td>6</td>
<td>GCTGTTGGC</td>
<td>GCCAACAGC</td>
<td>1697-1705</td>
</tr>
</tbody>
</table>

**Table 2.1** LNA-containing DNA oligonucleotide sequences and target RNA binding sites. LNA nucleotides are underlined.
these two states are extremely similar. We conclude that the refolded frameshift domain adopts essentially the same conformation as the more biologically relevant ex vitro state. This allows us to use heat denaturation to promote LNA disruption of stable helices while maintaining a native-like fold.

2.2.2 LNA binding experiments support the SHAPE-directed frameshift model

We designed 9- and 10-nt LNA oligonucleotides to bind to and disrupt each of the 3 helices unique to the SHAPE-directed frameshift model (Fig. 2.1 and Table 2.1). We designed one LNA for each helix strand for a total of two LNAs per helix. If the targeted helix existed and if LNA binding was able to out-compete native helical base-pairing, we expected to find increased SHAPE reactivity at the strand complementary to the LNA-bound strand.

The results of the six LNA binding experiments targeting the three helices unique to the SHAPE-directed frameshift model are shown in Fig. 2.4. The partner strand expected to be released from base-pairing by LNA binding is highlighted in grey. In all six cases, SHAPE reactivity increases in at least some nucleotides of the partner strand. These results strongly support the formation of these three helices in the full-length HIV RNA.

2.2.3 LNA binding can switch the SHAPE-directed alternate lower stem to the conventional lower stem.

Although a general increase in reactivity always occurs at the partner strand upon LNA binding, most LNAs also induce additional SHAPE reactivity changes in other regions of the frameshift domain. For example, while LNA 3 is designed to disrupt the
Figure 2.4 (A) and (B) Frameshift region SHAPE reactivities and predicted secondary structure models for HIV-1 RNA bound to LNAs targeting the slippery sequence helix.
Figure 2.4 (C) and (D) Frameshift region SHAPE reactivities and predicted secondary structure models for HIV-1 RNA bound to LNAs targeting the SHAPE-directed alternate lower stem.
Figure 2.4 (E) and (F) Frameshift region SHAPE reactivities and predicted secondary structure models for HIV-1 RNA bound to LNAs targeting the anchoring helix.
alternate lower stem and causes reactivity increases at the target’s partner, these increases are relatively minor and confined to nucleotides 1681-1683 (Fig. 2.4C, grey box). In contrast, more extreme changes in reactivity are present in other regions (Fig. 2.4C, red and yellow boxes). Nucleotides 1641-1643, highlighted in red, exhibit reduced reactivity relative to the no-LNA control. These nucleotides are predicted to be single stranded in the SHAPE-directed frameshift model, and the reduction in SHAPE reactivity upon LNA binding suggests the formation of new base-pairing interactions. Indeed, when the LNA 3-bound SHAPE data are used to direct folding of this domain, these nucleotides are predicted to form novel base-pairing interactions (Fig. 2.4C, purple lines), which largely correspond to the lower stem in the conventional two-helix model. Thus binding of LNA 3 switches the SHAPE-directed alternate lower stem to the conventional lower stem. Formation of this helix furthermore provides an explanation for why SHAPE reactivity increases are noted for only some of the partner nucleotides upon LNA 3 binding: many of these partner nucleotides refold to form part of the conventional lower stem.

Disruption of either strand of the anchoring helix by LNA 5 and LNA 6 results in SHAPE reactivity increases on the corresponding partner strand (Fig. 2.4 E, F). However, additional unexpected changes also occur. These changes closely resemble the changes seen upon LNA 3 binding to the alternate lower stem. SHAPE reactivities decrease for nucleotides 1641-1643 and increase for nucleotides 1590-1595, consistent with a transition from the alternate lower stem to the conventional lower stem. These results suggest that the anchoring helix stabilizes the alternate lower stem.
Disruption of the slippery sequence helix by LNA 1 binding also results in SHAPE reactivity changes that suggest a switch from the alternate to the conventional lower stem (Fig. 2.4A, red boxes). However, binding to the complementary side of the helix by LNA 2 does not result in a switch (Fig. 2.4B). LNA 2 targets nucleotides in close proximity to, and possibly overlapping with, the conventional lower stem. This partial steric occlusion could disfavor formation of the conventional lower stem.

2.2.4 Formation of the conventional lower stem destabilizes the slippery sequence helix.

Binding by LNAs 3, 5 and 6 all result in a switch from the alternate lower stem to the conventional lower stem (Fig. 2.4 C, E, F, red boxes) and may represent an interesting biological function. Binding of these LNAs also result in increased SHAPE reactivity at the slippery sequence helix (Fig. 2.4 C, E, F, yellow boxes). Remarkably, these changes occur upon disruption of the relatively distant anchoring helix by LNAs 5 and 6. In contrast, LNA 4, which destabilizes the alternate lower stem but does not induce formation of the conventional lower stem, does not cause significant SHAPE reactivity changes in the slippery sequence helix (Fig. 2.4D). These results suggest the alternate lower stem and slippery sequence stabilize each other.

2.2.5 Formamide denaturation experiments reveal the relative stabilities of the four frameshift structure helices.

We next examined the relative stabilities of the four main helices present in the SHAPE-directed frameshift element. Our strategy was to incubate HIV-1 RNA in a formamide-containing buffer. Due to the denaturing effects of formamide on nucleic acid (23), we expected to find increased reactivity towards 1M7 as the formamide concentration increased. Less stable helices should denature at lower formamide
Figure 2.5 (A) SHAPE reactivity profiles for formamide-denatured RNA (red and green) compared with ex virio reactivities (blue) and (B) average SHAPE reactivities for each helical strand as a function of formamide concentration. Helices defined by both the SHAPE-directed and conventional models are shown.
concentrations than highly stable helices and this transition from base-paired to single-stranded nucleotides can be detected by SHAPE reactivity.

Following a 20-minute incubation at 37°C in folding buffer, we added formamide and incubated the solution for a further 20 min before initiating a SHAPE reaction with 1M7. We varied the formamide concentrations from 0% to 60%, in increments of 10%. The resulting SHAPE reactivities show a general increase with formamide concentration, shown in Fig. 2.5A for 0%, 30% and 60% formamide.

Average SHAPE reactivities as a function of formamide concentration are plotted in Fig. 2.5B for all helix strands defined by both the SHAPE-directed and conventional two-helix models. Nucleotides on one side of the slippery sequence helix (1604-1611) have low reactivity in 10% formamide, but show markedly increased reactivity at formamide concentrations of 20% and above. However, only modest reactivity changes are apparent at high formamide concentration on the partner strand (1629-1636) (Fig. 2.5A, yellow). This demonstrates that the slippery sequence helix unfolds at 20% formamide, although one partner strand is able to form stabilizing interactions even in the presence of high denaturant concentration.

Nucleotides 1590-1595 are involved in the alternate lower stem and are unreactive in up to 10% formamide, but are very reactive at 20% formamide and greater. Partner nucleotides 1678-1683, which can pair with either the alternate or conventional lower stem, remain lowly reactive in up to 30% formamide, but then increase modestly in reactivity at higher formamide concentrations. Nucleotides in the conventional lower stem exhibit little change in reactivity in up to 30% formamide but increase after that. These data suggest that the alternate lower stem unfolds in 20%
Figure 2.6 Secondary structure models for formamide-denatured RNA.
formamide. At this concentration of denaturant, both its partner nucleotides and nucleotides involved in the conventional lower stem exhibit lower reactivities, and secondary structure modeling of the frameshift domain suggests the formation of the conventional lower stem (Fig. 2.6B). Upon further increases in formamide concentration, the conventional lower stem also unfolds (Fig. 2.6C). Nearest neighbor thermodynamic calculations indicate that the conventional lower stem is slightly more stable than the SHAPE-directed alternate lower stem (-8.4 kcal/mol and -8.2 kcal/mol, respectively). In light of both these results and LNA disruption results, it appears that formation of the alternate lower stem is largely dependent on stabilization from other elements in the domain and that as these interactions are disrupted by formamide, the slightly more stable conventional lower stem is able to form.

The anchoring helix and upper stem show very low SHAPE reactivities up to 40% formamide. Unfolding of the anchoring helix is first apparent in 50% formamide, although the upper stem remains unreactive even in 60% formamide (Fig. 2.6D). This implies that these two helices are highly stable and exist even under denaturing conditions. The high stability of the upper stem has been previously observed (24) and is thought to be attributable both to the high GC content of the helix and the stabilizing atypical ACAA tetraloop that caps the helix. We conclude from these denaturant experiments that the frameshift domain is bounded by two highly stable helices and that the intervening less stable helices have the propensity to both unfold and, in the case of the lower stem, refold into two alternative conformations.
Figure 2.7 (A) SHAPE reactivity profiles and (B) predicted secondary structure models for in virio compared with ex virio RNA.
2.2.6 SHAPE probing of *in vitro* genomic RNA reveals a less structured frameshift domain.

We next probed the frameshift region structure of HIV-1 RNA packaged inside authentic virions. Structural interrogation of this *in vitro* state, which includes all proteins packaged with the RNA in the virion, is possible because SHAPE reagents readily cross biological membranes (25). Our results (Fig. 2.7) show that the frameshift domain adopts a much less structured conformation *in vitro* as compared to the *ex vitro* and refolded states. In particular, nucleotides involved in the slippery sequence helix and lower stem, including both the conventional and alternate base-pairings, display higher SHAPE reactivities. In contrast, the anchoring helix and upper stem are still present and appear to be the only significant structural elements within the frameshift domain.

2.3 Discussion

In this study, we use LNA binding monitored by SHAPE to provide the first independent support for the three helices unique to the frameshift domain model proposed by SHAPE-directed folding of an entire HIV-1 genome (14). Our model shares a stable upper stem with the currently accepted two-helix model but has two primary modifications. First, this study supports alternative pairing partners for the conventional lower stem. Second, the SHAPE-directed model incorporates the standard frameshift region into a larger 140 nt domain. This larger domain includes base-pairing interactions involving the slippery sequence and an additional stable anchoring helix.
A key distinguishing feature between the present study and prior research on the frameshift structure is our use of full-length genomic RNA. This allowed us to structurally interrogate the frameshift element in its complete sequence context. We are thereby able to avoid biases created when a truncated sequence of an RNA is studied. In fact, most past studies of the HIV-1 frameshift, including the NMR verifications of the conventional model, used HIV-1 sequences that were shorter than the 140 nt domain. All of the three helices unique to the SHAPE-directed model are unable to form in these truncated sequences. Furthermore, two of these helices, the alternate lower stem and the frameshift sequence helix, appear to stabilize each other as disruption of one helix by LNA binding can destabilize or cause refolding of the other. Additionally, formation of the stable anchoring helix is required for formation of both these helices.

Interestingly, the frameshift domain adopts the same structure whether the genomic RNA is probed directly after extraction from virions or first heat-denatured and then refolded. This suggests that the differences between the SHAPE-directed model and the conventional model are due primarily to the presence of complete sequence context in the SHAPE-directed model rather than to the native folding pathway which, in general, likely influences folding of long RNAs (26).

While our LNA binding experiments strongly support our 140 nt frameshift domain, our formamide denaturation experiments also indicate that this domain contains both highly stable, likely static structures, as well as less stable helices with the flexibility to form alternative structures. The anchoring helix and upper stem are highly stable and exist even in the presence of high amounts of formamide denaturant.
Their stability even appears to hold across diverse biological states, as these two helices are the only frameshift domain helices that are present inside the packaged virion. Between these two elements, a greater diversity of structural states is possible upon perturbation of the folded domain through either LNA binding or formamide denaturation.

The present study examines the structure and thermodynamics of the frameshift domain but does not directly address the functional implications of the model. Nevertheless, we can use the structural changes induced by LNA binding to provide a rough approximation of frameshift domain structural dynamics as it is unwound and translated by the ribosome (Fig. 2.8). The anchoring helix is the first frameshift domain secondary structural element to be unwound by the translating ribosome (Fig. 2.8B). This causes major structural rearrangement throughout the frameshift domain. The lower stem switches from the SHAPE-directed alternate to the conventional conformation, and the slippery sequence helix is significantly destabilized, potentially releasing slippery sequence nucleotides from their basepairing interactions. It is unclear what the functional implications of this structural rearrangement may be, but we note that the importance of conformational switching has recently been demonstrated in ribosomal frameshifting in related retroviruses (27). Additionally, it seems clear from prior studies that the conventional lower stem is functionally important for frameshifting (10,13).

After unwinding the anchoring helix, the ribosome continues over the slippery sequence and encounters the conventional lower stem. LNA 2 binding data suggest that unwinding of the conventional lower stem results in a switch back to the alternate
Figure 2.8 Proposed model for frameshift domain unwinding during translation. A PI-induced C-to-U mutation (red) creates a potential secondary slippery sequence.
lower stem (Fig. 2.8C). The ribosome must then unwind the very stable upper stem and switch reading frames. At this point, all frameshift domain helices will have been unwound. However, the high stability of the anchoring helix raises the possibility of its renaturation after the ribosome has translated through it. The ribosome would then encounter the anchoring helix for a second time before exiting the frameshift domain (Fig. 2.8D).

Intriguingly, a UUUUCUU sequence (nucleotides 1676-1682) lies just upstream from this potential second encounter with the anchoring helix (Fig. 2.8D). In viruses resistant to protease inhibitors (PI), this C1680 is frequently mutated to U, resulting in a UUUUUUUU (Fig. 2.8D, red C → U) sequence that resembles the standard UUUUUUA slippery sequence at nucleotides 1631-1637 (28). This results in a cleavage site Leu to Phe mutation at the protein level which may increase cleavage efficiency and enhance the generation of functional protease. Additionally, Doyon et al. (29) hypothesized that this site could also act as a secondary slippery sequence that increases the overall amount of frameshifting, thereby increasing the relative amount of protease to compensate for reduced PI-induced protease activity. When the standard slippery sequence was inactivated by mutation, the mutant secondary slippery site, but not the wild-type, was able to stimulate frameshifting in in vitro translation assays. Furthermore, Doyon et al. (29) found evidence for frameshifting from this secondary site in virus-expressing cells. While these in vitro results have subsequently been confirmed independently, no evidence for frameshifting from this secondary site in cell culture was detected (30). However, these experiments used HIV sequences that were shorter than the 140 nt domain described here and therefore could not form the
anchoring helix. In future, it would be interesting to ask if, in the context of the complete domain, frameshifting can be enhanced by the PI-induced secondary slippery sequence with the anchoring helix acting as a secondary frameshift stimulatory stem.

In conclusion, this study confirms key structure elements specific to the SHAPE-directed frameshift model. We also demonstrate that this model has the ability to switch between the SHAPE-directed model and the conventionally accepted two-helix model. We hypothesize that this switch likely occurs as the ribosome unwinds the frameshift element, although the functional significance of this switch will need to be assessed in future work. This work underscores the influence of global sequence context on this important regulatory domain and implies that, at a minimum, all 140 nucleotides of the domain should be used when studying frameshifting in HIV-1.

2.4 Methods

2.4.1 HIV-1 virion production

HIV-1 virion particles were prepared as described previously (14,25). Briefly, HIV-1 strain NL4-3 (group M, subtype B) was used to infect a non-Hodgkin’s T cell lymphoma cell line (31). Virions were purified by subtilisin digestion and centrifugation through a 20% (w/v) sucrose cushion as previously described (32).

2.4.2 Extraction of RNA genomes from virions

Subtilisin-treated virions were lysed by incubation in a virion lysis buffer (50 mM Tris pH 7.5, 10 mM EDTA, 1% SDS, 100 mM NaCl, 10 mM DTT, 15 μL 20 mg/mL Proteinase K) for 30 min at room temperature. The digest was extracted four times with phenol/chloroform/isoamyl alcohol, followed by four extractions with pure
chloroform. The aqueous layer was brought to a NaCl concentration of 300 mM, precipitated in 70% ethanol, and stored at -20°C. To extract the RNA, we pelleted this solution by centrifugation at 14,000 rpm for 30 min, eluted off the ethanol layer, and vacuum dried the pellet for 1 min. We then resuspended the RNA pellet in a storage buffer (200 mM KOAc, 50 mM HEPES pH 8) for a final RNA concentration of about 400 nM. These aliquots were flash frozen in liquid N₂ and stored at -80°C.

2.4.3 Folding of ex virio RNA.

For each reaction, 1 µL 400 nM ex virio RNA was resuspended in a Mg²⁺-containing standard folding buffer (200 mM KOAc, 50 mM HEPES pH 8, 3 mM MgCl₂) to promote the stabilization of native-like interactions (total volume was 20 µL per reaction). We incubated this mixture at 37°C for 30 min and then proceeded immediately to the SHAPE modification step (Section 2.4.8).

2.4.4 Formamide denaturation

For the formamide denaturation experiments, 1 µL 400 nM ex virio RNA was resuspended in 5 µL folding buffer and incubated at 37°C for 20 min. A formamide-containing folding buffer (67% v/v deionized formamide, 200 mM KOAc, 50 mM HEPES pH 8, 3 mM MgCl₂) was then combined with the appropriate amount of standard folding buffer and added to the reaction mix to obtain the desired formamide concentration in a reaction volume of 20 µL. We performed these experiments in final formamide concentrations of 0% to 60%, in increments of 10%. We incubated this formamide-containing mixture for an additional 20 min at 37°C and then proceeded immediately to the SHAPE modification step (Section 2.4.8).
2.4.5 LNA oligonucleotide design

We used 9 and 10 nt LNA oligonucleotides (Exiqon) that were the reverse complements of HIV-1 sequences comprising helices proposed in the SHAPE-directed frameshift model. The specific sequences were chosen to avoid self-complementary sequences due to very strong LNA-LNA binding affinities. We also tried to avoid stretches of 3 or more Gs or Cs. These considerations resulted in LNA oligonucleotides that, in the case of the 10-nt anchoring helix bound most of the targeted strand. Additionally, when possible we included flanking single stranded regions as part of the targeted strand to facilitate LNA binding to structured helices. LNA oligonucleotide sequences and binding sites are shown in Table 2.1 and Fig. 2.1

2.4.6 LNA binding to genomic RNA

For each LNA binding experiment, 1 μL 400 nM extracted HIV-1 genomic RNA in storage buffer (200 mM KOAc, 50 mM HEPES pH 8), 2 μL 2 μM LNA, 4.6 μL TE, and 1.7 μL water were mixed and heated at 95°C for 5 min, and then snap cooled on ice. We then added 9.5 μL 2X storage buffer and 1.2 μL 50 mM MgCl₂ (20 μL total in folding buffer conditions) and incubated this mixture for 30 min at 37°C before proceeding immediately to the SHAPE modification step.

2.4.7 SHAPE modification of ex virio, LNA-bound, and formamide denatured RNA

The SHAPE modification was performed by adding 9 μL each of the folded RNA mixture to both 1 μL 40 mM SHAPE reagent (1M7 for standard SHAPE reactions and 1M6 and NMIA for differential SHAPE reactions) and 1 μL neat DMSO as a no-reagent control. The reaction mix was incubated for 4 min at 37°C. We then added 1 μL 50 mM
EDTA and performed a clean-up step to remove LNA oligonucleotides using the RNeasy Mini Kit (Qiagen) according to the manufacturer’s instructions. The resulting mixture (100 μL) was ethanol precipitated by the addition of 10 μL 2 M NaCl and 300 μL ethanol and was stored at -80°C until the primer extension step.

2.4.8 Primer extension and capillary electrophoresis detection of SHAPE adduct sites

SHAPE-modified RNA pellets were resuspended in 7 μL 0.5 X TE and 6 μL 0.4 μM VIC-labeled DNA primer (Applied Biosciences). Primer extension reactions for LNA binding experiments were performed using primers complementary to nucleotides 1750-1771 (primer 6.2, with sequence ATCGGCTCCTGCTTGAGAGG) and 2033-2054 (primer 7, with sequence CAATTATGGACAGGTAGG). We used primer 6.2 for formamide denaturation experiments and primer 7 for in virio experiments. In addition to primers 6.2 and 7, primers complementary to nucleotides 1522-1543 (primer 5.5, with sequence TTGGCTATGTGCCCTCTTTGC) were also used for the ~700 nt comparison between ex virio and refolded RNA. To promote primer binding, this mixture was heated at 65°C for 5 min, followed by 42°C for 2 min, and then cooled on ice. We then added 7 μL primer extension reaction mixture (4 μL 5X first strand buffer, 1 μL 0.1 M DTT, 1 μL 10 mM each dNTP, 1 μL 100 U Superscript III reverse transcriptase, Invitrogen). This mixture was incubated at 45°C for 10 seconds, 52°C for 20 min, 65°C for 5 min, and then cooled at 4°C. The resulting cDNA products were precipitated by addition of 60 μL ethanol and stored at -80°C for at least 30 min. We then centrifuged this ethanol solution at 14,000 rpm for 45 min, eluted off the ethanol layer, washed in 70% ethanol, pelleted, and then resuspended in 9 μL formamide. The cDNA solution was heated at 95°C for 3 min to aid in dissolving the pellet. 1 μL of a
NED-labeled dideoxycytosine sequencing ladder stock was added. Dideoxy sequencing reactions (GenomeLab Methods Development Kit; Beckman) were performed using plasmid pNL43 and NED-labeled primers. An Applied Biosystems 3500 capillary electrophoresis instrument was used to quantify the cDNA products.

2.4.9 Data processing

Raw capillary electropherograms were processed using the custom QuShape software as described in (33). Briefly, key processing steps include a mobility shift to correct for small differences in the electrophoretic mobility between the NED and VIC fluorescent dyes, and a signal decay correction to account for signal attenuation as distance from the reverse transcriptase primer binding site increases. Nucleotide positions are assigned to reagent and DMSO control traces by aligning the dideoxy sequencing ladder with the nucleotide sequence using the sequence alignment tool. Peaks in the reagent and DMSO control traces are integrated and scaled relative to each other such that the lowest peaks in the reagent trace, which correspond to lowly reactive nucleotides, are of similar magnitude to their corresponding DMSO control peaks. DMSO control peaks are subtracted from reagent peaks and the resulting SHAPE reactivities are normalized on a scale where a normalized reactivity of 1.0 is defined as the average intensity of the top 10% most reactive peaks, excluding a few highly reactive nucleotides taken to be outliers. The resulting reactivities span a scale from 0 to ~1.5, where 0 indicates no reactivity (and a highly constrained nucleotide) and reactivities >0.7 typically indicate highly flexible nucleotides.
2.4.10 Secondary structure modeling

SHAPE-directed models for the frameshift domain were created by incorporating SHAPE data into the RNAstructure folding algorithm (16,17). SHAPE data are added as experimental corrections to the nearest neighbor energy function (34) as follows:

\[ \Delta G_{SHAPE} (i) = m \ln(\text{SHAPE reactivity}(i) + 1) + b \]  \hspace{1cm} (2.1)

These pseudo-free energy terms are added for each base-paired nucleotide \( i \) and are weighted by the parameters \( m \) and \( b \). The parameter \( m \) weights a penalty for pairing nucleotides with high SHAPE reactivity, while the parameter \( b \), a negative value, accounts for the energetic bonus of pairing lowly reactive nucleotides. The choice of parameter can depend on the RNA. For HIV, we use values of \( m = 3.0 \) kcal/mol and \( b = -0.6 \) kcal/mol. The value for \( m \) is higher than values typically used for folding more structured RNAs (1.9-2.6) and is used to counteract the tendency for over-prediction of base-pairs by folding algorithms. To account for LNA binding, the LNA target site was made single stranded by imposing artificially high SHAPE reactivity values of 100. For LNA 5 and 6, which target the highly stable anchoring helix, base-pairing was also prohibited at partner nucleotides as their reactivity increases were taken to imply a single stranded state.

2.5 Acknowledgements

We are indebted to Howard Fried for suggesting and advising us on the formamide experiments. We thank Nathan A. Siegfried and Joshua S. Martin for helpful discussions on the LNA binding experiments. This work was supported by National Institutes of Health grant AI068462 (to K.M.W.). J.T.L. was supported by a National
Research Service Award (F30DA027364), the Medical Scientist Training Program (T32GM008719), and the Program in Molecular and Cellular Biophysics (T32GM008570).
2.6 References


CHAPTER 3
SHAPE-DIRECTED DISCOVERY OF POTENT SHRNA INHIBITORS OF HIV-1

3.1 Introduction

Interactions between RNAs and molecular ligands, proteins, and other RNAs govern numerous cellular regulatory processes. These targeting and binding events are strongly influenced by the structure of the target RNA (1). One biologically and clinically important example of intermolecular interactions between RNAs involves the RNA interference (RNAi) pathway (2). The RNAi pathway regulates gene expression in organisms from plants to humans (3) and immune defense via destruction of pathogenic RNAs (4). This pathway can be exploited to destroy pathogenic RNAs using synthetic short interfering RNA (siRNA) or short hairpin RNA (shRNA) expression vectors (5). The ~19 nucleotide guide strands generated from cellular transcripts or synthetic RNAs are loaded into the RNA-2 induced silencing complex (RISC). Base complementarity between the guide strand and target RNA signals the Argonaute protein in RISC to cleave the target RNA. Both siRNA and shRNA approaches have been used successfully to inhibit HIV-1 production and replication in cell culture (6-8).

The pathways for incorporation of siRNAs into RISC and their recognition of RNA targets are complex and not all sequences can be targeted efficiently. Most

---
randomly selected sequences are not efficiently repressed by siRNAs or shRNAs
directed against them (6, 9). Accurate identification of repressible sequences remains
an unmet challenge that has motivated the development of a wide variety of target
selection algorithms (9-22). Most algorithms were developed for selection of siRNAs
and are usually assumed to apply to shRNA design, although there is evidence that
siRNA and shRNA prediction require distinct, but overlapping, rules (23, 24).
Degradation of transfected siRNAs in vivo makes them primarily suitable for short-term
clinical applications, such as the treatment of acute infection. In contrast, for stable,
long-term suppression as likely required for chronic infections, including HIV-1, we are
especially interested in understanding the rules that govern shRNA-directed inhibition.

Reker Brake et al. (6) reported the ability of 84 shRNA constructs to inhibit HIV-1
production in cell culture. This set of inhibitors was attractive as a training dataset
because it was developed without using any si/shRNA selection rules. With a view
towards potential therapeutic applications, shRNAs were instead chosen based on
sequence conservation in the HIV-1 RNA. In addition, viral inhibition was assessed at
low, generally sub-saturating, shRNA transfection levels. As a result, viral inhibition
levels achieved by these shRNAs spanned a wide range of inhibition levels.

Correlations between virus production inhibition by shRNAs in this dataset (6)
and rankings given by previously described prediction algorithms (9, 12, 14-21) are
generally poor. This poor performance includes algorithms both that consider sequence
signatures in the target RNA sequence and that additionally incorporate
thermodynamic metrics to identify favorable interactions between an si/shRNA and its
target (Table 3.1).
| Algorithm                        | Reference | $| r |$  |
|---------------------------------|-----------|------|
| Approaches based primarily on sequence characteristics: |           |      |
| Amarzguioui                      | 17        | 0.33 |
| BIOPREDsi                        | 12        | 0.30 |
| Dharmacon                        | 9         | 0.15 |
| DSIR                             | 21        | 0.51 |
| i-Score                          | 20        | 0.30 |
| Katoh                            | 18        | 0.11 |
| Takasaki                         | 19        | 0.05 |
| Approaches that directly incorporate RNA structure metrics: |           |      |
| RNAxs                            | 14        | 0.28 |
| Sirna                            | 15,16     | 0.13 |
| Approaches based solely on RNA structure metrics, described in this study: |           |      |
| No constraints                   |           | 0.36 ($\Delta G_{	ext{target}}$) |
|                                  |           | 0.38 ($\Delta G_{	ext{total}}$) |
| With SHAPE                       |           | 0.72 ($\Delta G_{	ext{target}}$) |
|                                  |           | 0.61 ($\Delta G_{	ext{total}}$) |

**Table 3.1:** Correlation between HIV-1 inhibition and si/shRNA target prediction algorithms
We therefore sought to explore whether nucleotide-resolution information about an RNA target structure might make it possible to design potent shRNA inhibitors efficiently. We show that, if the underlying RNA structure is known with good accuracy, very simple thermodynamics-based rules yield excellent predictions for highly potent shRNA inhibitors of HIV-1 replication. Our prediction accuracy exceeds all current approaches when applied to the HIV-1 RNA genome, a notable result given that the rules developed in this work are much simpler than alternative approaches. In broad terms, this work shows how profoundly RNA structure influences biological function and emphasizes the importance of developing high-content models for an RNA fold to understand RNA-based therapeutics and biological mechanisms.

3.2 Results

3.2.1 Strategy

To derive new rules for designing shRNA inhibitors, we focused on one of the simplest possible models for the thermodynamics of the guide strand-HIV RNA target interaction. In this model (Fig. 3.1), the unstructured guide strand binds to a complementary region of the target RNA to form the guide strand-target duplex. We assumed that protein contacts with RISC maintained the guide strand in a single stranded conformation, poised to interact with the target RNA. This equilibrium is thus described by a free energy change, $\Delta G_{\text{duplex}}$, and is calculated using nearest neighbor thermodynamic rules (25, 26). $\Delta G_{\text{duplex}}$ depends only on the RNA sequence and is independent of RNA structure because the unstructured guide strand is assumed to interact with a fully unfolded site in the target RNA.
Figure 3.1: The guide strand-target RNA interaction equilibrium. RISC Argonaute protein is represented by a red oval, although only RNA-RNA thermodynamics were considered in this study.
In reality, the target RNA folds back on itself to form base paired secondary (and higher order) structures. Thus, a prerequisite for effective binding by RISC is that a site in the target RNA be unfolded to allow interaction with the guide strand. The strength of these pre-existing interactions is termed $\Delta G_{\text{target}}$ (Fig. 3.1). We consider a very simple model in which $\Delta G_{\text{duplex}}$ and $\Delta G_{\text{target}}$ dominate the equilibrium for forming a putative guide strand-target RNA interaction. The total free energy change $\Delta G_{\text{total}}$ of the interaction reflects a favorable contribution from duplex formation and the unfavorable cost of disrupting pre-existing structures in the target RNA:

$$\Delta G_{\text{total}} = \Delta G_{\text{duplex}} - \Delta G_{\text{target}}$$  \hspace{1cm} (3.1)

In contrast to $\Delta G_{\text{duplex}}$, which is independent of RNA structure, accurate calculation of the free energy change required to unfold the target RNA ($\Delta G_{\text{target}}$) depends critically on the RNA structure model. Conventional thermodynamics-based RNA secondary structure prediction algorithms typically attain accuracies of 50-70%; accuracies are at the lower end of this range as RNA length increases (26, 27). This level of accuracy provides a helpful overall glimpse of an RNA structure but, in the context of HIV-1 inhibition of shRNAs, is insufficient for consistent prediction of optimal guide strand binding sites (Table 3.1). Secondary structure predictions can be improved dramatically by incorporating additional experimental information. In particular, nucleotide-resolution measurements of molecular structure obtained from SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) chemical probing experiments (28, 29) can be incorporated as pseudo-free energy corrections into a thermodynamics-based RNA folding algorithm (30). The resulting RNA secondary structure models are generally highly accurate even for RNAs on the kilobase scale (27).
We therefore used a SHAPE-directed secondary structure model of an entire ~9 kb HIV-1 genome (31) to calculate Δ$G_{\text{target}}$.

### 3.2.2 Concentration dependence of shRNA inhibition.

Ideally, RNAi-mediated knockdown of a target gene should be accomplished using the lowest effective concentration of shRNA, for two reasons. First, the RNAi pathway contains a finite complement of protein mediators, which can become saturated by the expression of large amounts of exogenous siRNA. Such saturation is thought to be an important cause of off-target effects and of general cell toxicity (32, 33). Second, artificial production of large amounts of guide strand-RISC complexes can make it possible to target both optimal and sub-optimal sequences and thus obscure important differences in targetability.

We therefore sought to determine the minimum level of transfected shRNA-encoding plasmid that achieved inhibition of HIV-1 production in our cell-based assay (6). We titrated plasmids expressing five shRNA sequences over the range from 1 to 125 ng (Fig. 3.2). Individual shRNAs differ significantly in their ability to inhibit HIV-1 production and inhibition by a given shRNA increased as amount of transfected plasmid increased. All five shRNA sequences tested showed significantly stronger inhibition at 125 ng transfected plasmid versus 25 ng plasmid. In this work, we use 25 ng transfected plasmid because this amount showed a high level of knockdown for optimal sequences, was sub-saturating in our experimental system, and is therefore most appropriate for facilitating accurate identification of highly potent shRNAs.
Figure 3.2: Relative levels of virus production for five shRNAs. shRNA-encoding plasmids were transfected into 293T cells in amounts ranging from 1 to 125 ng per well. Each shRNA is labeled by the first nucleotide position of its binding site on the NL4-3 HIV-1 mRNA (see Appendix 1).
**Training set**

![Diagram](image)

**Test set**

**Figure 3.3:** HIV-1 genome locations of the 84 target sequences for the shRNAs in the ter Brake et al. dataset (6) used to derive design rules (top) and of the 26 sequences which matched the design rules defined in this work (bottom). Numbers indicate the position of the 5’ nucleotide of each shRNA target site in the HIV-1 NL4-3 mRNA.
3.2.3 Weak target folding energy characterizes effectively repressed sequences.

Local accessibility has been shown to be an important determinant for efficient si/shRNA-mediated knockdown of a target gene (14, 16, 34-37) (see $\Delta G_{\text{target}}$, Fig. 3.1). However, it is not obvious precisely which portion of the target RNA sequence needs to be accessible, or unfolded. An unfolding window is characterized by two parameters: its size and position. We calculated the free energy change, $\Delta G_{\text{target}}$, of base pair formation in the SHAPE-based HIV-1 RNA structure (31) as a function of window size and position. We then calculated the linear correlation coefficient, $r$, between these $\Delta G_{\text{target}}$ values and the experimental viral production inhibition values obtained by ter Brake et al. (6) for each of the 84 shRNAs in the training set (Fig. 3.3 and Appendix 1). The magnitudes of the correlation coefficients quantify the strength of the $\Delta G_{\text{target}}$ metric for predicting inhibition, whereas the window size and position parameters identify the target RNA window that should be unstructured for optimal shRNA-mediated inhibition (Fig. 3.4).

The strongest correlation was obtained for a 13-nt window that begins at position 14 of the target RNA (Fig. 3.4A, white box) and has an $r$-value of -0.72 (Fig. 3.5A). The negative correlation confirmed that more readily unfolded, or less structured, target windows were more readily silenced. The first 5-6 nucleotides of the position window overlapped closely with the "seed region", the guide strand sequence that interacts initially with the target RNA prior to complete unfolding of the target to form the $\sim$19-nt long duplex (38) (Fig. 3.1).

Unexpectedly, the 13-nt window also extended 7 nucleotides beyond the 3’ end of the target RNA binding site (Fig. 3.4B, grey box). This indicates that the most
**Figure 3.4:** Correlation coefficients ($r$) between calculated target folding energies, $\Delta G_{\text{target}}$, and experimental activity values for the 84 shRNAs in the training dataset. (A) Linear $r$-values as a function of target unfolding window size (vertical axis) and position (horizontal axis). Colors denote the relative strength of the correlation. A white box highlights the strongest correlation. (B) Optimal accessible target window. Window positions are numbered relative to the 5’ end of the guide strand binding site on the target RNA. The window size denotes the length of the window extending downstream of the window position value. The accessible target window that yielded the strongest correlation is shown by a grey rectangle.
SHAPE data

Figure 3.5: Relative viral production versus (A) target folding energies, $\Delta G_{\text{target}}$, and (B) total binding energies, $\Delta G_{\text{total}}$, for the 84 shRNAs in the training dataset. The optimal target unfolding window, identified in Fig. 3.4, was used for $\Delta G_{\text{target}}$ calculations. The SHAPE-directed HIV-1 secondary structure model (31) was used to calculate $\Delta G_{\text{target}}$. The same correlations for (C) $\Delta G_{\text{target}}$ and (D) $\Delta G_{\text{total}}$ calculated without using experimental constraints to estimate the secondary structure.
effectively targeted RNA sequences are characterized by a 13-nt unstructured window that includes the seed region binding site and a previously unrecognized requirement that extends ~7 additional nucleotides beyond the region directly bound by the guide strand in RISC.

3.2.4 Strong total binding energy characterizes effectively repressed sequences.

The formation of an approximately 19-nt duplex between guide strand and target is required for proper recognition and subsequent cleavage of the target by RISC Argonaute proteins. We estimated the strength of this binding as the overall binding free energy $\Delta G_{\text{total}}$ of the complete 19-nt interface. We then computed the correlation between $\Delta G_{\text{total}}$ and viral production inhibition for each of the shRNAs in the training set. Total binding energy correlated strongly with shRNA inhibition ($r = 0.61$) (Fig. 3.5B).

3.2.5 Strong thermodynamic correlations are specific to the SHAPE-directed RNA structure model.

Strong correlations between HIV-1 inhibition and the energetic cost of disrupting pre-existing structures in the viral RNA ($\Delta G_{\text{target}}$) and for overall strength of guide strand binding ($\Delta G_{\text{total}}$) were obtained when these free energies were calculated using the experimentally-supported model for an HIV-1 genomic RNA (31). This model was obtained using authentic genomic RNA gently extracted from virions and was used to represent structures at any targetable stage in the HIV-1 replication cycle. This SHAPE-directed model shares only ~45% of base pairs with a model obtained using the same nearest-neighbor parameters but without experimental data (data not shown).
Using the structure predicted without experimental SHAPE constraints, we obtained much weaker correlations between viral inhibition and any simple thermodynamic metric for guide strand interaction. The correlations between viral production values for the training set shRNAs and $\Delta G_{\text{target}}$ calculated for the 13-nt window or for $\Delta G_{\text{total}}$, which includes the guide strand-target duplex formation interaction, had modest $r$-values of -0.36 and +0.38, respectively (Figs. 3.5 C, D). These correlations, although poor, are actually at the higher end relative to current si/shRNA prediction algorithms (Table 3.1), even though these later approaches use more complex rules than the very simple thermodynamic approach outlined here (Fig. 3.1).

These results emphasize that target RNA secondary structure plays a profound role in RNAi activity, that current thermodynamics-only calculations do not recapitulate this contribution, and that SHAPE-directed secondary structure prediction provides significant additional information.

### 3.2.6 Experimental validation of shRNA design rules.

The strong correlations between inhibition of viral production and $\Delta G_{\text{target}}$ and $\Delta G_{\text{total}}$, calculated using the SHAPE-derived secondary structure for HIV-1 (Figs. 3.5 A, B), suggest two remarkably simple rules for shRNA design. First, the target RNA in the newly defined, optimal 13-nt window should have minimal pre-existing structure. Second, the total binding energy for the guide strand-target RNA interaction should be strong. We tested these two rules by designing a new, independent set of shRNA inhibitors and measured their ability to inhibit HIV-1 viral production. We calculated $\Delta G_{\text{target}}$ and $\Delta G_{\text{total}}$ for all possible 19-nt sequences in the 9,173 nt NL4-3 HIV-1 genome and initially required $\Delta G_{\text{target}}$ to be greater than or equal to 0 kcal/mol and $\Delta G_{\text{total}}$ to be
**Figure 3.6:** Prediction success rates of designed shRNAs. (A) Inhibition of HIV-1 production by shRNAs in the Test set. Target sequences were chosen to have accessible 13-nt windows ($\Delta G_{\text{target}} \geq 0$ kcal/mol) and strong overall binding energy $\Delta G_{\text{total}}$. Prediction success rates are shown for $\Delta G_{\text{total}}$ criteria of -25, -27 and -29 kcal/mol; $n$, the numbers of shRNAs meeting this threshold. (B) Inhibition of HIV-1 production by shRNAs in the 84-member Training set. (C) Prediction success rates for 16 shRNAs that failed the target accessibility criterion. A $\Delta G_{\text{target}} < 0$ corresponds to RNA target sites with some pre-existing structure.
less than -25 kcal/mol. Approximately 500 sequences (~5% of all possible 19-mers) satisfied these two rules. Of these, we tested 26 randomly-selected sequences (Fig. 3.3, bottom panel; and Appendix 2).

The percentages of shRNAs that achieved a defined level of HIV-1 inhibition were calculated for the initial training set and for the independently designed test sequences (Fig. 3.6A). Strikingly, 23 of the 26 shRNAs (88%) selected based on our rules reduced HIV-1 production by a factor of 2 or more. Moreover, this simple thermodynamics-based approach was especially successful at identifying highly potent inhibitors: 54% (14/26) of the designed shRNAs inhibited viral production by 90% or greater (see -25 kcal/mol data, Fig. 3.6A). Very simple thermodynamics-based selection rules thus yielded a dramatic improvement in selection of potent shRNAs (compare Figs. 6A and 6B).

The total binding energy plays a major role in governing shRNA inhibition. Increasing the stringency of $\Delta G_{\text{total}}$ from -25 to -27 and to -29 kcal/mol yields a monotonic increase in prediction success of the most potent shRNA inhibitors in the test set (Fig. 3.6A). At the -29 kcal/mol threshold, 82% (9/11) of designed shRNAs inhibit viral production by 90% or greater (Fig. 3.6A). In the HIV-1 genome, only about 2% of 19-mers are both completely accessible in the 13-nucleotide optimal window and bind at the ≤ -29 kcal/mol threshold.

To further evaluate the importance of the target accessibility rule, we also assayed 16 shRNAs that had total binding energies < -25 kcal/mol but did not conform to the rules developed here because they had nonzero target folding energies in the optimal 13 nt window. Of these sequences, 56% (9/16) reduce HIV-1 viral production
by a factor of 2 or more but only one of the sixteen shRNAs (6%) resulted in highly potent, greater than 90%, inhibition (black column, Fig. 3.6C). Poor inhibition by these sequences emphasizes the importance of an accessible optimal target window for potent RNAi-mediated repression.

3.2.7 shRNA repression and toxicity in human T cells.

Measuring shRNA-mediated inhibition by the 293T cell transfection assay used in this study has significant advantages because of its quantitative accuracy. However, an shRNA-based HIV-1 therapeutic will need to be effective in natural host immune cells. We therefore used a lentiviral vector to stably transduce a human T cell line, SupT1, with single copies of selected shRNA expression constructs identified in our test set. Cells were subsequently challenged with low and high doses of HIV-1 and the durability of inhibition was monitored (Fig. 3.7). In all cultures, HIV-1 production was either completely inhibited or was significantly delayed by shRNA expression, whereas HIV-1 replicated efficiently on cells transduced with the empty JS1 vector.

A key concern for therapeutic shRNA use is the potential for stimulating unintended off-target and cytotoxic effects. To quantify these effects, we transduced SupT1 cells with lentiviral constructs expressing this same set of shRNAs and measured the impact of lentiviral vector integration and shRNA expression using a competitive cell growth assay (39). No negative effect was caused by integration of the empty lentiviral vector, as observed previously (39). Of the shRNA constructs from our test set that were evaluated (Fig. 3.8), we observed no effects for shRNAs 1033, 1876, 4308, 4610, and 8302, and only minimal effects on SupT1 cell growth for shRNAs 8220 and 8221. Only one shRNA (4384) caused a significant reduction in cell growth (of ~4.5%).
**Figure 3.7:** Inhibition of HIV-1 replication in transduced SupT1 T cells. Cells were transduced with the empty JS1 lentivirus (negative control), JS1-shPol1 (a known effective inhibitor) (6), or lentiviral variants expressing 8 shRNAs randomly selected from among the most potent test set shRNAs. Cultures with CA-p24 amounts above a 1 ng/ml threshold were scored as positive for viral replication. The color code indicates when virus replication was apparent: within one week (white), two weeks (light grey), three weeks (dark grey), or no observable replication for up to 2 months (black).
**Figure 3.8:** Competitive cell growth curves of representative potent shRNAs from the test set. SupT1 T cells were transduced with lentiviral vectors expressing both shRNA and GFP and cultured together with non-transduced SupT1 T cells. To score effects induced by the lentiviral integration, cells were transduced with the empty JS1 vector expressing GFP but no shRNA. FACS measurements were used to quantify GFP+ and GFP− cell populations, corresponding to transduced and non-transduced cells, respectively. We plot GFP+/GFP− ratios (y-axis), which quantify cell growth defects (39). Cell samples were obtained twice weekly directly before cell passaging. One representative experiment of two repetitions is shown.
In sum, shRNAs identified in the quantitative cell-based assay using 293T cells (Fig. 3.6) also inhibit HIV-1 replication when stably transduced into T cells (Fig. 3.7) and generally show no or minimal impairment of cellular replication (Fig. 3.8).

3.2.8 Partial SHAPE information is sufficient to identify potent shRNA inhibitors.

The requirement for performing SHAPE to develop a high-content secondary structure model is the experimental cost of this approach for accurate shRNA design. SHAPE is rapidly becoming easier to perform and, with current mature technologies, it is straightforward to obtain structural data for ~500 nts (29). We thus evaluated whether obtaining partial SHAPE information might be sufficient to identify potent shRNA inhibitors. Prior experience emphasizes that there can be very strong "end effects" in RNA folding such that incorrect definition of the 5’ and 3’ ends of an RNA can cause large re-folding at internal sequences. We therefore used SHAPE data from two 500-nt regions, each flanked by 1,000 additional nucleotides of HIV-1 sequence in our folding analysis. Two test regions were chosen based on their spanning the largest number of shRNAs in our test set. We folded each of the two 2,500 nt regions (spanning nucleotides 601-3100 and 3201-5700), using SHAPE data for only the central 500 nucleotides (1601-2100 and 4200-4700). We then calculated the resulting free energies ΔG_{target} and ΔG_{total} for the test shRNAs in these regions. The energy values obtained using these partial folds are similar to energies obtained when using the complete genome model. In contrast, the structural models obtained without using SHAPE constraints yield free energy values that agree poorly with the SHAPE-directed models (Appendix 3). This analysis suggests that shRNA design is significantly improved through the inclusion of partial SHAPE data for a large RNA.
3.3 Discussion

Essentially all recognition processes involving RNA are critically dependent on the underlying base paired secondary (and higher-order tertiary) structure. The central role of accessibility at single target sites is well established for RNAi-mediated knockdown (14, 16, 34-37). Correspondingly, siRNA and shRNA prediction methods often incorporate estimates of target site accessibility into their algorithms. However, the structure of a long target RNA is generally incompletely understood \textit{a priori} and the resulting correlations between predicted and effective si/shRNAs are often poor (Table 3.1). Given this difficulty, newer algorithms have tended to become more complex and meld thermodynamic calculations with sequence signature and heuristic rules. However, these rules, applied to shRNA-mediated inhibition of HIV-1 production, do not consistently identify potent inhibitors.

In this work, we find that an extremely simple approach, involving the calculation of only two straightforward thermodynamic terms, significantly outperforms existing approaches when applied to inhibition of HIV-1 (Table 3.1). Our model considered only two simple RNA-RNA interactions central to the RISC ribonucleoprotein machinery (Fig. 3.1). Strong inhibition correlated with weak free energies of target folding ($r = -0.72$) within an optimal 13-nt window (Fig. 3.4b, grey bar) and with strong total binding energy ($r = 0.61$). These correlations were much weaker when SHAPE data was not used to direct calculation of the target HIV-1 RNA secondary structure model (Fig. 3.5 and Table 3.1), highlighting the requirement for an accurate target RNA structure in selection of shRNAs.
The correlation coefficient for our total binding energy metric is moderately weaker than that for the target folding energy metric ($|r| = 0.61$ versus $0.72$, respectively). However, requiring stronger $\Delta G_{\text{total}}$ improves the prediction success rate for sequences containing an optimal, fully accessible target window. Additionally, sequences with strong $\Delta G_{\text{total}}$ but less accessible optimal target windows were rarely potently repressed (Fig. 3.6C). Thus, a completely accessible optimal window appears critical for efficient repression and, once this condition is met, strong total binding energy improves prediction accuracies.

The correlations are consistent with models of target RNA association with the guide strand that involve an initial interaction with the seed region, followed by propagation to form a duplex containing all guide strand nucleotides (Fig. 3.1) (38, 40, 41). Our data also identify a previously unrecognized feature in which the optimal accessible target window extends 7 nucleotides downstream of the seed region binding site (Fig. 3.4). The required lack of secondary structure in this region of the target RNA may reflect unexplored interactions involving protein components of RISC.

Many RNAi design criteria are based on specific sequence signatures found more frequently in effectively repressed shRNA targets. Some of these signatures are consistent with the results of the present study. In particular, the observed preference for (more weakly pairing) A/U nucleotides at the 3’ end of the target binding interface (9, 42, 43), sometimes termed the asymmetry rule, likely corresponds to a qualitative sequence signature for an accessible seed region, which we quantify as $\Delta G_{\text{target}}$. Our work emphasizes that strong correlations can be obtained using structure-based
thermodynamic metrics without considering sequence characteristics, and may reflect a more direct physical basis for observed sequence effects.

In designing shRNAs to test our two structure-based design principles, we ignored other design criteria. Consequently, many of our successful test shRNAs violate sequence-based si/shRNA design principles identified in previous studies. These include having an A/U at position 10 of the guide strand to facilitate cleavage at this site (44) and avoidance of consecutive runs of identical nucleotides. Nevertheless, our tested shRNAs achieved highly potent HIV-1 inhibition (Fig. 3.6A), suggesting that these sequence-based design rules have limited applicability in this system.

There were two distinctive features of this study that may explain differences, and the relative success of our simple approach, as compared to prior work.

First, several prediction algorithms have emphasized the role of pre-existing target RNA structure in reducing RNAi efficiency (14-16, 22) and have obtained good siRNA predictions in the systems studied. However, we find that these measures of target accessibility do not correlate with shRNA-mediated knockdown for the training dataset used in our study (Table 3.1). The importance of structural accessibility and for an accurate model of secondary structure in the target RNA may vary with the targeted system and it appears to be especially critical in the context of shRNA-mediated inhibition of the highly structured (31) HIV-1 RNA genome.

Second, RNAi-mediated knockdown is highly sensitive to the effective concentration of the si/shRNA. Higher concentrations of shRNA yield increased inhibition (Fig. 3.2 and ref. (8)) but also potentially contribute to saturation of the RNAi machinery, yield off-target effects, and stimulate cytotoxic innate immune responses

86
(32, 33). We therefore focused on shRNA-mediated inhibition at sub-saturating amounts of transfected shRNA constructs. Importantly, when we probed cytotoxic effects induced by shRNA transduction of SupT1 T cells using a sensitive competitive cell growth assay (39), we found that all but one designed shRNA showed no or minimal reduction in cell growth.

The use of sub-saturating shRNA amounts may also permit the simultaneous expression of several different shRNAs and thereby reduce the chance of developing escape mutants. This is an especially important consideration for clinical applications when targeting rapidly mutating viruses such as HIV-1 with RNAi-based gene therapies (45, 46). Potential shRNA-based therapeutics will have to function in human T-cells, the natural HIV-1 host. While we used human 293T cells in our quantitative virus production inhibition studies, the potent shRNA inhibitors that we identified in 293T cells were generally able to inhibit HIV-1 replication in human SupT1 T-cells, consistent with the observation that determinants of shRNA repression appear to be conserved between evolutionarily disparate cell types (42).

In sum, an extremely simple model is sufficient to identify potent shRNAs that strongly inhibit HIV-1 production in cell culture, even when the shRNAs are used at sub-saturating concentrations. These sequences are strong candidate targets for further evaluation in pre-clinical anti-HIV assays. Critically, our approach emphasizes the profound role of RNA structure in tuning RNAi function and absolutely requires an accurate secondary structure model of the RNA target, as obtainable from SHAPE chemical probing information. The inclusion of partial SHAPE data, corresponding to that obtainable in a single experiment covering ~500 nts, also yields significant
improvements in identification of potent shRNA inhibitors. In the broadest terms, this work shows how profoundly RNA structure influences biological function and emphasizes that, if a high-content model for an RNA fold is obtained experimentally, deep insight into underlying mechanisms can be obtained.

3.4 Methods

3.4.1 Plasmid constructs

The shRNA expression plasmids pSuper-shRNA were constructed as described (47), were verified by sequencing, and were expected to yield 19-nucleotide guide strands upon processing by the cellular miRNA biogenesis pathway. The pLAI plasmid was used to express the HIV-1 isolate LAI (GenBank accession K02013). The pRL-CMV plasmid (Promega) expressing Renilla luciferase was co-transfected to control for transfection efficacy. For shRNA titration experiments, the pBluescript plasmid (Stratagene) was used to normalize plasmid amounts (6). To generate lentiviral vectors expressing the shRNAs, the H1-shRNA cassettes were cloned into the lentiviral vector JS1 (pRRLcptpgkgfppreSsin) (48), as described previously (6). JS1 harbors a GFP cassette for easy identification of transduced cells. For lentivirus production, the JS1 variants were co-transfected with the packaging plasmids pSYNGP, pVSVG, and 250 ng pRSV-rev.

3.4.2 Cell culture

Human embryonic kidney 293T adherent cells were grown in Dulbecco’s modified Eagle's medium (DMEM), supplemented with 10% fetal calf serum (FCS), 100 U/ml penicillin, and 100 μg/ml streptomycin in a humidified chamber at 37 °C and 5%
CO₂. The SupT1 T cell line was maintained in Advanced RPMI (Gibco, Carlsbad, USA) supplemented with L-glutamine, 1% FCS, 30 U/ml penicillin and 30 μg/ml streptomycin; cells were maintained at 37 °C and 5% CO₂.

3.4.3 Transfections and HIV-1 production experiments

For the HIV-1 production assay, co-transfections of pLAI and the shRNA vector were performed in a 24-well format. On the first day, 1.5 x 10⁵ 293T cells were seeded per well in DMEM without antibiotics. The next day, 250 ng pLAI, 25 ng of shRNA vector, and 1 ng pRL-CMV in Lipofectamine 2000 (Invitrogen, following the manufacturer’s protocol) were added. After 48 hours, samples of cell supernatant were taken for CA-p24 ELISA quantification and cells were lysed for measurement of Renilla luciferase activity (Renilla Luciferase Assay System, Promega). All transfections were carried out in duplicate and repeated twice. Negative controls were performed using an empty pSuper vector and with an shRNA directed against the firefly luciferase gene. Positive inhibition controls were performed using shRNAs directed against known efficiently repressed HIV-1 targets (LDR9, GagC, Pol1, and Nef19) (6). For shRNA titrations, 250 ng pLAI was co-transfected with 1 to 125 ng pSuper shRNA variants and 1 ng pRL-CMV. Variable amounts of pBluescript were added to yield equivalent DNA concentrations per transfection. All data from transfection experiments were corrected with Factor Correction to compensate for inter-experimental differences (49).

For lentivirus production, 293T cells were seeded in a 6-well-plate format one day prior to transfection to reach 70% confluency and subsequently transfected with 950 ng lentiviral plasmid JS1-shRNA or empty JS1, 600 ng pSYNGP, 330 ng pVSVg, and 250 ng pRSV-rev with Lipofectamine 2000, as described (6).
The HIV-1 virus stock was produced in 293T cells, which were seeded in a T75 flask to yield 70% confluency one day prior to transfection. 40 μg HIV-1 pLAI was transfected following the Lipofectamine 2000 protocol and supernatant was harvested at 48 hours post transfection. Cells were removed by centrifugation (4000 xg) and the supernatant was aliquoted and stored at -80 °C. The HIV-1 LAI virus stock was quantified for CA-p24 by ELISA measurement.

3.4.4 Lentiviral transduction, HIV- challenge experiments, and competitive cell growth assay.

First, lentiviral titers were determined via titration on SupT1 cells, scoring the percentage of GFP-positive cells by FACS (6). Then, a multiplicity of infection (MOI) of 0.15 was used for transduction of SupT1 T cells (6). Four days later, cells were sorted for GFP-expression by fluorescence-activated cell sorting. $2 \times 10^5$ transduced SupT1 cells were challenged with HIV-1 (0.01 or 0.05 ng CA-p24, corresponding to low and high titers, respectively) and monitored daily for up to 90 days by eye and light microscopy to score for cytopathic effects. Supernatants were collected in parallel three times per week to determine CA-p24 levels. Transduced SupT1 cells were screened for a negative impact on the cell growth as induced by lentiviral integration (JS1 control cells) and shRNA expression using the competitive cell growth (CCG) assay (39). In brief, SupT1 cells were transduced to obtain around 30% GFP+ cells. The GFP+/GFP− ratio was analyzed over a 30 day period by FACS measurement, and the ratio at day 3 was normalized to 1. Impact on cell growth can be converted as percentage reduction in cell growth (39).
3.4.5 Free energy calculations

Free energies were calculated using the OligoWalk algorithm in the RNAstructure package (50). OligoWalk takes an RNA secondary structure model as input and considers a bimolecular equilibrium between this folded RNA and every possible complementary RNA oligonucleotide of a user-defined length. The duplex annealing energy $\Delta G_{\text{duplex}}$ and target unfolding energy $\Delta G_{\text{target}}$ are calculated. We assumed that shRNAs were cleaved into 19-nt guide strands. The target unfolding energy $\Delta G_{\text{target}}$ was calculated as the energy required to break base pairs in the input secondary structure model (break local structure option in OligoWalk). The structural model was inputted in connect file (.ct) format. Our SHAPE-based secondary structure model differs from the published model (31) in two regions: the tRNA primer and a heuristically predicted pseudoknot are absent in our input file because OligoWalk only permits single molecule inputs and does not allow pseudoknots. We excluded shRNA target sequences in these two regions because removal of their binding partners resulted in falsely weak target folding energies and falsely strong total binding energies; these regions are highly inaccessible and thus poor candidates for RNAi targets (31). We also created a secondary structure model of the NL4-3 HIV-1 RNA using only thermodynamic parameters by folding the RNA in RNAstructure without any experimental constraints. Perl scripts were created to vary the length and position parameters by changing the input oligomer length and by shifting the target energy window relative to the guide strand binding site, respectively.

3.4.6 Correlation calculations for other algorithms

We calculated correlation coefficients between our 84 member training dataset
and siRNA design algorithm scores using the i-Score Designer tool (http://www.med.nagoya-u.ac.jp/neurogenetics/i_Score/i_score.html) for all methods except for RNAs and Sirna. Sirna scores were calculated using the online Sfold package (http://sfold.wadsworth.org). We applied the RNAs algorithm using Perl scripts generously provided by the Hofacker lab. RNAs predicts accessible sequences that are subsequently further ranked based on the worst scores among six different metrics. Twenty of our 84 sequences were predicted to be good inhibitors by RNAs. We calculated the linear correlation between the RNAs ranks of these 20 sequences and our experimental data.

3.4.7 Viral production datasets

We used fully independent datasets to train and test our rules (Fig. 3.3). Our training dataset was comprised of the shRNA sequences previously published by ter Brake et al. (6). Viral production values were obtained using relatively low plasmid transfection amounts (20 ng) and sequences were chosen without using si/shRNA selection rules; instead, sequences were selected based on sequence conservation among HIV-1 isolates. A total of 86 HIV-1 sequences were targeted by the shRNAs in this dataset. However, two sequences were, by coincidence, identical (Appendix 1, asterisks). In the current study, we averaged the viral production values for the identical sequences and used the resulting 84 sequences in our correlation calculations. The 26-member test dataset was created as described in the text by applying the two thermodynamic rules identified derived from the training dataset. As with the training dataset, no additional si/shRNA design rules were invoked.
3.5 Acknowledgements

This work was supported by National Institutes of Health grant AI068462 (to K.M.W.) and a Netherlands Organization for Scientific Research Division of Chemical Sciences (NWO-CW) TOP grant (to B.B.). J.T.L. was supported by National Research Service Award F30DA027364 and Medical Scientist Training Program T32GM008719; S.A.K. was supported by a fellowship from the Deutscher Akademischer Austausch Dienst (DAAD); and J.M.W. was a Postdoctoral Fellow of the UNC Lineberger Comprehensive Cancer Center.
3.6 Appendix 1

shRNA training dataset sequences and their relative viral production values. A relative viral production value of 1 represents no inhibition whereas low values represent strong inhibition. Two sequences were coincidentally identical and are denoted by asterisks.

<table>
<thead>
<tr>
<th>No.</th>
<th>NL4-3 starting position</th>
<th>guide strand core sequence</th>
<th>target RNA sequence</th>
<th>Relative viral production</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>326</td>
<td>CGCACCCAUCCUCUCUUU</td>
<td>AAGGAGAGAGAUGGGUGCG</td>
<td>0.47</td>
</tr>
<tr>
<td>2</td>
<td>327</td>
<td>UCGACCAUCCUCUCUCCU</td>
<td>AAGGAGAGAUGGGUGCGA</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>328</td>
<td>CUCGACCCAUCCUCUCCU</td>
<td>GAGAGAGAGAUGGGUGCGA</td>
<td>0.06</td>
</tr>
<tr>
<td>4</td>
<td>329</td>
<td>UCUCGCACCAUCCUCUCUC</td>
<td>GAGAGAGAGAUGGGUGCGA</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>330</td>
<td>CUCUCGCACCCAUCCUCUCU</td>
<td>AGAGAGAGAUGGGUGCGAG</td>
<td>0.06</td>
</tr>
<tr>
<td>6</td>
<td>331</td>
<td>GCUCUCGCACCAUCCUCUC</td>
<td>GAGAGAGAGAUGGGUGCGAG</td>
<td>0.32</td>
</tr>
<tr>
<td>7</td>
<td>332</td>
<td>CGUCUCGCACCAUCCUCU</td>
<td>AGAGAGAGAUGGGUGCGAG</td>
<td>0.16</td>
</tr>
<tr>
<td>8</td>
<td>333</td>
<td>AGCGUCUCGCACCCAUACUC</td>
<td>GAGAGAGAGAUGGGUGCGAG</td>
<td>0.21</td>
</tr>
<tr>
<td>9</td>
<td>334</td>
<td>GACGCUCUCGCACCAUCCU</td>
<td>AGAGAGAGAUGGGUGCGAG</td>
<td>0.09</td>
</tr>
<tr>
<td>10</td>
<td>1146</td>
<td>ACAUUUUUAACUUUUUUAAU</td>
<td>AAUAAAAUAAAGAAAGUG</td>
<td>0.73</td>
</tr>
<tr>
<td>11</td>
<td>1147</td>
<td>UACAUUUCUACUUUUAUAU</td>
<td>AAUAAAAUAAAGAAAGUA</td>
<td>0.7</td>
</tr>
<tr>
<td>12</td>
<td>1363</td>
<td>GUCGUCAUCCUUCUCUCUA</td>
<td>UAGAGAAAAUAAUGACAG</td>
<td>0.54</td>
</tr>
<tr>
<td>13</td>
<td>1364</td>
<td>UGCUUCAUCCUUCUCUCUU</td>
<td>AGAGAAAAUAAUGACAGCA</td>
<td>0.36</td>
</tr>
<tr>
<td>14</td>
<td>1365</td>
<td>AUGCUUCAUCCUUCUCUCU</td>
<td>GAAGAAAAUUGACAGCAU</td>
<td>0.14</td>
</tr>
<tr>
<td>15</td>
<td>1366</td>
<td>CAUGCUUCAUCCUUCUUUUCU</td>
<td>AAUAAAAUUGACAGCAUG</td>
<td>0.25</td>
</tr>
<tr>
<td>16</td>
<td>1623</td>
<td>UCCCCUUUUUUAUAGCCUG</td>
<td>CAGGCUAAUUUUUUGAGGA</td>
<td>0.78</td>
</tr>
<tr>
<td>17</td>
<td>1624</td>
<td>UCCCCUAAAAUUUAGCUU</td>
<td>AGGCUAAUUUUUUGAGGAA</td>
<td>0.75</td>
</tr>
<tr>
<td>18</td>
<td>1874</td>
<td>CGUUAUCCUCCUCUGCUGUGU</td>
<td>ACAGGACGAUGAUCAGC</td>
<td>0.03</td>
</tr>
<tr>
<td>19</td>
<td>1875</td>
<td>ACAGUUAUCAUUGCUCUGUG</td>
<td>CAGGACGAUGAUCAGU</td>
<td>0.14</td>
</tr>
<tr>
<td>20</td>
<td>1921</td>
<td>UACAUUUCUUGGUUCCCAU</td>
<td>UAGGAAACAAAUGAUA</td>
<td>0.37</td>
</tr>
<tr>
<td>21</td>
<td>1922</td>
<td>CUAUCAUUUUGGUUUCUA</td>
<td>UGGGAAACAAAAGUAUG</td>
<td>0.45</td>
</tr>
<tr>
<td>22</td>
<td>1923</td>
<td>CUAUCAUUUUGGUUUCUCC</td>
<td>GGAACAAAUGAUGAGG</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>3719</td>
<td>CUUGUUCUUCUCUCCAAUU</td>
<td>AUUGGAGAAUGGAAAAG</td>
<td>0.03</td>
</tr>
<tr>
<td>24</td>
<td>3720</td>
<td>ACUGUUCUUCUCUCUCCAA</td>
<td>UUGGAGAAUGGAAAAG</td>
<td>0.46</td>
</tr>
<tr>
<td>25</td>
<td>4085</td>
<td>CUCGCGACUUCUCUCGCAA</td>
<td>UUACAGGAAGAUGGCCAG</td>
<td>0.41</td>
</tr>
<tr>
<td>26</td>
<td>4086</td>
<td>ACUGGCGCAUCUUCUCUCGUA</td>
<td>UAGCAGGAAGAUGGCCAG</td>
<td>0.22</td>
</tr>
<tr>
<td>27</td>
<td>4196</td>
<td>UUUGGGAUGUUGGGAUAAUU</td>
<td>AUUCCUAAUCCUCCAAA</td>
<td>0.55</td>
</tr>
<tr>
<td>28</td>
<td>4197</td>
<td>CUUUGGGAUGUUGGGAUG</td>
<td>UUCCUCAAUCCUCCCAA</td>
<td>0.67</td>
</tr>
<tr>
<td>29</td>
<td>4322</td>
<td>CUUUCUUUUUAAAUGUG</td>
<td>CACAUUUUAAAGAAAAG</td>
<td>0.73</td>
</tr>
<tr>
<td>30</td>
<td>4323</td>
<td>CUCUUUCUUAAAAAUUGU</td>
<td>AAUAAUAAAGAAAAGG</td>
<td>1.05</td>
</tr>
<tr>
<td>31</td>
<td>4324</td>
<td>CCCCCCCUUCCUUUAAAAUG</td>
<td>CAAAUUUAAAGAAAAGG</td>
<td>0.68</td>
</tr>
<tr>
<td>32</td>
<td>4325</td>
<td>CCCCCCUUCCUUUUAAAAAUU</td>
<td>AAUAAUAAAGAAAAGG</td>
<td>0.69</td>
</tr>
<tr>
<td>33</td>
<td>4326</td>
<td>CCCCCCUUUCUUUUAAAAAUU</td>
<td>AAUAAUAAAGAAAAGG</td>
<td>0.62</td>
</tr>
<tr>
<td>No.</td>
<td>Sequence</td>
<td>Identity</td>
<td>Similarity</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>--------------------------------</td>
<td>----------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>CCCCCCUUUUUCUUUUAAAAA</td>
<td>UUUUAAAGAAAAGGGGGG</td>
<td>0.98 *</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>UCCCCCUUUUCUUUUAAAAA</td>
<td>UUAAAAGAAAAAGGGGGGA</td>
<td>0.95 *</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>AUCCCCCUUUUCUUUUAAAAA</td>
<td>UUAAAAGAAAAGGGGGGAU</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>AAUCCCCCUUUUCUUUAAAAA</td>
<td>UAAAAAGAAAAGGGGGGAUU</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>CAAUCCCCCUUUUCUUUAAAAA</td>
<td>AAAAGAAAAGGGGGGAUUG</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>CAUCCCCCUUUUCUUUAAAAA</td>
<td>AAAGAAAAGGGGGGAUUGG</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>CCAAUCCCCCUUUUCUUUAAAAA</td>
<td>AAGAAAAGGGGGGAUUGG</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>CCAAUCCCCCUUUUCUUUAAAAA</td>
<td>AGAAGAAAAGGGGAUUGG</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>CCAACCAUCCCCCUUUUUCUUUUC</td>
<td>GAAAAAGGGGGGAUUGGG</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>CCAACCAUCCCCCUUUUUCUUUUC</td>
<td>AAAAGGGGGGAUUGGGG</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>UUUCUUCUUCUUCGGACAGUA</td>
<td>UACAGUCAGGGGAAGAA</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>AAUCUCUUCUUCUUCGGACUG</td>
<td>ACAGUCAGGGGAAGAA</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>UAUUCUUCUUCUUCGGACUG</td>
<td>CAGUCAGGGGAAGAA</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>UUCUCUUCUUCUUCGGACUG</td>
<td>AAAAGGGGGGAUUGGGG</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>ACCCGAAAAUUUUUGAAUU</td>
<td>AAAUUCAAAAUUUUUGG</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>AAACCGAAAAAUUUGGAA</td>
<td>UCAAAAUUCUGGGUUAA</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>AAACAACCGAAAAAUUUGG</td>
<td>CAAAUUCUGGGUUUAAU</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>AAACCAACCGAAAAAUUUGG</td>
<td>AAAUUCUGGGGUUAAU</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>UGUAAUAACCCGAAAAUU</td>
<td>AAUUXUGCUGGGGUUAC</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>CUGUAAUAAACCCGAAAAAU</td>
<td>AAUUXUGCUGGGGUUAC</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>CCCCUCACCUUUCUCGAGG</td>
<td>CUCUGGAAAGGGAAGGGGA</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>GCCCCUCACCUUUCUCAGA</td>
<td>UCUGGAAAGGGAAGGGG</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>UGCCCUUCACCUUUCACUG</td>
<td>CUGGAAAGGGAAGGGGCA</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>UGCUCUUCACCUUUCUCACC</td>
<td>UGGAAAGGGAAGGGGAC</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>ACUGCCCCCUUCACCUUUC</td>
<td>GGAAGGUGAAGGGGCGAUG</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>UAUCUACUCUACCUUUCACC</td>
<td>GAAGGUAAGGGAAGGGCA</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>CUACUCUCUCUCUACCUUU</td>
<td>AAAGGUAAGGGAAGGGGCAG</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>ACUACUGCCCCCUUCACCUU</td>
<td>AGGUGAAGGGAAGGGCAGU</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>UACUACUCUCUACCUUUCACC</td>
<td>AGGUGAAGGGAAGGGCAGU</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>UUACUACUCUACUCUCCCUCACCU</td>
<td>GGUGAAAGGGAAGGGCAGUAA</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>UUACUACUCUACUCUCCCUCACCU</td>
<td>GGUGAAAGGGAAGGGCAGUAA</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>CUGCACAUCUGUUUUCAUAA</td>
<td>UAGGAAAACAGAUGGCAG</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>CUGCCCUACUCUCCCCUCACCAU</td>
<td>AUGGAAACAGAUGGCAG</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>ACCUCCGUACUCUGUUUUCACCAU</td>
<td>AGGAAAACAGAUGGCAG</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>CACUGCCUACUGUUIUCUUCACCAU</td>
<td>GGAAACAGAUGGCAGUG</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>GCUCUCUUCUCCUCAGGA</td>
<td>UCUACUUGGGAAGAAGG</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>CGCUUCUUCUCCUGCCAUAGG</td>
<td>CCUACUUGGGAAGAAGCG</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>CGCUUCUUCUCCUGCAUAG</td>
<td>CUAUGCAGGAAGAAGCGG</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>UCCCUCUUCUUCUGGCAUA</td>
<td>UAUGCAGGAAGAAGCGGA</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>UCCCUCCUUCUCCUGCAU</td>
<td>AUGCCGAAGAAGCGAGG</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>CCCAUAGUGCUUCCUCGUGC</td>
<td>CAGCAAGAAGCAUAGGGG</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>GCCCAUAGUGCUUCCUCGCU</td>
<td>AGCAGGAAGCAUAGGGG</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>AGCCCAUAGUGCUUCCUCGUC</td>
<td>GCAGGAAGCAUAGGGG</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGCCCAUAGUCUUCCUG</td>
<td>CAGGAAGCACUAUGGGGC</td>
<td>0.77</td>
</tr>
<tr>
<td>77</td>
<td>7337</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>8599</td>
<td>CCCCCUUUCUUUAAAG</td>
<td>CUUUUAAAGAAAGGGG</td>
<td>0.46</td>
</tr>
<tr>
<td>79</td>
<td>8600</td>
<td>CCCCCUUUCUUUAAAG</td>
<td>UUUUAAAGAAAGGGG</td>
<td>1.15</td>
</tr>
<tr>
<td>80</td>
<td>8601</td>
<td>CCCCCUUUCUUUAAAG</td>
<td>UUUUAAAGAAAGGGG</td>
<td>1.07</td>
</tr>
<tr>
<td>81</td>
<td>8602</td>
<td>UCCCCCUUUCUUUAAAG</td>
<td>UUUAAAAGAAAGGGGA</td>
<td>0.96</td>
</tr>
<tr>
<td>82</td>
<td>8603</td>
<td>GUCCCCCUUUCUUUAA</td>
<td>UUUAAAAGAAAGGGGAC</td>
<td>1.2</td>
</tr>
<tr>
<td>83</td>
<td>8604</td>
<td>AGUCCCCCUUUCUUUUA</td>
<td>UAAAAGAAAGGGGACU</td>
<td>0.93</td>
</tr>
<tr>
<td>84</td>
<td>8605</td>
<td>CAGUCCCCCUUUCUUU</td>
<td>AAAAGAAAGGGGACUG</td>
<td>1.4</td>
</tr>
<tr>
<td>85</td>
<td>8606</td>
<td>CCAGUCCCCCUUUCUUU</td>
<td>AAAGAAAGGGGACUGG</td>
<td>1.05</td>
</tr>
<tr>
<td>86</td>
<td>8607</td>
<td>UCCAGUCCCCCUUUCUU</td>
<td>AAAGAAAGGGGACUGGA</td>
<td>1.18</td>
</tr>
</tbody>
</table>
3.7 Appendix 2

shRNA test dataset core sequences designed using the two thermodynamic rules identified in this study and their relative viral production values.

<table>
<thead>
<tr>
<th>No.</th>
<th>NL4-3 starting position</th>
<th>guide strand core sequence</th>
<th>target RNA sequence</th>
<th>Relative viral production</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1021</td>
<td>UCACUCCCCUUUGGUUCUC</td>
<td>GAGAACAAAGGGGAGUGA</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>1026</td>
<td>CUAUGUCACUCCCUUGGG</td>
<td>CCAAGGGGAAUGACAUAG</td>
<td>0.55</td>
</tr>
<tr>
<td>3</td>
<td>1033</td>
<td>GUGCCUGCUCAGUCAUGUCUC</td>
<td>GAAGUGACAGAACAGAGA</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>1876</td>
<td>UACUGUAUCUCUGCUCCUC</td>
<td>AGGAGCAGAUGACAGUA</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>1877</td>
<td>AUACUGUAUCUCUGCUCCUC</td>
<td>GGAGCAGAUGACAGUA</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>1882</td>
<td>UUCCUAUAACUGUAUCAUC</td>
<td>AGAUAUAAGAAAGAGAA</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>1883</td>
<td>CUUCUAUAACUGUAUCAUC</td>
<td>GAUGAUACAUGAAAGAAG</td>
<td>0.02</td>
</tr>
<tr>
<td>8</td>
<td>1884</td>
<td>UCUCUAUAACUGUAUCAUC</td>
<td>AUAUAUGAAAGAGAAA</td>
<td>0.3</td>
</tr>
<tr>
<td>9</td>
<td>1959</td>
<td>UCAUACUGUCUUACUUUGA</td>
<td>UCAAGUAAGACAGUAAGA</td>
<td>0.77</td>
</tr>
<tr>
<td>10</td>
<td>3721</td>
<td>UACUGUAUCCAUUCCUCCA</td>
<td>UGGAGGAAUGCAGAAGA</td>
<td>0.05</td>
</tr>
<tr>
<td>11</td>
<td>3722</td>
<td>CUACUGUGUCAUUCCCUCUC</td>
<td>GGAGGAAAGAAGAAGA</td>
<td>0.02</td>
</tr>
<tr>
<td>12</td>
<td>3723</td>
<td>UCUACUGUGUCAUUUCCUC</td>
<td>GAGGAACAAAGAGAGA</td>
<td>0.05</td>
</tr>
<tr>
<td>13</td>
<td>3724</td>
<td>AUCUCACUGUCAUUUCCU</td>
<td>AGAUAUAACAGAACAGA</td>
<td>0.15</td>
</tr>
<tr>
<td>14</td>
<td>3725</td>
<td>UACUCACUGUCAUUUCCCUC</td>
<td>GAAAUAAGAAGAAGAACA</td>
<td>0.27</td>
</tr>
<tr>
<td>15</td>
<td>3726</td>
<td>UUAUCACUGUCAUUUUCUUC</td>
<td>GAAUAUAACAGAAGAUA</td>
<td>0.06</td>
</tr>
<tr>
<td>16</td>
<td>4308</td>
<td>UUGUAGGAAUAACUGCCCA</td>
<td>UGGCAUAUAAGAAGAACA</td>
<td>0.08</td>
</tr>
<tr>
<td>17</td>
<td>4380</td>
<td>AUGCUGUGUGCUAUAUGUGU</td>
<td>ACAUAUAAGCAGAAGAAC</td>
<td>0.29</td>
</tr>
<tr>
<td>18</td>
<td>4384</td>
<td>UUGUAGUGUGUGCUAUAUU</td>
<td>AAUAACAGACAGAACAAA</td>
<td>0.11</td>
</tr>
<tr>
<td>19</td>
<td>4386</td>
<td>GUUGUUGUAUGUGUCUGCUAG</td>
<td>UAGCAACAGACAAAAAC</td>
<td>0.01</td>
</tr>
<tr>
<td>20</td>
<td>4610</td>
<td>GUCUACUGUCACACACAAUUC</td>
<td>GAUUGUGUGCAAGAAGAC</td>
<td>0.06</td>
</tr>
<tr>
<td>21</td>
<td>7184</td>
<td>CUUCUCAUAUGUGUCCCUCA</td>
<td>UAGAGGAAUUGGAGAAGA</td>
<td>0.59</td>
</tr>
<tr>
<td>22</td>
<td>7191</td>
<td>UAAUAUCACUUCCUCAAUUUG</td>
<td>CAAUGAGGAGAGAAUAUA</td>
<td>0.4</td>
</tr>
<tr>
<td>23</td>
<td>7193</td>
<td>UAAUAUAUACUCCUAUUCAUA</td>
<td>AAUAAGGAGAAGAUAUA</td>
<td>0.19</td>
</tr>
<tr>
<td>24</td>
<td>8220</td>
<td>UGGUCUCUCACGCUACUGCU</td>
<td>AGCAGUAGCUGAGAGGACA</td>
<td>0.03</td>
</tr>
<tr>
<td>25</td>
<td>8221</td>
<td>CUGUCUCUGCAGCUACUGCUC</td>
<td>GCAGUAGCUGAGAGGAGA</td>
<td>0.05</td>
</tr>
<tr>
<td>26</td>
<td>8302</td>
<td>UCCUUUCAAGCCCCUGUCU</td>
<td>AGACAGGCUUCAAGGAA</td>
<td>0.03</td>
</tr>
</tbody>
</table>
3.8 Appendix 3

Free energy values for test shRNA sequences calculated using two 2,500 nucleotide folds of HIV-1 genome segments. SHAPE data were included for only the central 500 nucleotides of each fold.

<table>
<thead>
<tr>
<th>Fold</th>
<th>shRNA</th>
<th>Partial fold (2500 nt)</th>
<th>Complete fold without SHAPE (9173 nt)</th>
<th>Complete fold with SHAPE (9173 nt)</th>
<th>Relative viral production</th>
</tr>
</thead>
<tbody>
<tr>
<td>601-3100</td>
<td></td>
<td>ΔG\text{target}</td>
<td>ΔG\text{total}</td>
<td>ΔG\text{target}</td>
<td>ΔG\text{total}</td>
</tr>
<tr>
<td>1876</td>
<td></td>
<td>0</td>
<td>-28.4</td>
<td>-4.7</td>
<td>-19.7</td>
</tr>
<tr>
<td>1877</td>
<td></td>
<td>0</td>
<td>-27.9</td>
<td>-6.5</td>
<td>-18.7</td>
</tr>
<tr>
<td>1882</td>
<td></td>
<td>0</td>
<td>-26</td>
<td>-9.2</td>
<td>-14.3</td>
</tr>
<tr>
<td>1883</td>
<td></td>
<td>0</td>
<td>-27</td>
<td>-9.5</td>
<td>-16.7</td>
</tr>
<tr>
<td>1884</td>
<td></td>
<td>0</td>
<td>-26</td>
<td>-8.4</td>
<td>-16.8</td>
</tr>
<tr>
<td>1959</td>
<td></td>
<td>0</td>
<td>-21.5</td>
<td>-9.2</td>
<td>-16</td>
</tr>
<tr>
<td>3201-5700</td>
<td></td>
<td>ΔG\text{target}</td>
<td>ΔG\text{total}</td>
<td>ΔG\text{target}</td>
<td>ΔG\text{total}</td>
</tr>
<tr>
<td>4308</td>
<td></td>
<td>0</td>
<td>-31.8</td>
<td>-6.7</td>
<td>-15.7</td>
</tr>
<tr>
<td>4380</td>
<td></td>
<td>0</td>
<td>-27.6</td>
<td>-4.5</td>
<td>-24.3</td>
</tr>
<tr>
<td>4384</td>
<td></td>
<td>0</td>
<td>-27.4</td>
<td>-1.4</td>
<td>-29.2</td>
</tr>
<tr>
<td>4386</td>
<td></td>
<td>0</td>
<td>-29</td>
<td>-3.5</td>
<td>-30.8</td>
</tr>
<tr>
<td>4610</td>
<td></td>
<td>0</td>
<td>-28.4</td>
<td>-17</td>
<td>-18.3</td>
</tr>
</tbody>
</table>
3.9 References


CHAPTER 4
CONCLUSIONS AND CLINICAL RELEVANCE

4.1 Structural organization of the HIV-1 Gag-Pol frameshift element

In this work, we used LNA binding monitored by SHAPE to confirm a novel SHAPE-directed model of the HIV-1 group M Gag-Pol frameshift domain. The frameshift element comprises a larger, more complex domain of 140 nucleotides than the currently accepted model (1). We further discover that this domain is organized into both highly stable helices that are conserved across biological states, as well as less stable intervening helices that are capable of adopting two alternative conformations. The formation of the domain depends upon the existence of all 140 nucleotides of the domain. This larger domain should therefore be used in future frameshifting studies.

4.2 Structural requirements for effective shRNA targeting of HIV-1

Using a SHAPE-directed secondary structure model of an entire HIV-1 genome, we discovered that extremely simple thermodynamic rules can predict how well sequences within the HIV-1 genome can be inhibited by shRNAs (2). These rules are highly dependent on having an accurate, high resolution secondary structural model of the HIV-1 genome, underscoring the important role model accuracy can play for inferring biological functions. This work additionally makes the novel observation that effectively targeted HIV-1 sequences tend to have an unstructured region that extends
beyond the inhibitor binding site. These results motivate further research into the mechanistic significance of this novel structural requirement.

### 4.3 Applications to RNA structure probing

The LNA binding method used for verifying the SHAPE-directed frameshift element could be broadly applicable to similar efforts to ascertain the existence of specific RNA helices. This method requires heat-denaturing the RNA in the presence of LNA to achieve sufficient binding to stable helices. This heating step restricts the relevance of this method to RNA regions that can adopt the native fold following heat denaturation. For large RNAs extracted from biological environments, such as the *ex vivo* HIV-1 RNA studied here, this necessitates a control experiment where SHAPE reactivities from the heat denatured and refolded RNA are compared to reactivities from the biological state. However, this method should be generally applicable to *in vitro* transcribed RNAs that typically undergo a heat denaturation step prior to refolding and structural interrogation.

Formamide denaturation of RNA analyzed by SHAPE probing could also be broadly useful for studying the relative stabilities of RNA structural elements. Additionally, identification of highly stable structural elements may provide an additional constraint for structure prediction of large, complex RNAs.

### 4.4 Clinical relevance

The rapid development of drug resistance and lack of an existing cure for HIV make the development of novel therapeutic approaches an important endeavor. In this work, we investigated the RNA genome itself as a new potential antiretroviral target.
We confirm a novel SHAPE-directed model of an important potential target in the HIV-1 genome, the Gag-Pol frameshift element. Frameshifting is the process by which all viral-encoded enzymes are generated and is therefore essential for viral fitness (3). As frameshifting depends critically on the RNA structure of the frameshift domain, this work should provide a more accurate framework for further mechanistic and therapeutic studies of this important drug target.

Additionally, we uncover key structural signatures of HIV-1 target sites that are effectively repressed by RNA interference (RNAi)-based inhibitors. Despite their attractiveness as a potential antiretroviral strategy, RNAi-based therapeutic approaches can saturate the cellular RNAi machinery and trigger undesirable immune-mediated cytotoxicity (4,5). Potential RNAi-based inhibition of HIV-1 should therefore include only the most highly potent shRNAs. The structural characteristics determined in this work should aid in ongoing efforts to identify these highly potent inhibitors. Additionally, some of the potent shRNA sequences identified in this work could be suitable candidates for pre-clinical studies.
4.5 References


