Development and evaluation of peptide mimetic inhibitors of HIV-1 replication

Structure of the conserved core region of the lincRNA Cyrano

Alisha Jones

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Gabriele Varani, Chair

Robert Synovec

Matthew Bush

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Development and evaluation of peptide mimetic inhibitors of HIV-1 replication.
Structure of the conserved core region of the lincRNA Cyrano.

Alisha Jones

Chair of the Supervisory Committee:

Professor Gabriele Varani

Department of Chemistry and Biochemistry

An estimated 1.2 million people in the United States are living with the human immunodeficiency virus (HIV), as reported by the Center for Disease Control (CDC). Because mutations constantly occur throughout the HIV genome, a cure has yet to be found to eradicate the virus. Instead, treatment with Highly Active Antiretroviral Therapy (HAART) effectively postpones the progression of HIV to AIDS (Acquired Immune Deficiency Syndrome) by eliminating most circulating viruses through drugs that target reverse

transcription, integration, or HIV-1 fusion and entry. While this combination therapy has been shown to be very effective, it is unable to target viral reservoirs in patients, allowing emerging viruses to rapidly repopulate the body if therapy is interrupted. Thus, an important gap in current HIV treatment is the absence of drugs that block the emergence of the virus from latently infected cells, principally residing amongst resting CD4+ cells.

The viral transactivator protein (Tat) is required for viral gene expression for both the exponential growth of the virus and the activation of integrated, but latent, proviral genomes. Since the emergence of the virus from latency relies on Tat-dependent transcription, inhibitors of Tat-function are expected to be potent blockers of viral escape from latency. My thesis explores three approaches to developing potent, proteolytically stable, selective inhibitors of HIV-1 viral replication.

Structurally constrained cyclic peptides were identified that bind to HIV-1 TAR RNA with low nanomolar (1-50 nM) affinity and specificity. The compounds are proteolytically stable, non-toxic, potent inhibitors of viral replication that act by a new dual mechanism involving interference with viral reverse transcription as well as transcriptional elongation. Using peptide mimetic chemistry and structure-based design, these peptides were further

developed to generate new inhibitors with increased binding affinity and cellular potency. The pharmacological potential of the lead structures was improved by incorporation of non-canonical amino acid side chains. My primary focus was determining the *in vitro* binding affinity, selectivity and specificity of a small library of improved peptides, followed by elucidation of the tertiary structure of the lead peptide-RNA complex. The lead peptide, JB-181, bound selectively to HIV-1 TAR RNA at 28 pM, was specific to HIV-1 TAR RNA in the presence of other competing, structurally similar RNAs, and its 1.6 Å, NMR resolved tertiary structure revealed more intimate contacts between the peptide's hydrophobic residues and the HIV-1 TAR RNA relative to previously designed peptides of its class.

My second focus was to evaluate the activity of a new class of peptide mimics – γ-AA peptides (based on the γ-PNA backbone) designed to target and inhibit the TAR-Tat interaction. These γ-AA peptides can project the same number of functional groups as peptides of equivalent length, suggesting that they could structurally mimic an RNA-binding protein. They can be modified with virtually limitless potential by introducing a wide variety of functional groups and are resistant to proteolytic degradation. A γ-AA peptide analogue of Tat residues 48 – 57 was developed and demonstrated to bind to HIV TAR RNA with low nanomolar affinity.

My third focus was to implement the *in silico* free energy perturbation (FEP) method to design new peptide inhibitors of HIV-1 transcription, particularly those inclusive of nonstandard amino acids. The first peptide-RNA complex modeled was the cyclic JB-181-TAR complex. It had an RMS value of 2.96 Å with the lowest energy NMR resolved structure of the peptide-RNA complex. The free energy of the FEP model (7.95 kJ) agreed well with the experimental free energy of binding (10.4 kJ). Despite successfully modeling the JB-181-TAR complex, the CHARMM force field, at the time, inadequately supported RNA molecules, and prevented us from further utilizing the method for the development of other JB-181 peptide derivatives. Recently, however, the CHARMM force field was updated to support RNA molecules, and was released to the public. Further studies with this method to help further determine which nonstandard amino acid modifications may provide for a better binding capacity against HIV-1 TAR RNA, by inspection of relative trends in free energy of binding, are currently underway.

The novelty of the dual inhibitory mechanism, the proteolytic stability, the development of a computationally dependent screening method, and the improved pharmacological activity exhibited by these antiviral leads provides a unique approach to antiviral targeting. Furthermore, by interfering with the maintenance of infection in viral reservoirs, such

compounds would be particularly attractive for use in combination treatment against HIV-1 strains resistant to current drug treatment.

University of Washington

Abstract

Structure of the conserved core region of the lincRNA Cyrano

Alisha Jones

Chair of the Supervisory Committee:

Professor Gabriele Varani

Department of Chemistry and Biochemistry

Two decades of genome sequencing have uncovered tens of thousands of biologically important non-coding RNAs with largely unknown function. Like other non-coding RNAs (ncRNA), long intervening noncoding RNAs (lincRNAs) do not encode proteins, but are clearly identified by capping and polyadenylation, unlike other long noncoding RNAs. They regulate key biological processes such as transcription and chromosomal inactivation, but how this occurs is unclear. The 4.5 kb lincRNA Cyrano was discovered in zebrafish and

regulates nasal and eye development during embryogenesis. While there is little sequence conservation from one species to the next across much of the 4.5 kb Cyrano transcript, a highly conserved 300-nucleotide region is found in mammalian and fish genomes, including human and mouse. Mutations in this conserved region cause developmental defects in zebrafish embryos in embryogenesis.

Using free energy minimization folding, comparative sequence analysis, and Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE), I determined the secondary structure of the 300-nucleotide conserved region of Cyrano. Secondary structure provides insight to how this RNA folds and functions during embryogenesis, albeit through RNA - RNA or protein – RNA interactions, as well as form the start for three-dimensional structural elucidation.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE12-56082.

I am extremely thankful for all of the undergrads and rotating graduate students who have helped me with my projects over the past five years. I am also very thankful for the support, guidance, and mentoring that I've received from Gab, Matt, and Karlotta. I don't think I would be where I am today without them.

DEDICATION

To my family and friends

TABLE OF CONTENTS

Part 1: Development and Evaluation of Peptide Mimetic Inhibitors of HIV-1 Replication

Chapter 1: Introduction: the HIV-1 TAR Tat Complex

RNA is increasingly being appreciated as a therapeutically attractive target for chronic and infectious diseases [1, 2], but it is notoriously difficult to identify pharmaceutically attractive lead molecules because of the pharmacological limitations of oligonucleotide analogues [3] and the limited potency and specificity of small RNA-binding molecules [4, 5]. As for inhibitors of protein-protein interactions, it is generally difficult to identify small molecules that inhibit large macromolecular interfaces. Structured peptides that mimic protein functional epitopes have been recently shown to be effective therapeutic agents toward proteins that have so far been "undruggable" using small molecule chemistry [6]. These molecules are particularly attractive against macromolecular targets, such as RNA, where small molecule chemistry is more difficult than for traditional enzymatic targets [7]. Improvements in affinity and specificity, provided by the much larger molecular interfaces of peptides, can compensate for the more difficult pharmacology of peptides.

The HIV-1 TAR RNA-Tat complex is one of the most extensively researched RNA-protein interactions because of its involvement in transcriptional activation and essential role in viral replication. After HIV-1 DNA has been integrated into the host's cellular genome, the virus enters a stage of latency, or with the aid of the cellular RNA polymerase II (RNAP II), proceeds with transcriptional elongation of HIV-1 RNA. Very shortly after the initiation of transcriptional elongation, two transcriptional replication inhibitors, NELF (negative elongation factors) and DSIF (DRB sensitivity-inducing factor), stall RNAP II to prevent further transcription of viral HIV-1 RNA. The delay from the initiation of transcription to the stalling of RNAP II allows for the

transcription of 59-nucleotide HIV-1 RNA, called the transactivator response element (TAR). This RNA is found at the 5' end of all nascent viral transcripts.

A virally encoded transcriptional activator, Tat, and its cellular co-factor, the transcription elongation factor-b (PTEF-b), specifically binds to the nascent TAR RNA [8, 9]. This stable ternary complex is then recruited to the stalled RNAP II where the activated CDK9 kinase component of PTEF-b phosphorylates the carboxyl terminal domain (CTD) of RNAP II, the NELF and the DSIF, to promote transcriptional elongation of the HIV-1 RNA [10] (Figure 1).

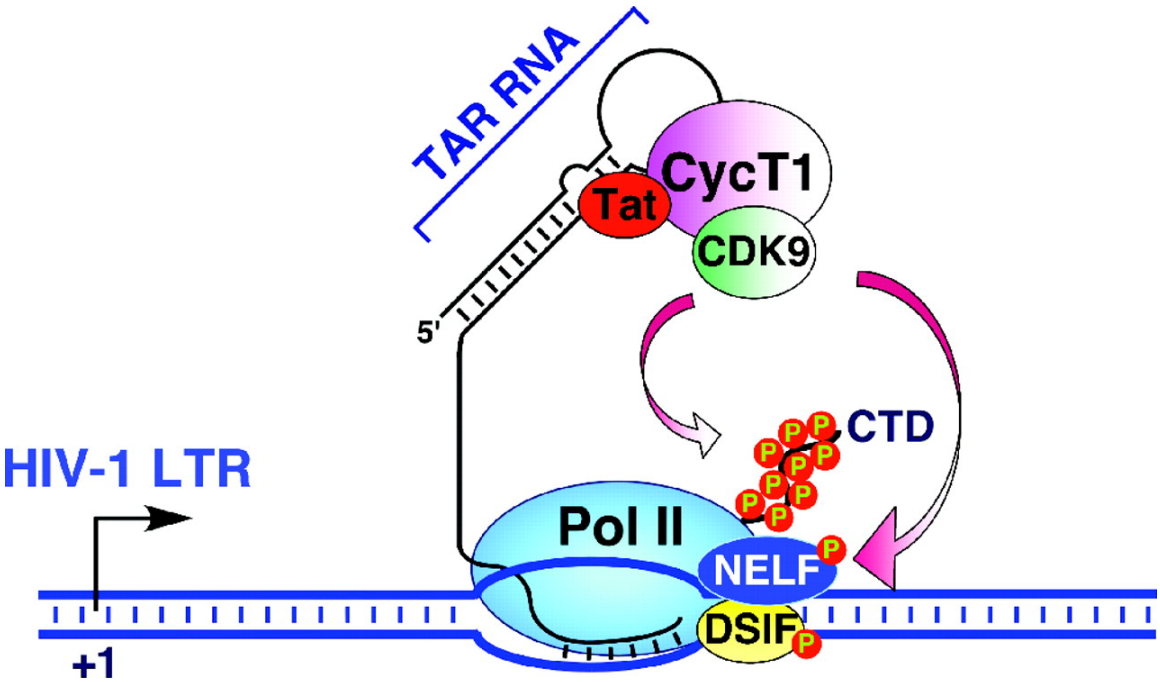


Figure 1: Schematic illustration of the tertiary complex formed by HIV-1 TAR RNA, HIV Tat protein, cyclin T1 and CDK9. The complex is recruited to the stalled Pol II, and via phosphorylation, inactivates NELF and DSIF, and activates elongation through the CTD [10].

While Highly Active Antiretroviral Therapy (HAART) allows for effective postponement of HIV-1 progression by eliminating most circulating viruses through drugs that target reverse transcription, integration, or HIV-1 fusion and entry [11], this combination has two drawbacks. First, it is unable to target latent viral reservoirs in patients, allowing emerging viruses to rapidly repopulate the body when therapy is interrupted [12]. Second, upon treatment, there is an emergence of mutations causing the development of new subtypes of viral RNA [12]. Thus, the effectiveness of antiretrovirals is dependent upon its ability to inhibit a diverse range of HIV-1 isolates as well as blocking the emergence of the virus from latently infected cells.

Because transcriptional elongation of HIV-1 is dependent on the TAR-Tat interaction [9, 10, 13], I hypothesized that HIV-1 TAR RNA is highly conserved. An alignment of 3,122 HIV-1 TAR RNA isolates (retrieved from the Los Alamos database: http://www.hiv.lanl.gov/content/index) spanning a wide range of subtypes as listed in the figure below, revealed 91.1% pairwise percent identity of HIV-1 TAR RNA (Figure 2).

45_cpx	D	O	06_cpx	19_cpx
A	F	46_BF	08_BC	20_BG
A1	F1	01_AE	09_cpx	26_U
A2	G	02_AG	10_CD	27_cpx
B	H	03_AB	11_cpx	32_06A1
C	N	04_cpx	12_BF	unclassified

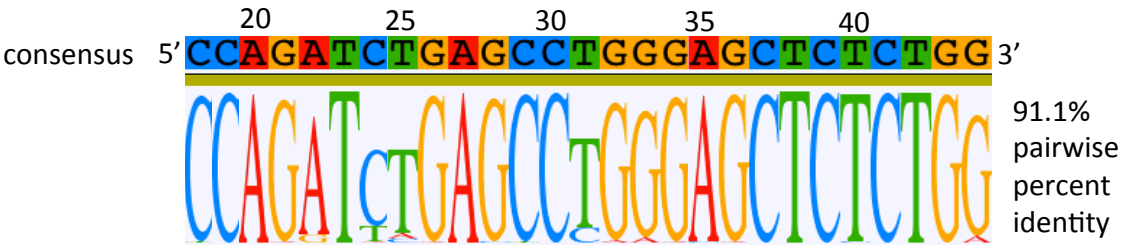


Figure 2: HIV-1 TAR is strongly conserved. The enclosed box lists all of the subtypes of HIV-1 included in the alignment of 3,122 TAR sequences. The consensus logo reflects the alignment of the HIV-1 TAR isolates; there is 91.1% pairwise percent identity of the isolates relative to the consensus sequence.

Given the 91.1% TAR RNA conservation, targeting the binding domain of TAR is therefore a promising target for the development of new antiviral agents through the disruption of the TAR-Tat interaction. This would inhibit viral replication at both latent and active stages of infected cells, despite the emergence of mutations.

The pharmacological inhibition of the TAR-Tat interaction has be pursued by a number of academic laboratories and pharmaceutical companies over the last 20 years with only limited success. Although a number of early TAR RNA-binding compounds can inhibit HIV transcription, the majority of the antiviral activity attributed to these types of compounds was due to off-target inhibition of cell entry by highly cationic compounds binding to the HIV envelope [14-17]. Small molecules that were evaluated as TAR RNA-directed antivirals have also proved to be only

partially active and relatively non-specific including analogs of existing aminoglycoside antibiotics or DNA-binding structures, and compounds discovered in proprietary libraries [1, 4, 5, 18-25]. The most potent and lowest molecular weight series of small molecules capable of blocking the TAR-Tat interaction were a set of biaryl compounds identified at RiboTargets. These compounds all had molecular weights well below 500 Da and were able to block TAR-Tat interactions in the 50 nM range, but unfortunately, they entered cells poorly and had only modest antiviral activities.

A key innovation underlining the success of blocking the TAR-Tat interaction are cyclic peptides that mimic the conformation of the native Tat protein bound to TAR; these have been demonstrated to be potent inhibitors of this critical RNA-protein interaction [26]. The use of peptidomimetics has been proven to be effective for discovering new inhibitors of protein-protein and, more recently, protein-DNA interactions using constrained stapled peptide mimics of alpha helices.

In the next three chapters, I will describe three approaches I utilized to further develop and evaluate peptidic inhibitors of HIV-1 replication via the TAR-Tat interaction. This includes peptoid chemistry, structure based design, and *in silico* free energy perturbations.

Chapter 2: γ-AApeptides as Peptidic Inhibitors of the HIV-1 TAR Tat Interaction

An abbreviated version of this chapter appeared in Organic Biomolecular Chemistry (2011).

Section 1: Introduction

In order to develop inhibitors of TAR-Tat interaction, significant efforts have been dedicated to synthesize and evaluate short peptides that can mimic Tat protein and disrupt Tat binding to TAR [5, 26-33]. Among them, oligopeptidomimetics such as oligocarbamates, oligoureas, β-peptides, peptoids and templated cyclic peptides were considered, since these structures are resistant to proteolytic degradation [26, 32, 34-36]. However, more than a decade's exploration has not led to any clinical candidates, in part because a structure of the HIV-1 TAR RNA-Tat complex remains to be determined, due to its highly dynamic conformation [37]. Recently, the Cai research group developed a new class of peptide mimics – γ-AApeptides [38], based on the γ-PNA backbone [39]. These γ-AApeptides can project the same number of functional groups as peptides of equivalent length, suggesting that they could structurally mimic an RNA-binding protein. They can be modified with virtually limitless potential by introducing a wide variety of functional groups and are resistant to proteolytic degradation [38]. Their potential biomedical application has been demonstrated by their capability to disrupt the p53-MDM2 protein-protein interaction [38]. To further explore the applications of γ-AApeptides, we demonstrate here that a γ-AApeptide analogue of Tat 48–57 can bind to HIV TAR RNA with nanomolar affinity. The results indicate that γ-AApeptides are valid peptide mimics of RNA binding proteins, and they can potentially be further developed to modulate RNA-protein interactions.

The arginine-rich segment of HIV-1 Tat (residues 48–57) makes direct contacts with the TAR trinucleotide bulge region, and is the key determinant of the TAR-Tat interaction [37]. Many oligopeptidomimetic inhibitors have been designed based on this fragment of Tat [26, 30, 33, 35, 36, 40, 41]. Since HIV Tat 48–57 adopts an extended conformation [42], we hypothesized that γ-AApeptide γ-**AA1** (Figure 3) would be able to mimic HIV Tat 48–57. γ-**AA1** and Tat 48–57 have identical molecular weight and project exactly the same functional groups; the relative positions of these functional groups are similar to each other when the peptide conformation is extended. To test this hypothesis, the Cai group synthesized γ-AApeptide γ-**AA1** and γ-AApeptide γ-**AA2** (a truncated sequence mimicking Tat 48–53). A control HIV Tat 48–57 peptide, **P1,** was also prepared for comparison.

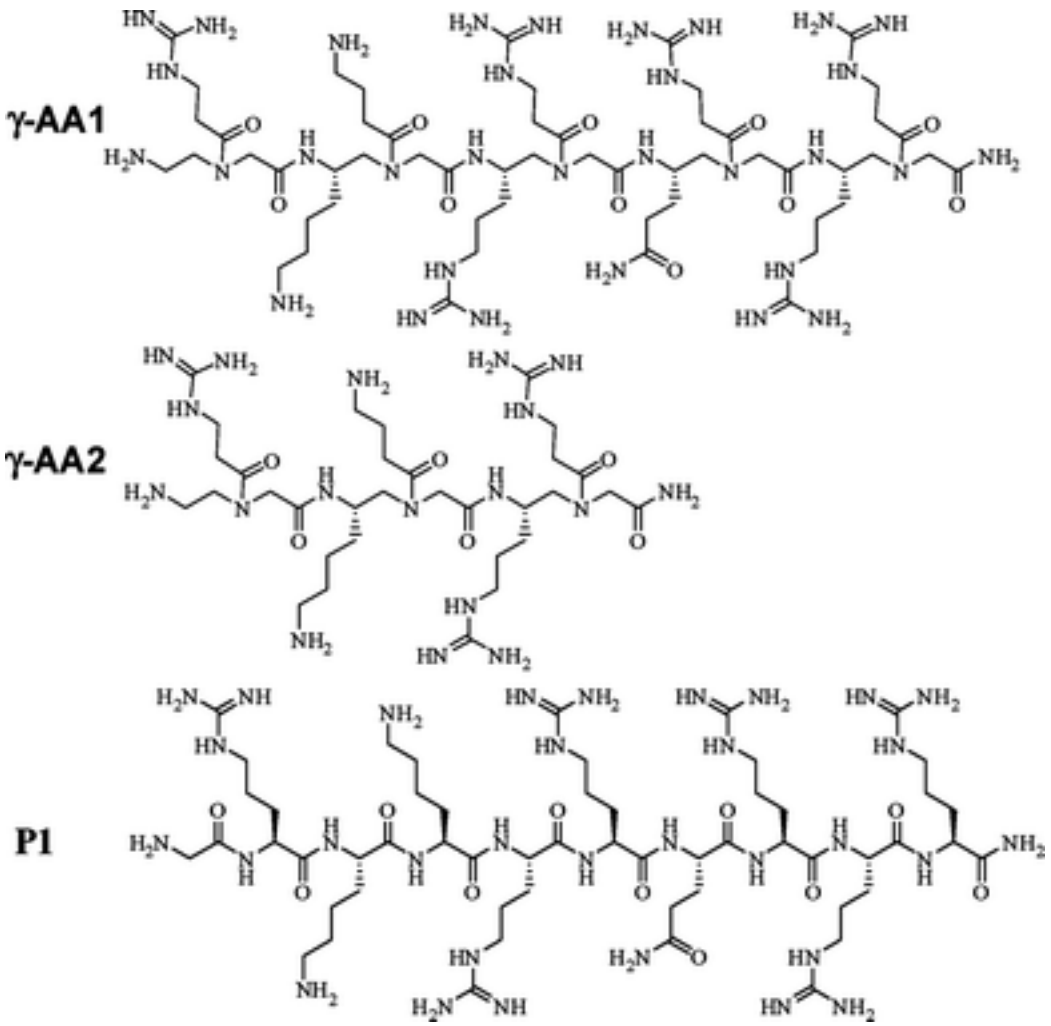


Figure 3: γ-AApeptides γ-AA1 and γ-AA2, and a control, P1 (HIV-1 Tat 48-57)

Section 2: Synthesis of γ-AApeptides

The synthesis of γ-AApeptides was carried out by the Cai research group using manual solid-phase synthesis from Fmoc-protected γ-AApeptide building blocks, a method recently developed by their group to synthesize AApeptide sequences [38, 43]. A Fmoc protected amino aldehyde **1** reacted with benzyl glycinate to form secondary amine **2**, which was acylated by either γ-Boc-amino butyric acid or di-Boc-guanidinopropionic acid to give **3** (Figure 4). Subsequent hydrogenation provided Fmoc protected γ-AApeptide building blocks **4**. These γ-AApeptide building blocks were assembled on solid phase, and the desired sequences were cleaved from the solid support, purified by HPLC and characterized by MALDI.

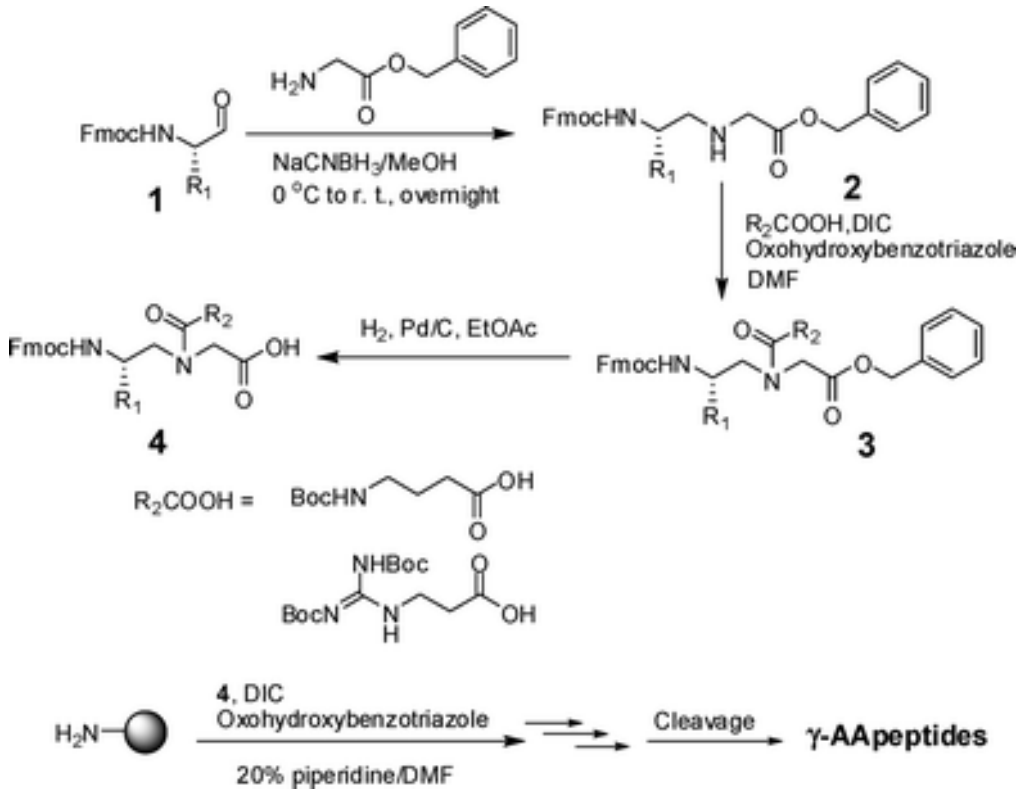


Figure 4: Synthesis of γ-AApeptide building blocks of γ-AApeptides

Section 3: Results and Discussion

To investigate whether γ-AApeptides could mimic the Tat 48–57 peptide **P1**, I tested their binding to HIV-1 TAR RNA by measuring dissociation constants, K_D's, using EMSA (electrophoretic mobility shift assay); the closely related BIV TAR RNA (bovine immunodeficiency virus) was used as a control for specific binding. Both HIV-1 and BIV are lentiviruses, and the functions of Tat and TAR are conserved; the sequences and secondary structures of HIV-1 and BIV TAR are also highly similar (Figure 5). Because of these similarities, we anticipated that these Tat γ-AApeptide mimetics should bind tightly to both HIV-1 and BIV TAR by recognizing their TAR-Tat binding regions.

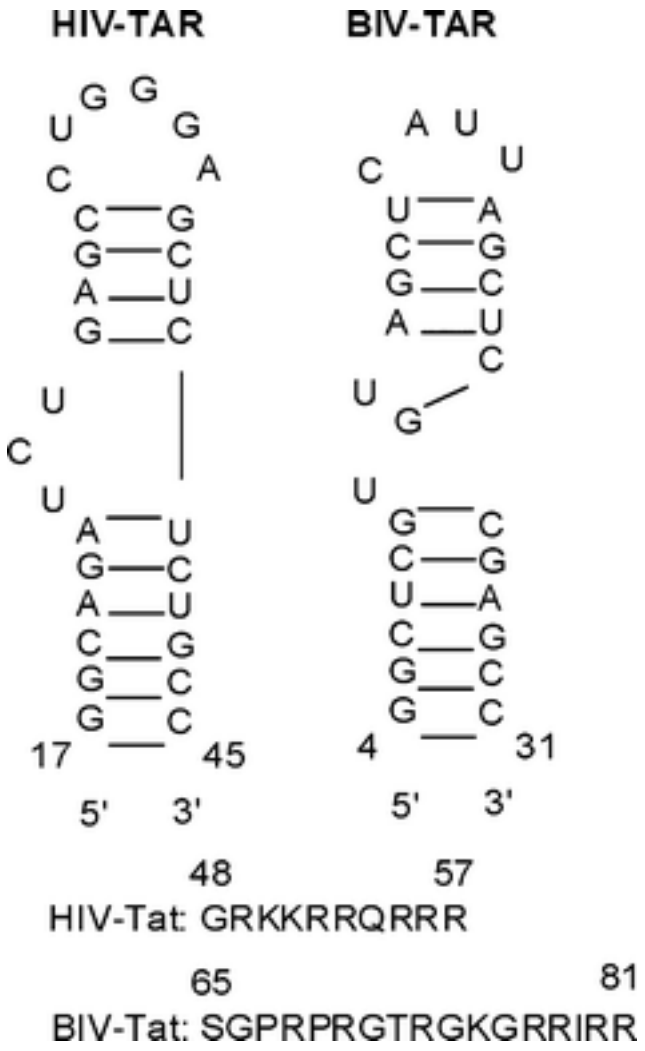


Figure 5: Secondary structures of HIV-1 and BIV TAR RNA, and partial sequences of HIV-1 (48–47) and BIV Tat protein (65–81); these fragments are largely responsible for the interaction with their respective TAR RNAs.

The EMSA experiments demonstrated that γ-AApeptide γ-**AA1** (Figure 6) binds to both HIV-1 TAR and BIV TAR RNAs with nanomolar affinity. Furthermore, binding in the low μM regime is retained even in the presence of 25,000-fold excess of tRNA (Figure 6), suggesting that the interaction is specific for the TAR RNA structures, in that even large amounts of excess tRNA fails to completely abolish complex

formation. Binding affinities are listed in Table 1 and compared with those of γ-AApeptide γ-**AA2** and Tat 48–57 peptide, **P1**, which were also obtained by EMSA.

Table 1: Summary of the affinity of different peptide and peptide mimetic sequences for their interaction with HIV-1 and BIV TAR, as determined by EMSA.

Mimic	**K_D (HIV), nM**	**K_D (BIV), nM**
AA1	166 ± 85	300 ± 25
AA2	>33000 (smeared band)	>33000 (smeared band)
P1	166 ± 85	333 ± 41.5

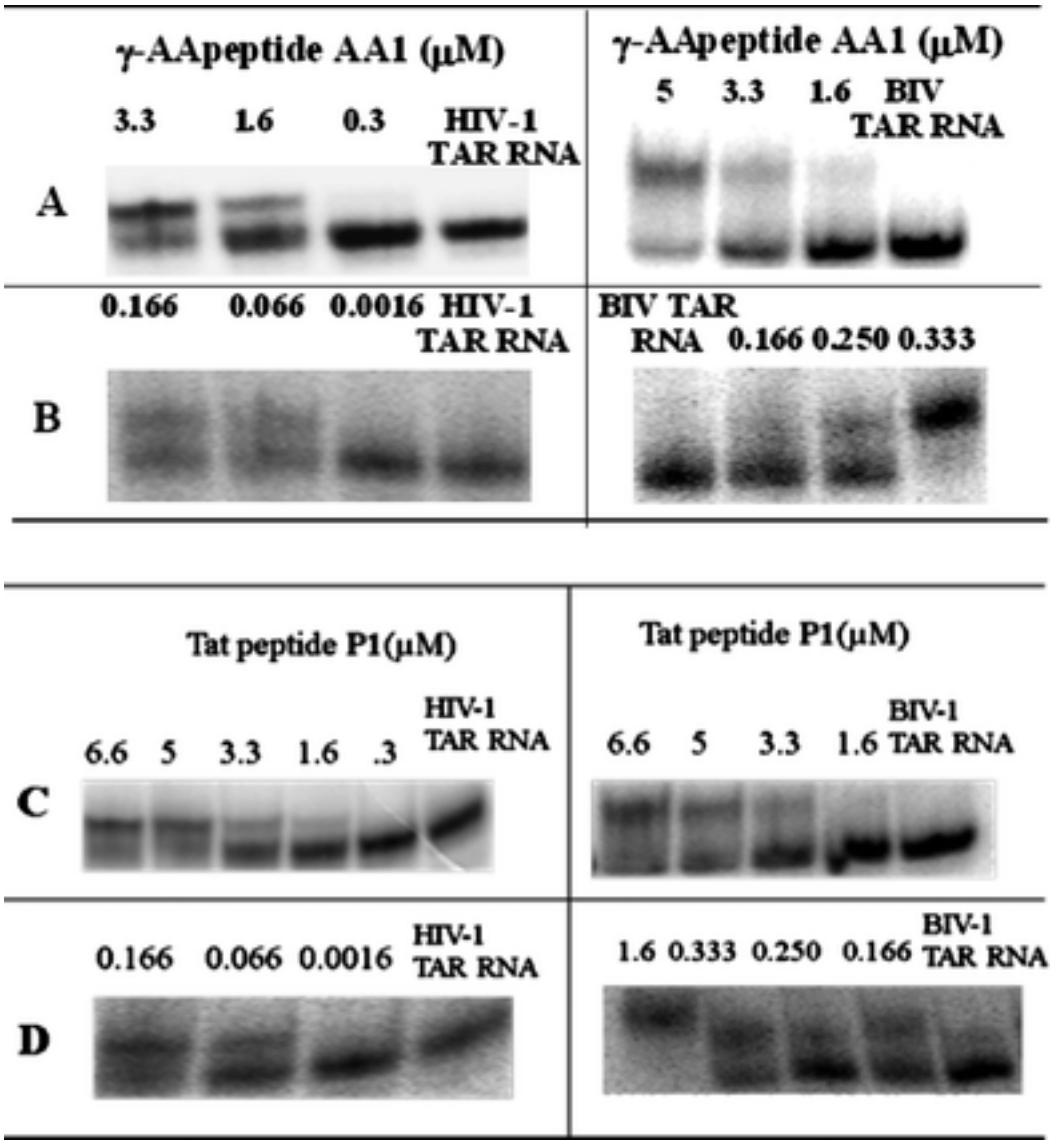


Figure 6: Binding of γ-AApeptide γ-**AA1** to HIV-1 and BIV TAR RNAs assayed by EMSA. A) Binding of γ-AApeptide 1 to HIV-1 (0.4 nM) and BIV TAR (0.4 nM); the buffer contains a 25,000-fold excess of tRNA to reduce non-specific binding. B) Binding of γ-AApeptide γ-**AA1** to HIV-1 (0.4 nM) and BIV TAR (0.4 nM) RNAs; the buffer contains a smaller excess of tRNA (250-fold) to measure K_D's more accurately. C) Binding of Tat peptide **P1** to HIV-1 (0.4 nM) and BIV TAR (0.4 nM); the buffer contains a 25,000-fold excess of tRNA to reduce non-specific binding. D)

Binding of Tat peptide **P1** to HIV-1 (0.4 nM) and BIV TAR (0.4 nM) RNAs; the buffer contains a smaller excess of tRNA (250-fold) to measure K_D's more accurately.

K_D values shown in Table 1 were calculated from data as shown in Figure 6, B and D, where a 250-fold excess of tRNA was present. In the absence of tRNA, band smearing was observed. This smearing may occur due to partial dissociation of complexes during gel electrophoresis, as well as to the likely presence of non-specific, lower affinity complexes between the peptides and the TAR RNAs, which were reduced or eliminated in the presence of the tRNA, resulting only in the formation of a specific peptide-TAR complex. The EMSA results show that γ-AApeptide γ-**AA1** binds to HIV TAR with K_D of 166 nM, 2-fold more tightly than binding to BIV TAR (K_D = 300 nM). As expected, Tat peptide **P1** can bind to HIV and BIV TAR tightly with comparable K_D's of 166 nM and 333 nM, respectively. Interestingly, γ-AApeptide γ-**AA2**, although carrying a few comparable positively charged side chains, failed to provide any binding capability to both HIV and BIV TAR under our experimental conditions. The EMSA results show that γ-AApeptide γ-**AA1** binds to HIV-1 and BIV TAR RNAs as tightly as the Tat-derived **P1** peptide 48–57, even if the γ-AApeptide backbone is more flexible than the conventional peptide backbone.

Although structural information is not yet available for the complete HIV-1 TAR RNA-Tat complex, we previously used NMR to investigate the conformational change of HIV TAR when binding to Tat and other small molecules and peptides [40, 44, 45]. In the presence of Tat and other ligands, the bulge region of TAR undergoes a local conformational rearrangement and forms a more stable structure (Figure 7). This folding process can be induced by any ligand containing a guanidinium group and even by the single amino acid analogue argininamide. However, the interaction of

this guanidinium group with TAR is not the only source of binding affinity and specificity for Tat recognition. NMR studies demonstrated that there are multiple points of contacts between base functional groups and phosphate groups of HIV TAR and amino acid residues of Tat. These interactions contribute not only to the affinity of the interaction, but also to its specificity.

Based on the experimental results, we postulate that γ-**AA1** binds to HIV TAR by mimicking the Tat peptide **P1**, because they possess the same side functional groups. Following the mechanism of TAR-Tat interaction, γ-**AA1** could recognize the bulge of HIV TAR using one of the guanidinium groups, which would re-fold the HIV TAR bulge and define the precise positioning of critical functional groups in the major groove. Compared to α-peptides, half of the C═O groups have been relocated to side chains, which leads to increased conformational freedom for the γ-AApeptide backbone. As a result, γ-**AA1** may be able to more easily adjust its conformation in the TAR- γ-AApeptide complex by mimicking that of the Tat peptide so as to achieve optimal binding [37].

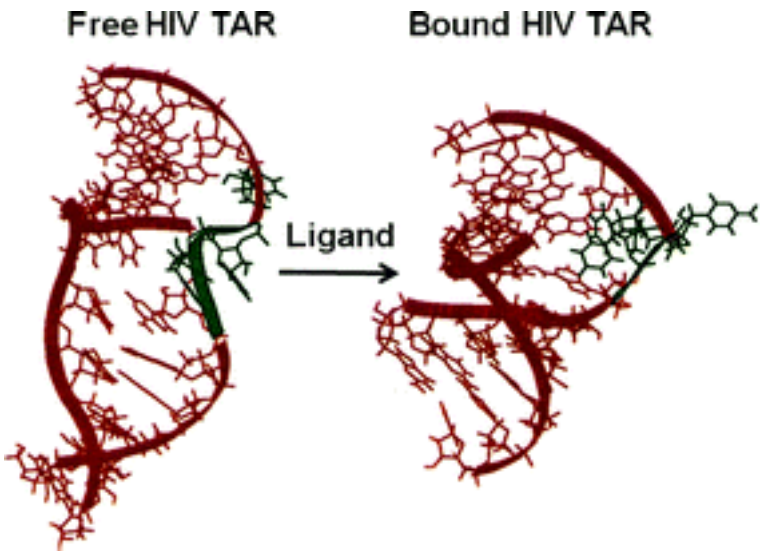


Figure 7: Schematic representation of the mechanism of HIV-1 TAR RNA-Tat recognition (adapted from reference [44]). Binding of Tat re-folds the bulge of TAR into a locally different conformation.

Satisfactorily, the truncated γ-AApeptide γ-**AA2** has completely lost its binding capability to both HIV-1 and BIV TAR RNAs, strongly supporting the vitally important presence of three neighboring guanidino functional groups for binding. This result also further suggests that the TAR- γ-AApeptide interaction is not purely driven by electrostatic interactions, since the truncated γ-AApeptide γ-**AA2** retains as many guanidino functional groups as small molecules that were shown to bind to TAR with nM affinity [5]. The sequence is highly positively charged, yet it shows no interaction with TAR RNA in the presence or absence of tRNA. This result indirectly shows the importance of multiple points of interactions between HIV TAR and γ-**AA1**, similar to the HIV TAR-Tat interaction. It is also noteworthy that there is a small binding preference of γ-**AA1** for HIV-1 compared to BIV TAR (2-fold), which may be due to γ-**AA1** being the mimic of HIV-1 Tat protein and not of BIV Tat. Altogether, these

observations provide a starting point for the rational design of more potent and selective RNA-binding γ-AApeptides in the future.

Section 4: Conclusions

Our collaborators have developed a new peptide mimetic structure, the γ-AApeptides, which mimics the Tat peptide, and I have evaluated their nanomolar binding affinity to TAR RNAs. Our findings suggest that γ-AApeptide structures are capable of binding to RNAs by mimicking RNA-binding proteins. These structures can be developed further to probe or disrupt RNA-protein interactions in the future. This is a further demonstration of the promising biological activity of γ-AApeptides, that have already been shown to disrupt protein-protein interactions [38, 43]. Due to their resistance to proteolysis, convenient synthesis and limitless diversification, there is potential for γ-AApeptide **AA1** to be further optimized to identify new anti-viral leads.

Although the γ-AA peptides demonstrated low nanomolar binding with HIV-1 TAR RNA and modest specificity relative to BIV TAR RNA, it is well known that linear structures are very flexible, a characteristic leading to a reduction in selectivity and specificity of their targets. This was demonstrated in the *in vitro* and *in vivo* studies of HIV-1 RNA with CGP 642222 (Figure 8), a peptoid designed to mimic the binding domain of Tat [36].

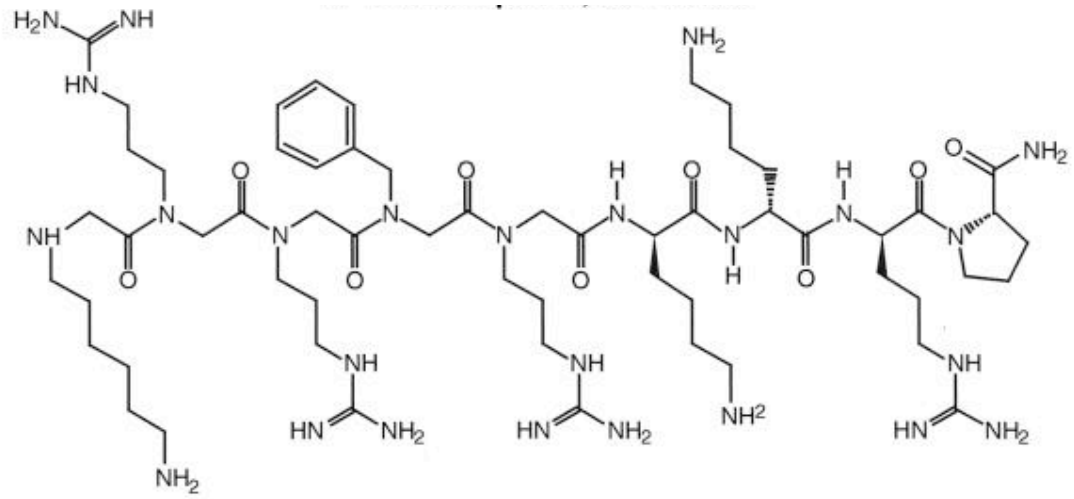


Figure 8: Peptoid CGP 64222 designed to mimic the Tat binding site for TAR RNA.

Despite nanomolar binding to HIV TAR, due to a lack of rigidity, this peptoid was not able to selectively and specifically target HIV-1 RNA.

The same result was evident when EMSAs could not be successfully performed sans tRNA with the γ-AA peptides. As seen in the EMSA gel images, clearly resolved complex bands were only visible in the presence of 25K fold excess tRNA. The tRNA competes away weak complexes for HIV-TAR RNA (hence the weaker binding relative to the binding observed in the presence of 250 fold excess tRNA), resolving the complex bands. Future development of these peptide mimetics could involve cyclization of the backbone to force them into more rigid structures (Figure 9).

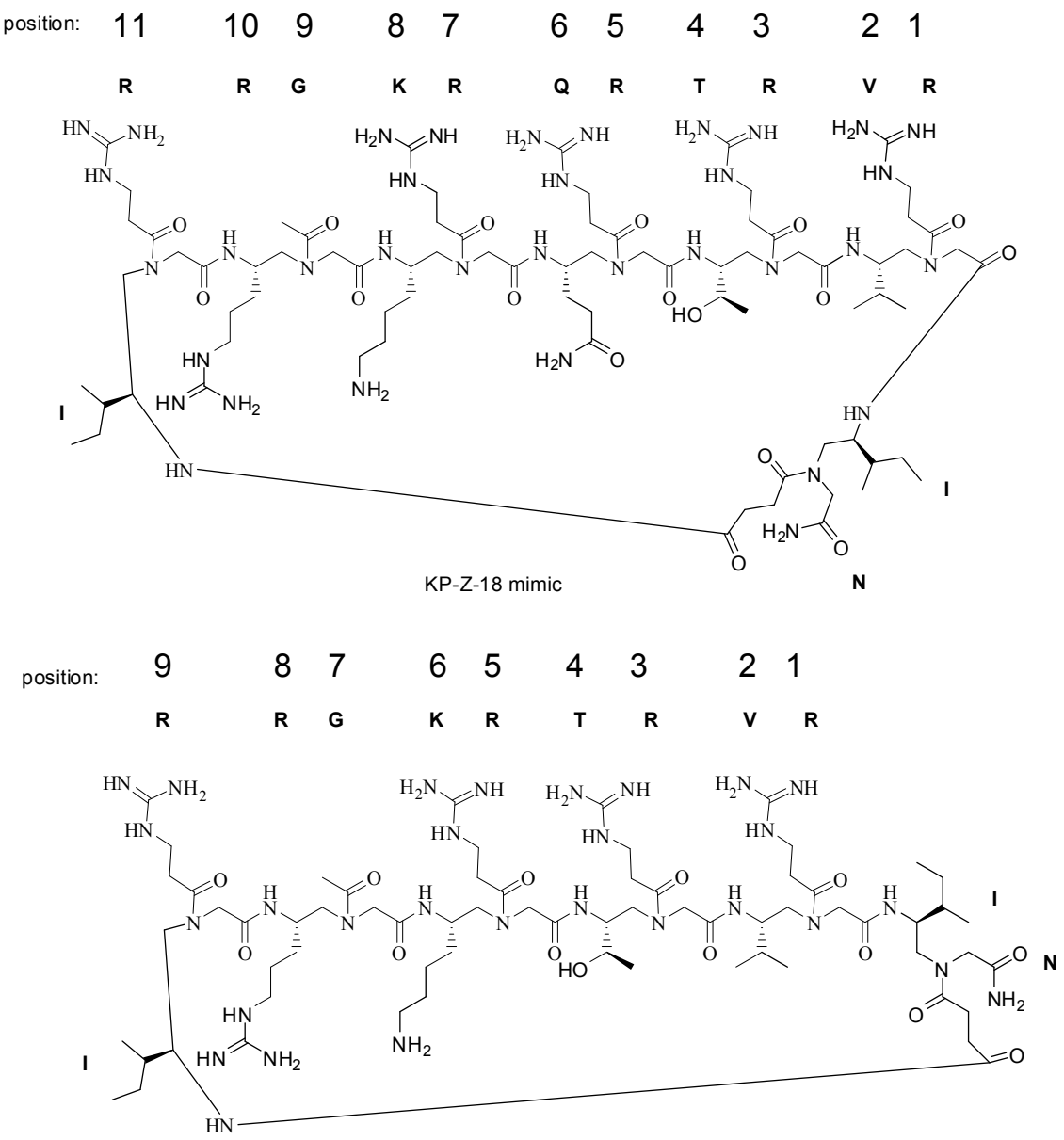


Figure 9: Two proposed cyclic peptide mimics with γ-PNA backbone. They are derived from KP-Z-18 and L-22, two cyclic peptides that bind to HIV-1 TAR RNA.

These proposed peptides are derived from L-22 and KP-Z-18 [26, 40], two cyclic peptides that bind to HIV-1 TAR with binding constants of 30 nM and >5000 nM, respectively. These two mimics would serve as a control for testing the design, and

further development of the sequences would contain other functional groups that improve the binding affinity of the peptide mimics with HIV-1 TAR RNA.

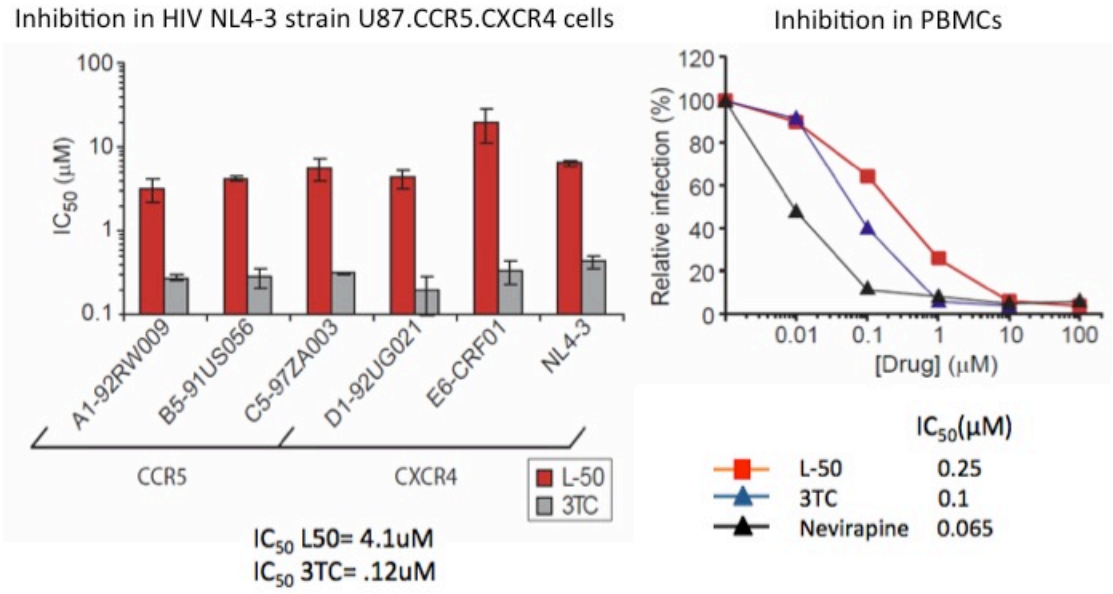
Chapter 3: A Highly Specific Picomolar Ligand for HIV-1 TAR RNA

A manuscript describing this work is pending publication

Section 1: Introduction

Linear peptide analogues of Tat have been previously used to target HIV-1 TAR RNA and inhibit Tat binding [36, 46]. However, linear peptide analogues were highly flexible, and bound to a wide range of RNAs nonspecifically, preventing the development of clear structure-activity relationships. It was hypothesized and confirmed that conformationally constraining the peptides improves their binding capacity to HIV-1 TAR RNA [26].

Small β-hairpin mimics of 12 residues, stabilized by a heterochiral D-pro-L-pro dipeptide template that strongly favors a type-II β-turn backbone conformation and therefore formation of a β-strand, were designed to mimic the binding domain of BIV-1 Tat [47, 48]. In the absence of a structure for the HIV TAR-Tat complex, the cognate Bovine Immunodeficiency Virus (BIV) TAR-Tat interaction was mimicked [27, 49, 50], for which a structure exists. The most potent peptides generated after a few rounds of synthesis, testing, and rational re-design exhibited low nanomolar *in vitro* binding activity (1-30 nM) against HIV TAR, and, when subjected to a variety of antiviral and mechanistic studies, were shown to penetrate cells readily and to inhibit Tat-dependent transactivation in primary lymphocytes against a variety of isolates that were both CCR5 and CXCR4 tropic (Figure 10) [26, 29]. The antiviral activity of these peptides in primary blood lymphocytes was 10-fold less potent than the FDA approved antiviral drug nevirapin (Figure 10), but arose from a completely different mechanism. The peptides inhibited, at the same time, Tat-dependent transcriptional elongation as well as initiation of reverse transcription (Figure 11) [29].



Mimetic	Position 1	2	3	4	5	6	7	8	9	10	11	12	$K_{d\ (nM)}$ (HIV)	$K_{d\ (nM)}$ (BIV)
L-22	R	V	R	T	R	K	G	R	R	I	R	I	30	5
L-50	R	V	R	T	R	G	K	R	R	I	R	R	1	1
L-51	R	T	R	T	R	G	K	R	R	I	R	V	5	50

Figure 10: (Top) L-50 is a potent inhibitor of replication across divergent HIV-1 subtypes. It has an IC_{50} of 4.1 μM, just ten fold less than the FDA approved nucleoside reverse transcription inhibitor 3TC. In peripheral blood mononuclear cells, the IC_{50} of L-50 was improved to 0.25 μM, and is just ten fold less than 3TC. (Bottom) Three lead peptide sequences.

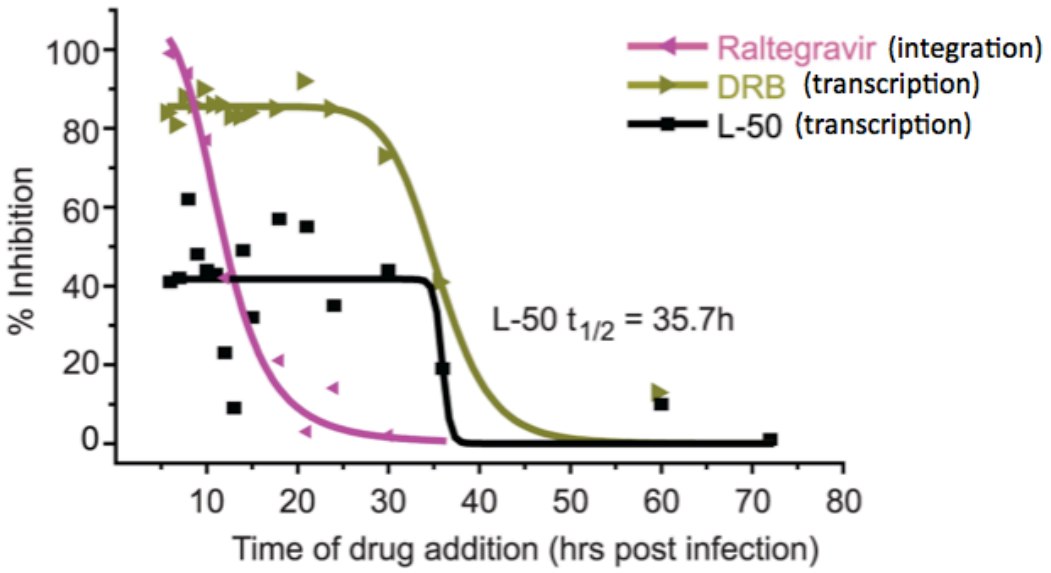


Figure 11: L-50 inhibits after integration, similar to Raltegravir, an integrase inhibitor, and DRB, an inactivator of the positive elongation factor-b (PTEF-b).

The peptides reported so far bound to RNA potently, but did not discriminate between related RNAs well. They discriminate 10,000 fold against tRNA, but bound with comparable potency to HIV and BIV TAR [25, 40]. We hypothesized this characteristic limits their potency in cells. In order to further improve the activity of this promising class of antivirals, we examined the structure of the complex between peptide L-22 (cyclo-R V R T R K G R R I R I dP P) and HIV TAR RNA [26] to further improve its activity, under the hypothesis that if binding of the L-22 peptide to HIV TAR RNA could be improved, potency would increase accordingly.

We now report the discovery, within a small, highly focused peptide library, of very potent, low pM and highly selective inhibitors of the HIV-1 TAR RNA-Tat interaction designed using structure-based principles. The increased binding activity and specificity leads to improvements in antiviral potency, which is now comparable to AZT. The activity, specificity and cellular potency of these peptides are unprecedented amongst molecules that bind to RNA and suggest that this chemistry

could have broad applications with other potentially attractive, but so far undruggable, RNA drug targets.

Section 2: Results

Structure-based design of side chain substitutions – We previously reported the discovery of three peptides, called L-22, L-50 and L-51, which bound to HIV TAR with nM activity and inhibited HIV replication by a TAR-dependent mechanism. To further improve its binding activity, under the hypothesis that this would also improve the antiviral potency, we examined the structure of the L-22-TAR complex [26] and generated a new library of 100 peptides containing single and multiple peptide substitutions. However, none of these substitutions, based on the 20 traditional amino acid side chains, improved the binding activity significantly. When we re-examined the structure and structure-activity relationships, we noticed that two arginine side chains, Arg1 and Arg11, appeared to be too long to form optimal interactions with the RNA while retaining good stereochemistry with extended, nearly planar side chains (Figure 12).

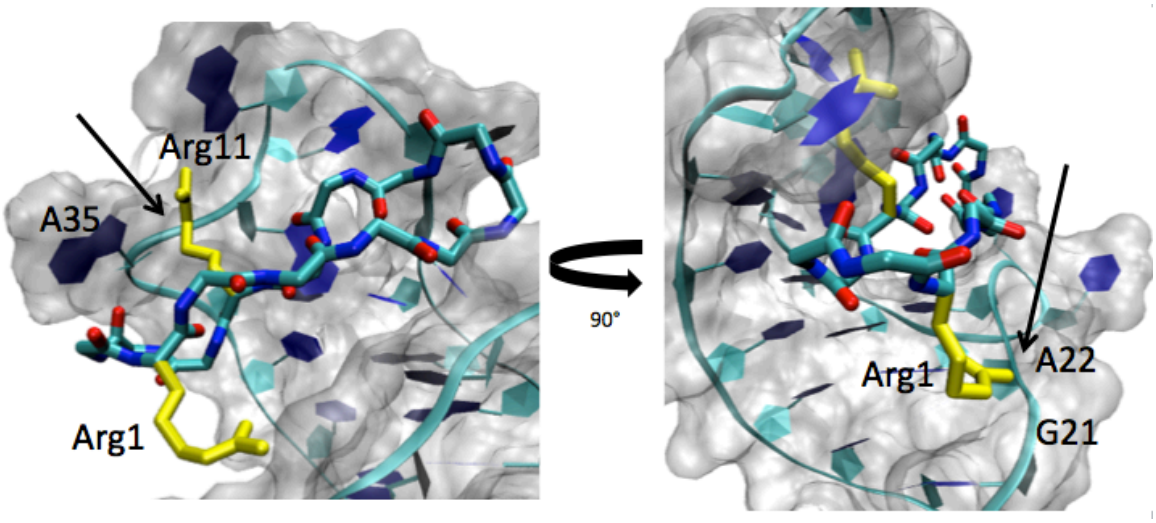


Figure 12: Two amino acid side chains (Arg1 and Arg11) form non-optimal contacts with the TAR RNA in the structure of the L-22 peptide-RNA complex. Arg11 prevents the peptide from sitting deeper into the groove by lying along the backbone of nucleotide A35, while Arg1 also blocks the peptide from sitting deeply into the major groove; it lies against the phosphate backbone between G21 and A22.

As a consequence, the core of the peptide was 'pushed' away from the RNA, causing multiple sub-optimal interactions between the two molecules. Thus, we introduced non-standard amino acids which carried guanidinium (noR, nor-arginine) or amines (Dab, diamino butyric acid and O, ornithine) groups at the top of the side chain shortened by one (or two in the case of ornithine) methylene group within a small library of peptides (nine) carrying substitutions at these two positions.

Evaluation of binding affinity and specificity - Peptide binding to HIV-1 TAR and other structurally similar RNAs (as controls for specificity) was evaluated using both electrophoretic mobility shift assays (EMSAs) and a 2-amino purine fluorescence-detected binding assay [51]. This second assay was introduced when we observed that the activity of some of the peptides was below the practical detection limit of EMSA (about 1 nM; we estimate the lowest concentration of radioactively labeled RNA that can be accurately quantified to be a few hundred picomolar, although lower quantities can be detected). Thus, we substituted nucleotide C24 of HIV-1 TAR RNA with the highly fluorescent 2-amino purine nucleotide analog (2-AP), which provides a sensitive probe for peptide interactions with HIV TAR. Titrating the peptides into the fluorescently labeled RNA causes a change in the conformation of the bulge, which leads to a change in 2-AP fluorescence. Previous studies identifying base positions in HIV TAR RNA that are essential for high-affinity Tat peptide/argininamide, and evaluation of the structure of L-22 complexed with HIV TAR demonstrated that modifying C24 to 2-AP would be non-perturbing. However, to further validate this second assay, we evaluated the binding affinity of three peptides by both methods (Table 2).

Table 2: Comparison of dissociation constants observed using EMSA and 2-AP fluorescence assays for three cyclic peptides JB185, JB186 and L-50.

Peptide	EMSA (nM)	Fluorescence (nM)
JB-186	0.3 – 1.6	1.17
JB-185	0.3 – 1.6	0.4
L-50	0.5 – 1.5	0.56

The most potent peptide, JB-181 (cyclo-Dab V R T R K G R R I noR I dP P) was found to bind at 28 ± 4 pM, while other peptides in this library demonstrated binding affinities between low picomolar (JB-185) and low nanomolar (JB-186, JB-58 and JB-59, JB-190) (Figure 13, Table 3).

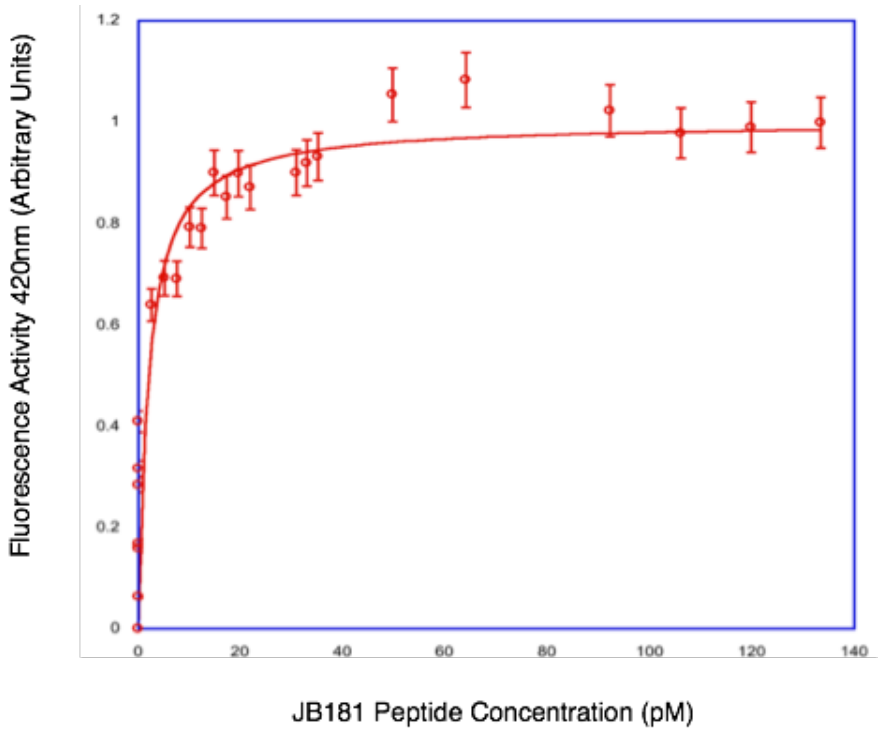


Figure 13: Results of a titration of JB-181 peptide into 2-AP HIV TAR RNA. Fluorescence intensity (thousands of counts per second) emitted at 421 nm is plotted against peptide concentration (pM). Red dots represent data points with 5% error bars. The relative standard deviation for all titration experiments was ~22.0%. Data points that fall outside of the smooth line reflect experimental error (poor mixing/not enough time for equilibrium to be reached). The smooth curve shows the line of best fit through the points giving a binding affinity of 27.5 ± 3 pM.

Table 3: Binding affinity for a small library of peptide derivatives of L-22 binding to HIV-1 TAR RNA as determined by fluorescence assays. Asterisks indicate data collected only with EMSA.

Peptide	1	2	3	4	5	6	7	8	9	10	11	12	K_D (HIV), nM
L-22*	R	V	R	T	R	K	G	R	R	I	R	I	30.0 ± 5.00
JB-181	Dab	V	R	T	R	K	G	R	R	I	noR	I	0.028 ± 0.004
JB-182	Dab	V	R	T	Q	K	G	R	R	I	I	I	>5000
JB-184	Dab	V	R	T	Q	K	G	R	R	V	noR	V	>5000
JB-185	Dab	V	R	T	R	K	G	R	R	I	R	L	0.45 ± 0.1
JB-186	Dab	V	R	T	R	K	G	R	R	I	noR	V	1.2 ± 0.16
JB-58*	O	V	R	T	R	K	G	R	R	I	R	I	3.00 ± 0.3
JB-59*	Dab	V	R	T	R	K	G	R	R	I	R	I	1.00 ± 0.3
JB-190	Dab	V	R	T	R	G	K	R	R	I	noR	I	1.56 ± 0.52

In addition to the introduction of nonstandard amino acids, the K-G turn of L-22 was changed to a G-K turn in one of the new peptides, JB-190. Other peptides synthesized in the L-22 library containing a G-K instead of a K-G turn (L-50 and L-51) (along with 1 or 2 amino acid substitutions) had significant improvements in binding activity with HIV TAR (Figure 10). We hypothesized a G-K turn may also drastically improve the binding capacity of the peptides containing nonstandard amino acids with HIV-1 TAR, however binding remained in the low nanomolar range at 1.56 nM with JB-190. It is possible that it was not the G-K turn that made L-50 and L-51 better binding partners to HIV-1 TAR, but the substitutions present in positions 2 and/or 12.

Two peptides, JB-182 and JB-184 did not bind to HIV-1 TAR RNA at concentrations

lower than 250 nM. We hypothesize that an arginine in position 5 is required for the stability of TAR's structure in the major groove, primarily around the base triple (U23, A27, U38). This will be further discussed in Chapter 4. Thus, we have identified ligands that bind to RNA with unprecedented pM activity whose affinity for TAR is very sensitive to even small changes in side chain identity.

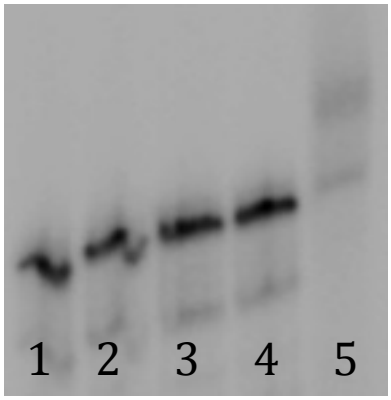
We evaluated the specificity of JB-181 stringently by comparing binding towards two very closely related RNAs: BIV TAR and 7SK RNA. The BIV TAR RNA is structurally and functionally homologous to HIV-1 TAR and serves as the most stringent control for selectivity: these peptides were originally designed after all to mimic BIV Tat. Satisfactorily, binding of JB-181 to BIV TAR was observed to occur with K_D of 4 nM. Thus, the improvement in binding was observed only for HIV-1 TAR, and not for even the very closely related BIV TAR RNA (Table 4).

Table 4: Binding affinity of JB-181 with competitor RNAs, BIV TAR and 7SK RNA as determined by EMSA.

RNA	K_D, nM
HIV-1 TAR	0.028 ± 0.004
HIV-1 TAR + 25K excess tRNA	0.250 ± 0.050
BIV TAR	4.00 ±1.00
7SK	2500 ± 1300

The 7SK RNA is also structurally similar to HIV-1 TAR. As a transcriptional regulator, when bound to PTEF-b and HEXIM proteins, 7SK RNA inactivates the PTEF-b kinase. Tat binds to 7SK, displacing the HEXIM proteins, thereby making the PTEF-b active for positive transcription of HIV [52, 53]. Furthermore, 7SK RNA forms a binding site for Arginine, like TAR [53]. Despite these structural and functional similarities, JB-

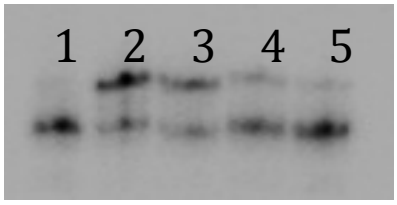
181 binds to 7SK with a very modest 3 μM binding constant (Figure 14).



Lane 1: Free ^{32}P-7SK RNA
Lane 2: 333 pM JB-181
Lane 3: 8.3 nM JB-181
Lane 4: 16.6 nM JB-181
Lane 5: 3.3 μM JB-181

Figure 14: EMSA of *in vitro* binding activity of 181 with ^{32}P radiolabeled 7SK RNA. From left to right, the concentration of JB-181 increases. Smearing in lane 5 indicates weak micromolar binding.

We further evaluated the selectivity of JB-181 for HIV-1 TAR RNA by introducing competitor 7SK and BIV TAR RNAs into EMSA experiments with HIV TAR (Figure 15).

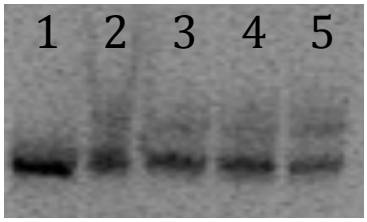


Lane 1: Free ^{32}P-HIV TAR RNA
Lane 2: Control 181 + HIV TAR RNA
Lane 3: HIV TAR RNA + 75 pM 7SK RNA
Lane 4: HIV TAR RNA + 3.75 nM 7SK RNA
Lane 5: HIV TAR RNA + 20.5 nM 7SK RNA

Figure 15: Competition assay results with 7SK RNA: ^{32}P radiolabeled TAR concentration held constant at 37.5 pM with increasing amounts of unlabeled 7SK RNA titrated in.

While holding the concentration of ^{32}P radiolabeled HIV-1 TAR RNA constant at 37.5 pM, we introduced increasing amounts of unlabeled 7SK RNA. Only when the concentration of 7SK RNA exceeded that of HIV-1 TAR by >100 fold did we observe any diminished binding. Conversely, we introduced unlabeled HIV-1 TAR RNA to ^{32}P

radiolabeled BIV TAR RNA with JB-181 (Figure 16); as the concentration of HIV TAR RNA in solution increased, binding between JB-181 and BIV TAR rapidly diminished due to the peptide binding in a better capacity to unlabeled HIV TAR.



Lane 1: Free ^{32}P-BIV TAR RNA
Lane 2: BIV TAR RNA + 3 nM HIV TAR RNA
Lane 3: BIV TAR RNA + 2 nM HIV TAR RNA
Lane 4: BIV TAR RNA + 1 nM HIV TAR RNA
Lane 5: Control 181 + BIV TAR RNA

Figure 16: BIV TAR RNA is labeled and binds at approximately 3 nM (lane 5). Going from right to left, as HIV TAR RNA is introduced, binding diminishes between radiolabeled BIV and JB-181.

***In vitro* peptide activity against variants of HIV-1 TAR found in various HIV subtypes -** Development of antiretroviral drugs primarily target subtype B HIV-1 strains, such as the NL4-3 lab strain used in our studies. However, the B subtype comprises only 10% of current HIV infections. Subtypes A, C, D and CRF01_AE account for 80% of all viruses [54]. HIV-1 TAR RNA is very highly (91.1%) conserved in the alignment of 3122 HIV-1 TAR RNA sequences from the Los Alamos database, but some mutations are found within this RNA. We evaluated these peptides' span of activity against a variety of TAR variants taken from primary HIV-1 isolates, which contained a) substitutions of conserved nucleotides that are critical for Tat binding and transactivation [55], and b) substitutions observed over a period of time in HIV-1 patients exposed to treatment [56].

Substitutions of conserved nucleotides critical for Tat binding

Base pairs directly above the bulge (GC26/39), and the U23 nucleotide located in the bulge of HIV-1 TAR RNA are critical points of contact for Tat binding and *in vivo* transactivation activity [55]. The GC26/39 is nearly 100% pairwise conserved and disruption of this base pair causes a severe reduction in *in vivo* HIV-1 activity. Modifying this base pair to C26/G39, which is never observed in any isolate, reduces the binding capacity only slightly to 74 pM, for JB-181. The AU22/40 base pair is strongly conserved across HIV-1 isolates but it is rarely mutated to UA. JB-181 retained low picomolar binding (100 pM) for this mutation.

U23 is necessary for binding of Tat and we found it to be critical for JB-181 as well. We observed no binding when we substituted U23C or simply removed U23, but it is rarely mutated anyway (less than 0.2%). It is not clear how HIV replicates with this mutation. The HIV-1 TAR sequence of a patient followed longitudinally for four years (see below) contained the U23C substitution, but it reverted back to U23. Mutations of U25 are more common (70.1% pairwise percent identity) and this base was not reported to play a significant role in Tat binding. However, binding of JB-181 to the U25C mutant was reduced to approximately 1 nM. In some (but not all) calculated structures, this residue is tucked under Lys6, suggesting that its mutation may disrupt interactions formed by this amino acid side chain. Finally, C24 is maintained with 63.2% pairwise percentage identity and does not participate in binding to Tat [51]. We changed this base to 2-amino purine for our assays, and binding data obtained with EMSA are consistent with low picomolar binding, indicating that a pyrimidine to purine substitution was non-perturbing.

Substitutions observed over a period of 3 years in HIV-1 patients

A second interesting set of mutants occurring throughout the stem and loop of HIV-1 TAR were observed over a 3 years study of four early stage asymptomatic HIV-1 patients [56]. Many of the substitutions did not exhibit co-variation in base pairs and would be predicted to disrupt or destabilize the secondary structure of HIV-1 TAR. We evaluated the binding affinity of the JB-181 through the progression of the mutations in two patients.

TAR contained variations in the loop sequence, G32A and G36A, after 122 days in the first patient. The binding capacity of JB-181 was retained to better than 166pM suggesting that any purine in these positions support peptide binding. After 850 days, the TAR sequence was mutated again to include the double substitution A35G/G28C; binding to JB-181 was reduced to 10 nM by this substitution. It is likely that disruption of a critical contact between Arg5 of the peptide and nucleotide G28 would cause a decrease in binding. This sequence might nonetheless not be functional for the virus, since TAR reverted to the native sequence after a few days.

The HIV-1 TAR sequence of a second patient contained G33A, G34A and G21A mutations after 425 days of treatment. While G21 has not been shown to affect HIV-1 replication [55], modification of this base without co-varying its base-paired C would weaken the RNA secondary structure. Binding of JB-181 was essentially lost (>70 nM) with this mutant but the TAR sequence in the patient reverted to the consensus sequence.

These data further support previously presented data of the ability of the peptide to inhibit replication across divergent subtypes [29] and suggest that JB-181 would

remain active against the majority of these mutations.

Structure of the JB-181 HIV-1 TAR complex - In order to understand the molecular basis for the much more potent activity of the new peptide, we determined the 3D NMR structure of the JB-181-TAR complex. Using the experimental constraints listed in Table A1 (Appendix A), the 1.6 Å NMR structure of the 12+2 residue peptide, JB-181, bound to the 29 residue transactivation binding domain of HIV-1 TAR RNA confirmed that JB-181 binds to HIV-1 TAR in the upper portion of the major groove, like the other peptides of this class [26, 40]. The lysine-glycine turn is positioned upward, toward the loop-helix junction (residues 26-29, 36-39), while the D-pro-L-pro dipeptide scaffold is positioned downward toward the lower portion of the stem (residues 17-22, 40-45) (Figure 17).

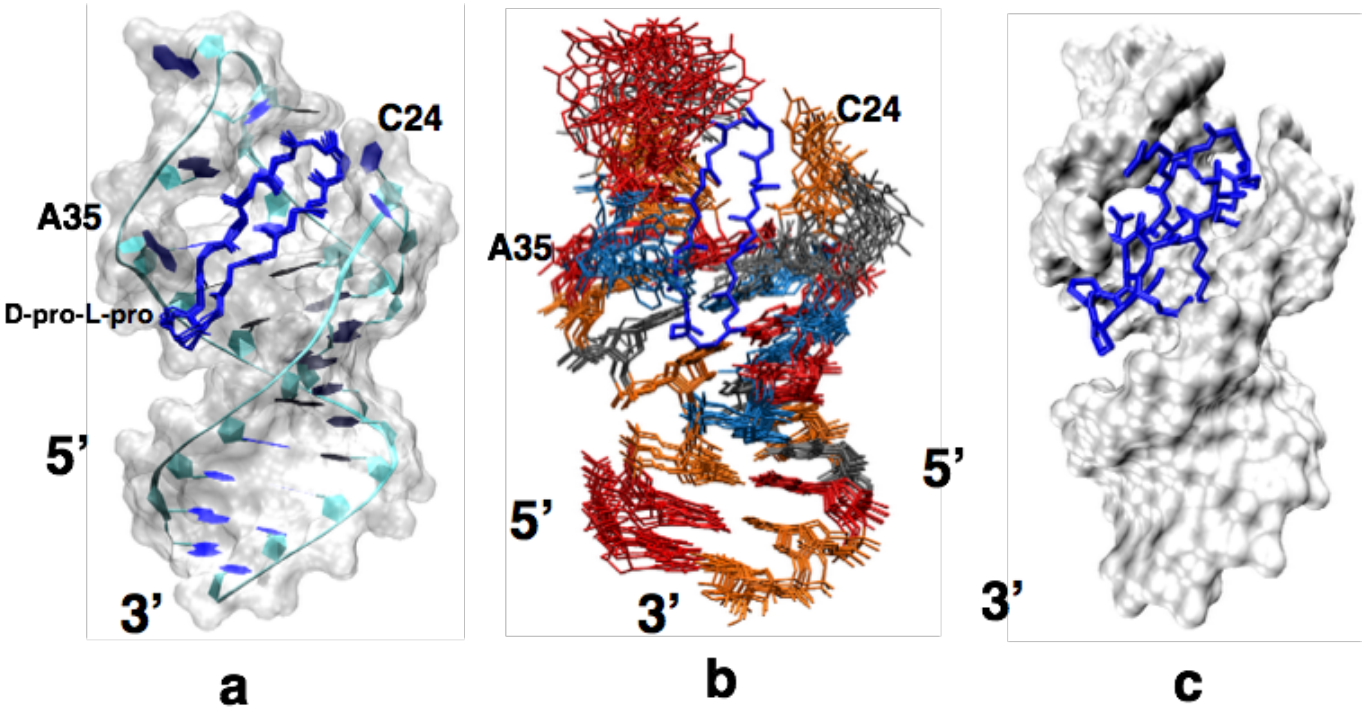


Figure 17: Structure of JB 181 bound to HIV-1 TAR RNA, as determined from the restraints summarized in Table A1. A) The peptide backbones from the ten lowest energy structures are superimposed. B) The RNA from the ten lowest energy structures are superimposed. C) Lowest energy structure of the complex.

Even before formal determination of the structure of the complex, a superficial examination of NOESY spectra of each complex showed that JB-181 made more intimate contacts with TAR RNA compared to L-22 (Figure 18).

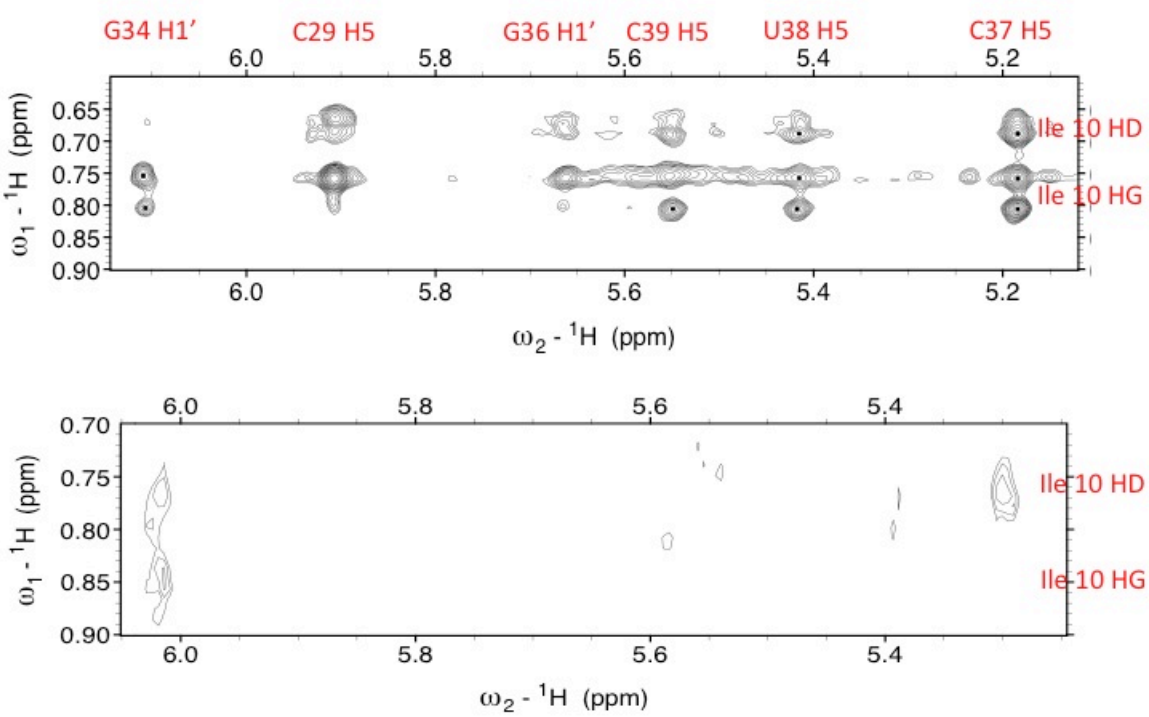


Figure 18: Intermolecular NOEs between isoleucine residues 10 and 12 and RNA protons located in the major groove (residues 36-39) are much more numerous in the JB-181 complex (top) than in the L-22 complex (bottom). The only conserved intermolecular NOEs are between gamma and delta protons of isoleucine 10 with G34 H1'. This supports the hypotheses that shortening the side chain of residues 1 and 11 allows the peptide to sit deeper in the major groove.

The most notable differences supporting this conclusion was the improvement of some previously broad resonances and the presence of many more intermolecular NOEs between two critical hydrophobic isoleucine residues (10 and 12) and the RNA bases in the major groove (residues 36-39), which were not present or very weak in the L-22 complex, indicating that the JB-181 peptide sits much deeper in the major groove. Using VMD [57], we calculated the average distance from Ile10 γ protons to

the amino group of C37 in both complexes; the average distance between these atoms, in the lowest energy L-22 TAR complexes, was 5.2 Å, while the distance is much shorter, 2.7 Å, in the JB-181-TAR complex, supporting our hypotheses that shortening the side chains would bring the peptide and RNA closer together.

Multiple intermolecular NOEs between the γ and δ hydrogens of Ile10 and Ile12 support the positioning of the peptide deep into the groove and packing against the U-A-U base triple. The base triple interaction between U23 and A27/U38 is supported by strong proton-proton resonances between the A27 amine protons with both imino groups (H3) of U38 and U23, which were not visible with L-22 (Figure 19).

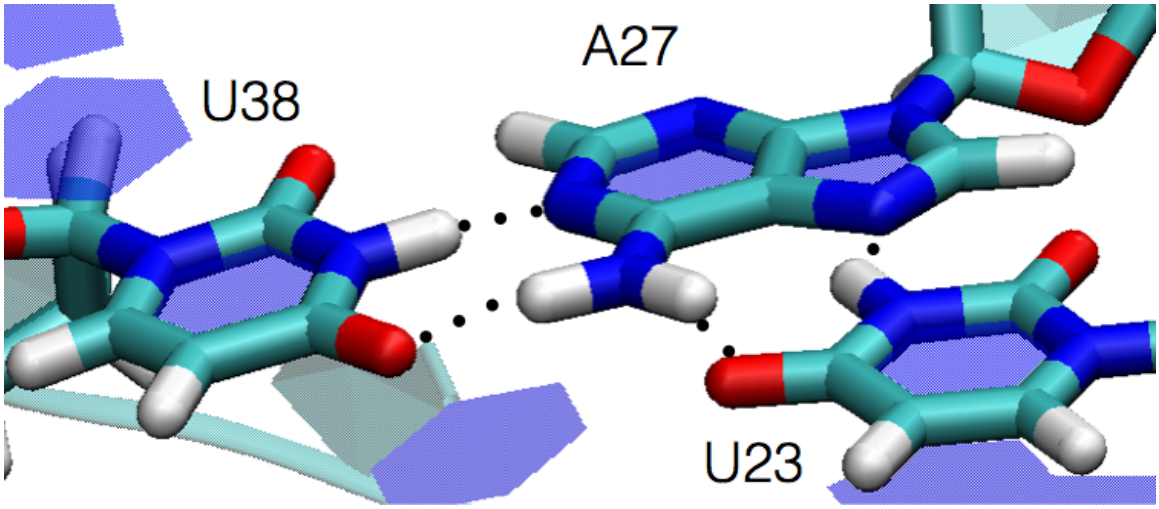


Figure 19: Base triple between U38, A27, and U23 is maintained in the JB-181 TAR RNA NMR structure.

The intimate contacts between the Ile 10 side chain and the RNA drives the formation of the base triple, as observed for the BIV TAR-Tat structure and designed into our peptide mimics of the BIV Tat peptide; HIV-1 Tat has no equivalent Ile residue within its RNA binding domain (Figure 20).

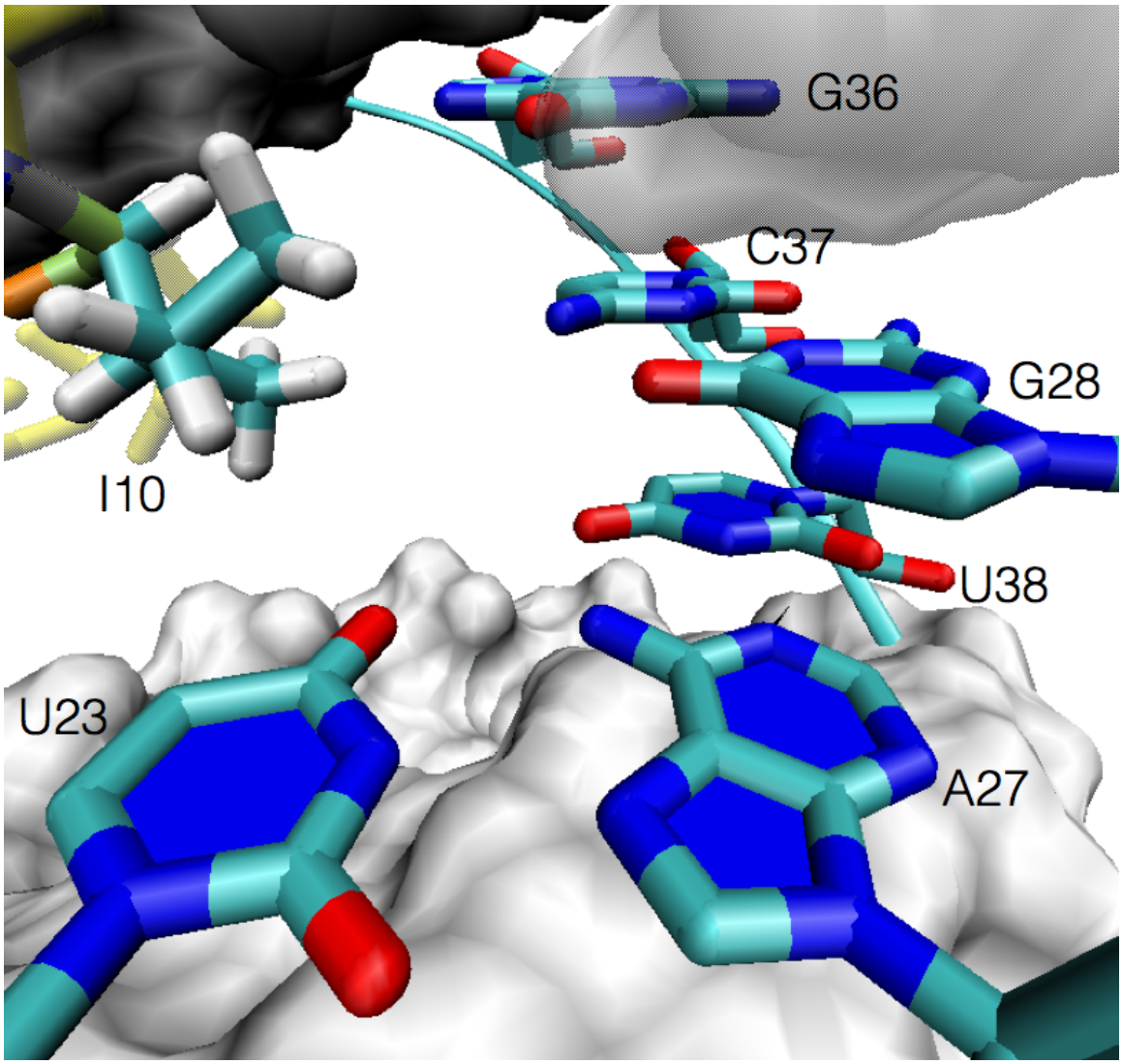


Figure 20: Isoleucine 10 stabilizes the base triple. Intermolecular NOEs are present between its gamma and delta protons and protons of residues G36, C37, G28, U38, A27, and C29.

The apical loop is flexible in the free RNA and retains some dynamics even in the complex with JB-181. However, hydrogen bonding interactions formed by the guanidinium groups of Arg8 and Arg9 with the bases of the single stranded C30, G32 and G33 nucleotides stabilize the structure of the loop compared to the free RNA and to the L-22 complex. These interactions 'pull' these bases down toward the U-C-U

bulge, covering the peptide and burying it into the major groove (Figure 21).

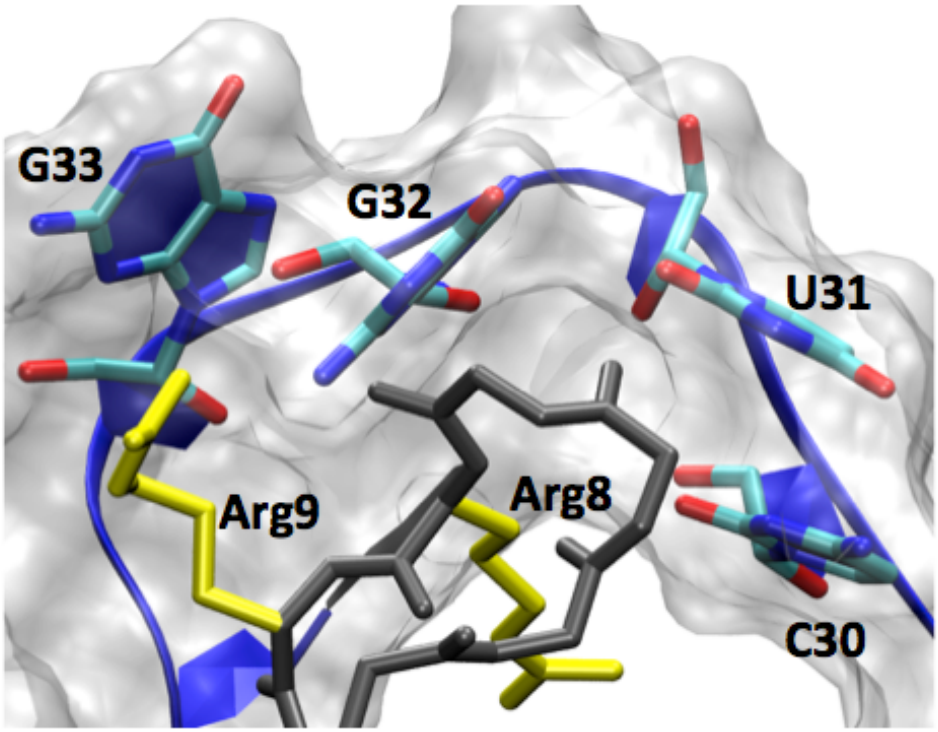


Figure 21: Bases G33, G32, and U31 are positioned above the peptide, acting to hold and push the peptide deeper into the major groove of HIV-1 TAR RNA.

The bases of G34 and A35 are also more precisely positioned than before by newly observed strong intermolecular NOEs in the complex. A35 is flipped in, away from the other loop residues and is drawn down toward the lower portion of the helix toward the U-C-U bulge, where it occupies a pocket formed by peptide residues Pro14, D-Pro13, Nor11 and Val2 (Figure 22).

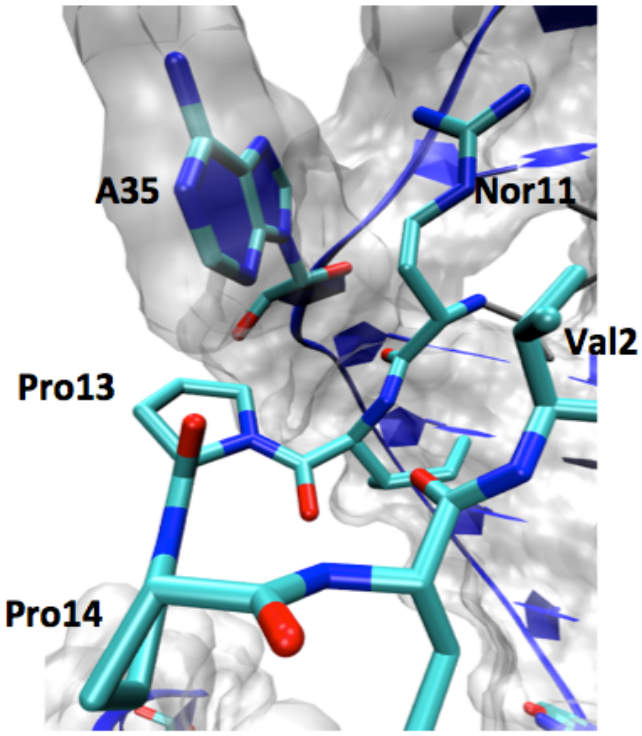


Figure 22: A35 is positioned away from the rest of the loop and is pulled down over the peptide; it sits in a pocket created by Nor11, Pro13, Pro14 and Val2.

G34 is drawn instead behind the peptide, its heterocyclic rings somewhat perpendicular to Arg9 and its H8 proton pointing down toward Ile10. The positioning of G34 provides increased space for the peptide to slide deeper into the groove. In the L-22 structure, the base was positioned in such a way that the heterocyclic rings of G34 blocked Arg9, and effectively prevented the rest of the peptide from sliding deeper into the major groove (Figure 23, 24).

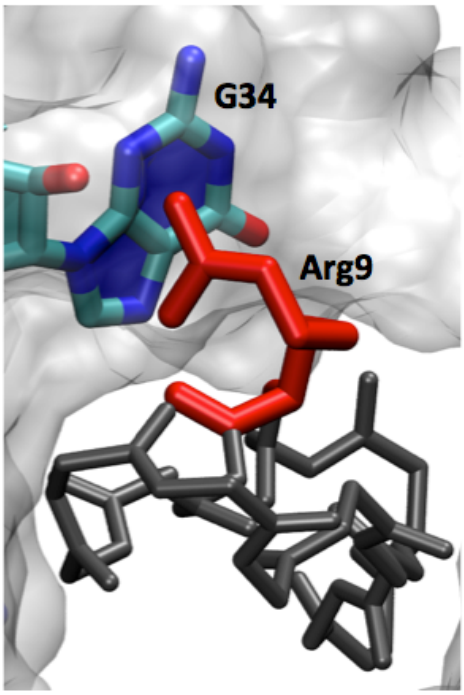


Figure 23: In the L-22 structure, G34 blocks Arg9, and effectively the rest of the peptide, from making more intimate contacts with the RNA.

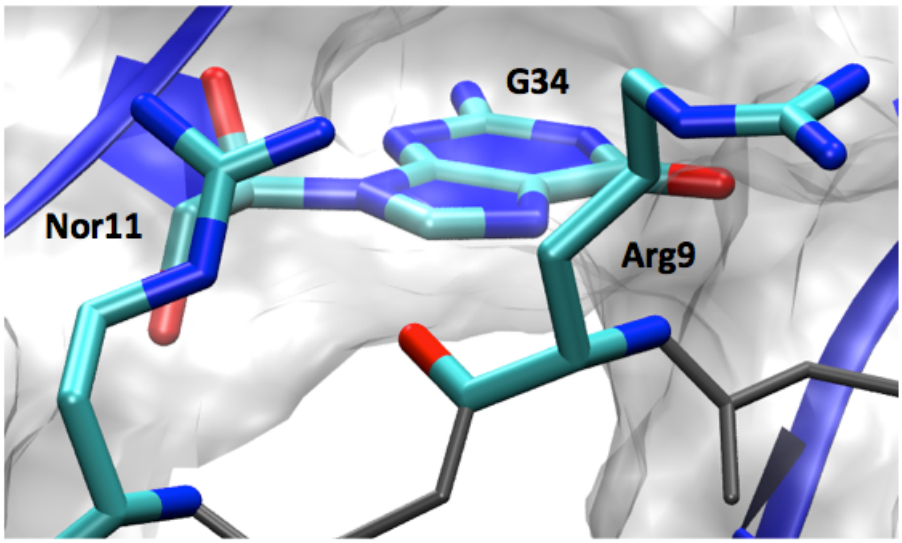


Figure 24: In the JB-181 complex, G34 is positioned perpendicular to Arg9 and Nor11, with its H8 pointed toward the peptide. This allows the peptide to slide closer to the groove while the other loop residues are pulled down over the peptide.

Antiviral activity

L-22-derived peptides inhibit HIV-1 replication in CD4+ T cells.

We tested whether differences in binding affinity would affect the ability of the peptides to block HIV-1 replication in CD4+ T cells (Figure 25).

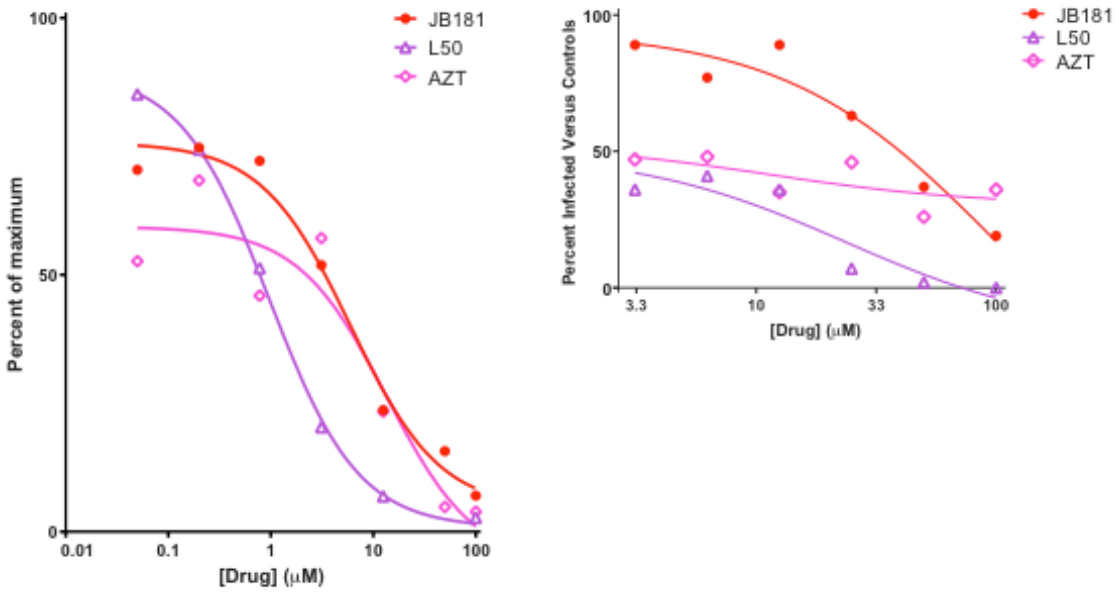


Figure 25: Peptide derivatives of L-22 were tested for their ability to inhibit HIV-1 spread in CD4+ T cells derived from peripheral blood mononuclear cells (PBMCs), using AZT as control. Whole CD4+ T cells purified from PBMCs were infected with an HIV-1 clone expressing Nef-IRES-GFP at an MOI of 0.1. Two days post-infection, a subset of cells were measured for GFP by flow cytometry while the remaining cells were incubated with increasing concentrations of the peptides and AZT. 5 days post-infection, cells were again measured for GFP by flow cytometry. Percent inhibition was plotted by taking the percent cells infected on day 2 as 100% and the percent cells infected on day 5 in the absence of drug as 0%.

CD4+ cells were collected from human PBMCs and infected with GFP+ HIV-1 viruses. Peptides were added to the infected cells two days after infection rather than prior to infection in order to measure virus spread (which would require both reverse transcription and Tat-mediated transcription) rather than just establishment of infection (which would only require reverse transcription). Peptides that bound to TAR with poor affinity (JB-182 and JB-184) and those that had low nM binding affinities had different antiviral activity. While the majority of the peptides blocked spread at IC50 values between 19.1 μM and 39.8 μM, JB-182 and JB-184 blocked less than 50% of virus spread (Table 5, Figure 26). These results suggest that the peptides that bind poorly to TAR are not as effective at inhibiting viral replication.

Table 5: Inhibition of HIV-1 spreading infection as determined by GFP fluorescence in infected cells.

Inhibitor	IC_{50}, mM
JB-181	39.8 ± 0.5
JB-182	>100
JB-184	>100
JB-185	33.9 ± 0.2
JB-186	38.9 ± 0.2
JB-58	19.1 ± 0.4
JB-59	22.9 ± 0.36
L-50	<1
AZT	<1

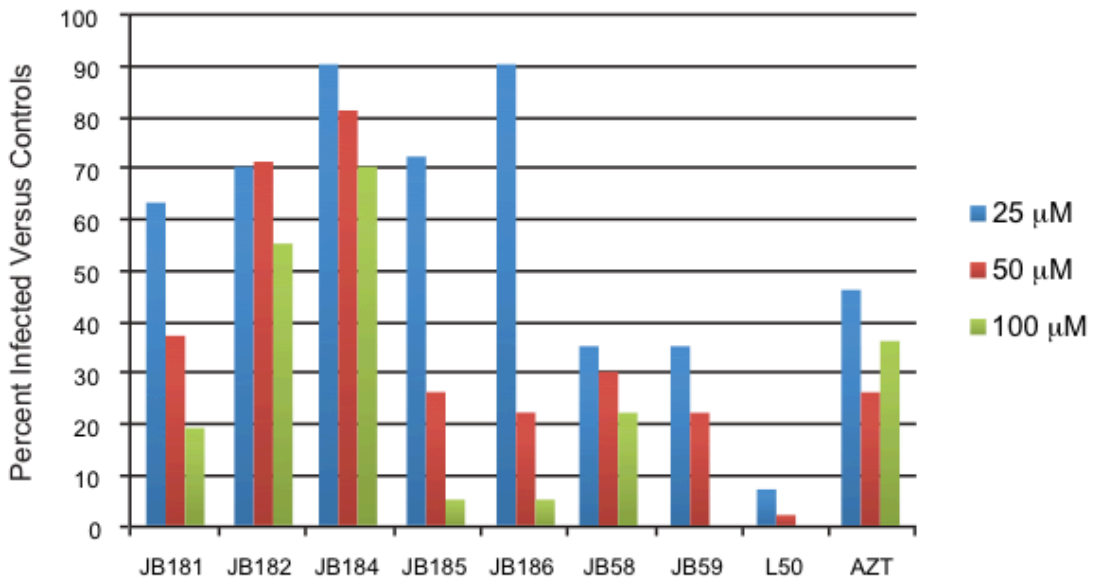


Figure 26: Peptide derivatives of L-22 with AZT as a control were tested for their ability to inhibit HIV-1 (−) strand strong stop DNA synthesis in an in vitro transcription assay. Increasing concentrations of each peptide and AZT were incubated with a subgenomic HIV-1 RNA containing the 5'-untranslated region (from TAR through the primer binding site) and a ^{32}P-radiolabeled DNA primer. Following peptide binding, HIV-1 RT was added to each reaction for 30 min at 37°C. Reactions were separated on denaturing gels, and bands representing full length (−) strand strong stop DNA products were quantified following autoradiography via phosphorimager. Plots represent average of experiments in triplicate.

L-22-derived peptides inhibit reverse transcription irrespective of binding affinity.

L-50 was previously demonstrated to block reverse transcription and Tat-mediated transcription, and thus inhibited viral replication through a TAR-dependent mechanism [29]. We therefore tested whether higher affinity binding to TAR led to more potent activity against reverse transcription. An *in vitro* reverse transcription assay revealed that there is no correlation between binding affinity, when this is strong enough, and inhibition of reverse transcription, e.g., the peptide with the highest binding affinity, JB-181, had an IC_{50} of 6.0 ± 0.3 μM, while the two peptides with the lowest binding affinities, JB-182 and JB-184 had lower IC_{50} values at 1.2 ± 0.2 μM and 1.16 ± 0.02 μM, respectively (Table 6).

Table 6: *In vitro* inhibition of HIV-1 reverse transcriptase as determined by incorporation of ^{32}P-ATP into reverse transcription product from HIV-1 subgenomic RNA template.

Inhibitor	IC_{50}, mM
JB-181	6.0 ± 0.3
JB-182	1.2 ± 0.2
JB-184	1.16 ± 0.02
JB-185	1.6 ± 0.2
JB-186	1.4 ± 0.2
JB-58	1.6 ± 0.2
JB-59	0.61 ± 0.36
L-50	0.97 ± 0.31
AZT	13.5 ± 0.6

There is also no evidence for an inverse correlation, as peptides like JB-185 with increased binding affinity compared to L-22 had IC_{50} values (1.6 ± 0.2 μM) comparable to JB-182 and JB-184. Thus, increases in binding affinity beyond nM did not affect inhibition of reverse transcription in a reconstituted in vitro assay.

Section 3: Discussion

The TAR-Tat interaction in HIV has long been described as a promising target for the development of new antiviral drugs [10, 55, 58], but it also represents a paradigm for the discovery of drug-like molecules that bind to RNA [2, 59]. In this regard, there is still considerable skepticism as to whether drug-like molecules can be found that bind to RNA with the kind of affinity and specificity required for pharmaceutical applications, especially to simple secondary structures such as TAR instead of more complex tertiary sites provided by the ribosome [1, 2]. Furthermore, many questions remain as to whether molecules synthesized in the past and reported to interfere with viral replication did work *in vivo* by a TAR RNA-dependent mechanism [22, 60].

We have previously reported the design, synthesis and characterization of a class of cyclic peptide mimetics of Tat that bound to TAR with nM affinity and acted *in vivo* by inhibiting both Tat-dependent transcriptional activation and initiation of reverse transcription in a TAR-dependent manner [26, 29]. However, the most potent molecules of this class, called L-50 and L-22, did not discriminate well between different RNAs. This biophysical characteristic, together with cell penetration, limited their potency in cell-based assays [29]. By introducing non-natural amino acid side chains, we have now improved binding to reach the low pM range without improving affinity for off target RNAs even as closely related as BIV TAR. The increase in affinity and specificity led to improved antiviral potency in primary lymphocytes, which is now as strong as that of the FDA-approved drug AZT. We have not investigated the mechanism of action of JB-181, but its considerable similarity with L-50 and L-22 strongly suggest that it is the same as for the previous peptides.

Before introducing non-canonical side chains, we synthesized and assayed about 100

peptides to improve upon the scaffold of L-50 but did not observe any significant improvement in binding. We thus resorted to a more careful examination of the 3D NMR structure of L-22 complexed with HIV-1 TAR [26]. We observed that two arginine residues, in order to maintain good side chain stereochemistry, appeared to "push" the peptide away from the RNA. By shortening the arginines by one or more methylene groups, thus introducing nonstandard amino acids into the peptide, we were able to maintain the essential interaction contacts provided by the guanidinium groups and improve activity with the library of peptides presented in this article.

Unexpectedly, we also observed a further closing of the RNA loop above the peptide. In fact, the most remarkable result of this peptide modification is how the RNA structure molds itself around the peptide to an even greater extent than we noticed with L-22. The apical loop reorganized itself to not only have A35 flipped down and over the peptide, but conserved RNA bases in the loop fold over the peptide and bury it deeper into the major groove, thus contributing to improved affinity and specificity. These additional points of contacts arise in a critical region of the structure, the binding site for cyclin T1, and involve conserved positions that we were unable to target in the past by extending the size of the peptide by four amino acids [40].

Constrained peptides of this class are not only potent ligands, but are also able to penetrate the cell membrane, exhibit low cytoxicity, and are proteolytically stable [29]. They have been developed as antibacterial compounds and are now in clinical evaluation [61]. The peptides have also been demonstrated to be effective against various major strains of HIV-1 TAR [29] and indeed we have shown that the JB-181 peptide generally retains strong binding activity against mutants found in even rare isolates. Thus, it is likely that these molecules could be developed into bona-fide

clinical candidates, although much work remains to be done to optimize their pharmacological potential. More broadly, peptides of this class might represent a useful addition to the class of structures that can be used to target structured RNAs well beyond TAR.

Chapter 4: *In silico* Free Energy Perturbation: Modeling Cyclic Inhibitors of HIV-1 Replication

Section 1: Introduction

Since the 1980s, implementation of the thermodynamic perturbations method has been used to predict free energy of binding for biomolecular complexes [62-65]. One of the earliest applications of this method was with the enzyme thermolysin complexed with a pair of phosphoamidate and phosphonate inhibitors. The calculated free energy difference by thermodynamic perturbation was reported at 4.21 + 0.54 kcal/mol. The experimental value for this same complex, 4.1 kcal/mol, agreed very well with the calculated free energy difference [66].

Over the past 30 years, the methodological advancements in force field development have made it possible to computationally study structural conformation, small molecule and protein binding to both protein and RNA, and even the effects of solutions (i.e. pH and salt concentrations) on reactions [67-69]. I utilized this method to study our cyclic peptide inhibitors (Chapter 3) complexed with HIV-1 TAR RNA.

Although structure-based design has led to the development of potent picomolar inhibitors of HIV-1 replication via the HIV-1 TAR RNA and Tat interaction [26], optimizing the selectivity and specificity of these peptide inhibitors is a trial-and-error process. By way of free energy perturbations, implemented with molecular dynamics, we hypothesized that we could determine the free energy of binding for peptide inhibitors of HIV-1 replication. Establishing a trend of energies for pre-

existing complexes would guide us in the future selection of which peptides in our library were more rational choices for synthesizing.

In this chapter, I will discuss our approach utilizing free energy minimization to computationally model the JB-181 TAR RNA complex. I will also discuss further developing this molecular modeling method, via free energy perturbation calculations to predict relative free energy of binding for peptide-HIV TAR complexes.

Section 2: Background and Theory

NAMD (Nanoscale Molecular Dynamics program [70]), complemented with VMD (Visual Molecular Dynamics [57]) provides a free platform for free energy perturbation (FEP) calculations. The *in silico* methodology supported by NAMD and VMD is the dual topology approach [67, 71, 72], a statistical mechanical method used to calculate the free energy difference between two similar states: a, the reference, and b, the target. The FEP approach involves the concurrent, but independent transformation of the reference to the target, which is known as the forward reaction, and the target to the reference, known as the backward reaction. The backward reaction is run for validation.

The free energy can be calculated from the partition function, Q, which is a sum of all the possible Boltzmann weights of all the energy levels of the reference and target systems. The partition function can be expressed as:

$$Q=1/N!h^{3N} \iint e^{-(H(p^N,q^N))/RT}\, dp^N\, dq^N$$

where N represents the number of particles, h is Planck's constant, R is the universal gas constant, and T is temperature. p^N describes the coordinates of interacting atoms, while q^N represents the momenta of all the particles in the system. The difference in free energy, ΔG, between states a and b is expressed as:

$$\Delta G=G_b-G_a= -1/\beta \ln Q_b/Q_a$$

where $\beta = 1/k_bT$ and k_b is the Boltzmann constant.

The free energy difference between the direct, one step, transition from state a to b, and vice versus, is impractical and yields an inaccurate free energy difference.

Instead, a series of intermediate steps between the two states is required for a more accurate calculation. Thus, the parameter λ is introduced to allow for the reversible transformation between the reference and target states. The partition function can now be related to the Boltzmann distribution as a function of λ:

$P(p^N q^N \lambda) = \text{"exp"}[-\beta H(p^N q^N \lambda)]/(\int dp^N dq^N \text{"exp"}[-\beta H(p^N q^N \lambda)])$

$\equiv 1/Q_\lambda \ \text{"exp"}[-\beta H(p^N q^N \lambda)]$

where $\mathcal{H}\, p^N q^N \lambda$ is the Hamiltonian characterizing the intermediate states defined by λ. The FEP free energy difference for the forward and reverse transformations, respectively, can be expressed as:

$\text{"exp"}(-\beta\Delta G) = \langle \exp(-\beta\Delta U) \rangle_a$ and $\text{"exp"}(+\beta\Delta G) = \langle \exp(+\beta\Delta U) \rangle_b$

where ΔU is the difference in potential energies $U(p^N\lambda = 1) - U(p^N\lambda = 0)$. $\langle \ldots \rangle_a$, denotes an ensemble average of configurations represented by Hamiltonian $\mathcal{H}_a$.

In the concurrent, but independent, transformation between **a** and **b**, the atoms in the molecules are classified by three groups:

a) the atoms that are not involved in the transition from **a** to **b**

b) the atoms representing state **a**, and

c) the atoms representing state **b**.

Throughout the free energy perturbation calculations, changes are only made locally; intermolecular potentials are not altered. Because system **a** and **b** must be similar in order for a sensible free energy to be calculated, this is an ideal model for mutating amino acids in a peptide scaffold to determine the free energy associated with the peptide-RNA complex, relative to the parent L-22 complex.

Free energy can be expressed in terms of the dissociation constant:

$\Delta G = - RT \ln K_D$

We hypothesized that if we were able to determine low energy inducing peptides, relative to L-22, we would effectively be able to achieve tighter binding peptides with HIV-1 TAR RNA. We further hypothesized that tighter binding would lead to better *in vivo* inhibition activity, as described in Chapter 3. Additionally, because we were able to improve binding with the use of nonstandard amino acids, we wanted to include those in our calculation predictions as well. We first began with the transformation of L-22 to JB-181 complexed with HIV-1 TAR RNA.

Section 3: Methods

Setup

We downloaded and installed both NAMD 2.10 and VMD 1.9.2 from http://ks.uiuc.edu/Research. NAMD is a molecular dynamics simulation package, which utilizes the CHARMM force field. VMD is a tool used for visualizing and analyzing the results of molecular dynamics simulations. The '*In silico* alchemy: A tutorial for alchemical free-energy perturbation calculations with NAMD' was used as a guide for conducting the FEP simulations [71].

Figure 27 is a general flowchart indicating the steps taken to perform the FEP simulations. In the next section, I will describe each of these steps in detail.

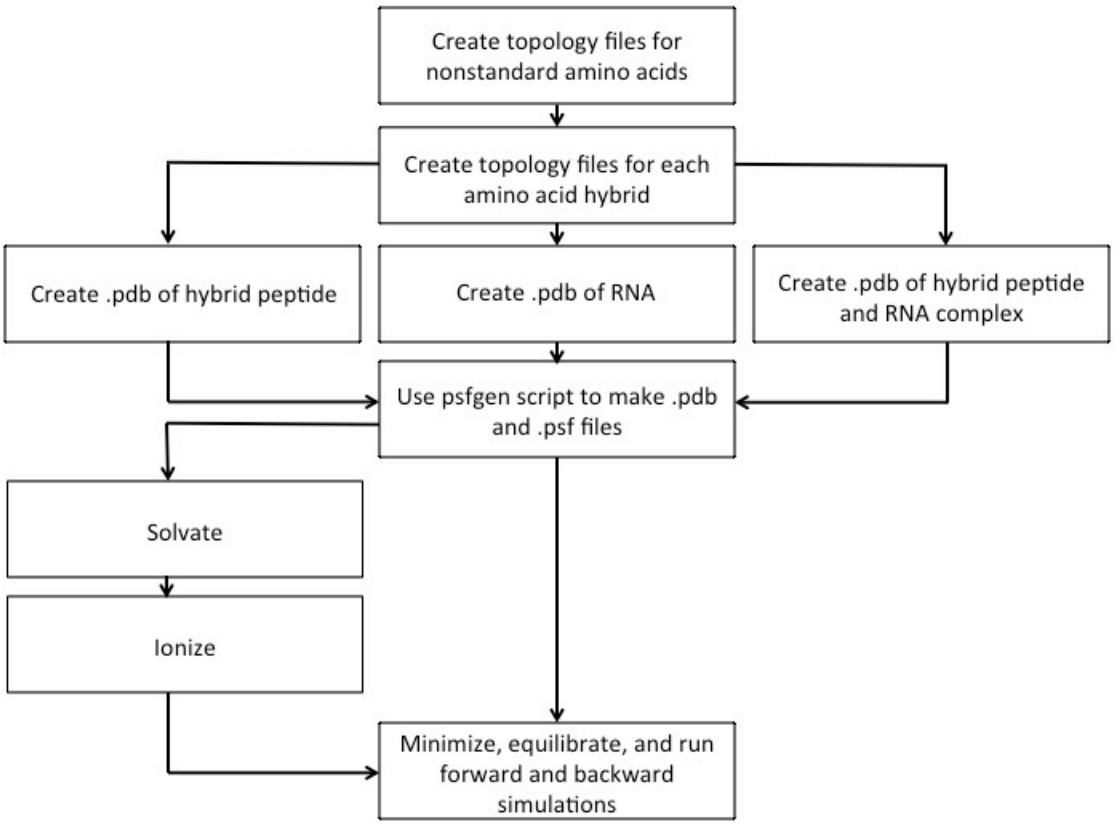


Figure 27: Flow chart of the steps taken to run the FEP simulations.

System Setup: generation of topology files

Before utilizing the dual topology approach for energy minimization and FEP simulations, topology files had to be manually generated for the nonstandard amino acids in JB-181 (diamino butyric acid (DAB) and norarginine (NOR)), D-proline (DPR), and the other nonstandard amino acids that we were interested in substituting into our peptide scaffold. These nonstandard amino acids include DL 2-3 -diaminopropionic acid (DAP), L-ornithine (ORN), L-2-amino-3-aureidopropionic acid, D-5-hydroxylysine, Lys(N3)-OH, and L-homoarginine. These files were added to the general topology file utilized by the NAMD software. The nonstandard amino acid topology files for DAB, NOR, and DPR are listed in Appendix A. Hybrid topology files

representing the transition of one amino acid to another, in this case R to diaminio butyric acid (R2B), and R to norarginine (R2N), were created. These are listed in Appendix A.

Model 8 of the L-22 TAR pdb (2KDQ) was randomly selected and modified to contain R2B in place of arginine 1, and R2N in place of arginine 11. The new .pdb file was split into two separate .pdb files: one containing solely RNA and the other containing just the hybrid peptide. This was done because both the RNA and the peptide entities have to be solvated an ionized before running calculations.

From here, the NAMD psfgen script was modified with pdb aliases (atoms from the hybrid residues had to be redefined) and the new topology files were added. This script was then used to make both a .psf and a .pdb file for each the RNA and the peptide.

An alchFile, a .fep file, that indicates which class each atom belongs (recall there are those involved in the transformation, appearing or vanishing, and those that remain unchanged) was created from the previous .pdb file for the hybrid peptide. The RNA .pdb remained unchanged.

Using the autosolvate plug-in of VMD, both the RNA and the hybrid peptide were solvated to contain a water box that extended 5 Å within each direction around the complex.

Following solvation, VMD's autoionize feature was used to ionize both the RNA and the peptide in $MgCl_2$. The concentration was set at 5 mM, similar to the concentration of $MgCl_2$ present during the binding mobility shift assays. The minimum distance of

ions from the solute was set to 5 Å. The minimum distance between individual ions was set to 5 Å.

The peptide and the RNA were then minimized and equilibrated individually, and then combined together in one .pdb file to form the complex. The complex was minimized and equilibrated again. Minimization sets the number of iterations over which to vary atom positions to search for a local minimum in the potential. It eliminates bad initial contacts and reinitializes the velocities to the desired target temperature. Following minimization and equilibration, we calculated the free energy change involved in the two point mutation of L-22-HIV-1 TAR to JB-181-HIV-1 TAR, as illustrated in the thermodynamic cycle below (RNA not shown). The calculation was conducted both *in vacuo* and in a hydrated system (solvated) (Figure 28).

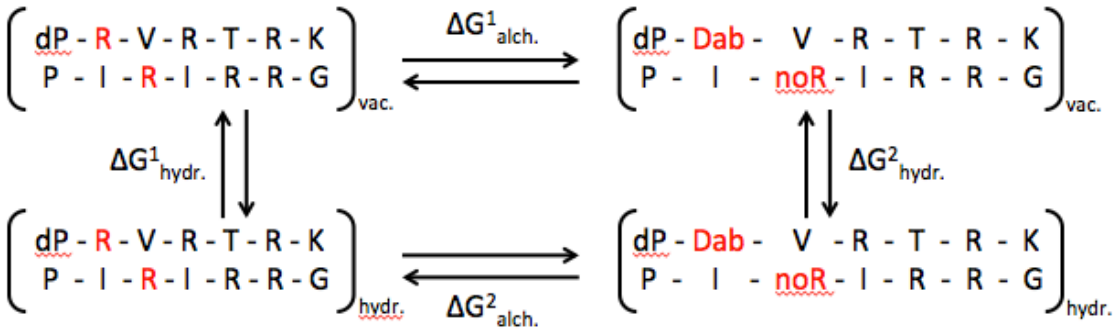


Figure 28: Thermodynamic cycle run in the alchemical free energy perturbation calculations.

The FEP simulation was carried out using the forward-shift.namd and the backward-shift.namd scripts. The final ΔG(λ) was calculated using the provided delta.awk script.

Section 4: Results

The free energy yielded from the FEP simulation was 7.95 kJ, similar to the experimental free energy of binding, 10.4 kJ (Figure 29).

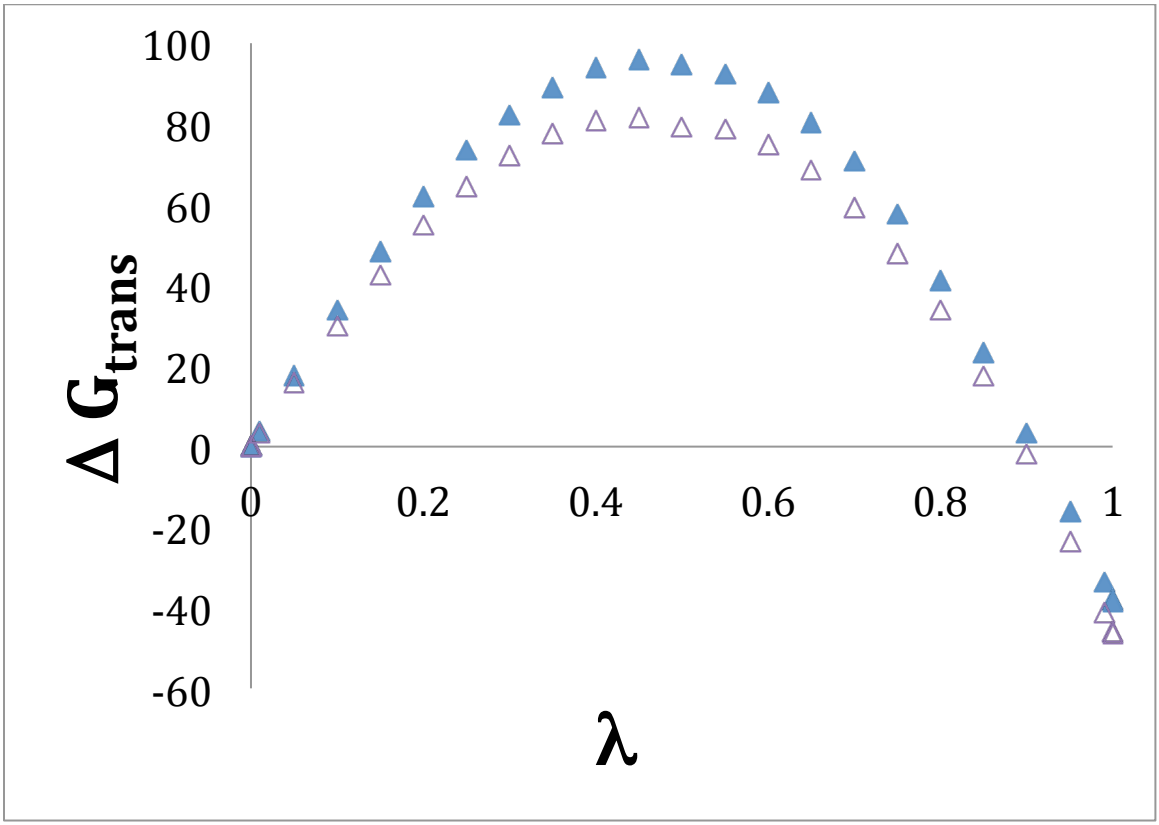


Figure 29: Forward and reverse run for JB-181 TAR, L-22 TAR transition. Blue represents the forward reaction, and purple represents the reverse reaction.

After calculating the free energy associated with the transformation of L-22 to JB-181, we obtained a predicted structure of JB-181 bound to HIV TAR. The predicted structure of JB-181 from L-22 is illustrated below (Figure 30), with highlights of the nonstandard amino acid changes.

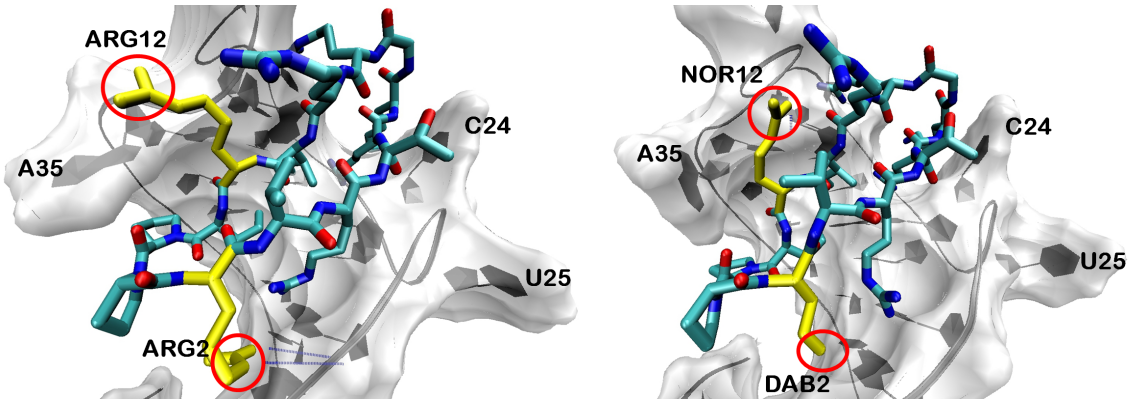


Figure 30: On the left is L22, with arginine 12 and arginine 2 highlighted in yellow. On the right is the FEP model of JB-181, after mutating arginine 12 to norarginine, and arginine 2 to diamino butyric acid.

Because FEP simulations require the energies between the final and initial states to be similar, the overall structure of the peptide oriented in the complex matches the orientation of other peptides of this class. The FEP structure of the 12+2 residue peptide, JB-181, bound to the 29 residue transactivation binding domain of HIV-1 TAR RNA revealed that JB-181 remained bound to HIV-1 TAR RNA in the upper portion of the major groove. The lysine-glycine turn is positioned upward, toward the loop-helix junction (residues 26-29, 36-39), while the D-pro-L-pro dipeptide linkage is positioned downward toward the lower portion of the stem (residues 17-22, 40-45).

Close examination of the modeled structure predicted that Dab2 would sit inside the phosphate backbone of G21 and A22, instead of resting against it, from the outside, as in the L-22 TAR structure. The FEP model also positioned Nor12 pocketed just below the phosphodiester backbone of A35; its shorter side chain allows for the peptide to sit deeper in the groove.

Structural comparison - Once I finished determining the NMR structure of JB-181 with HIV-1 TAR RNA, I compared it to the FEP model and observed great similarity between the complexes. The RMS value between the one of our lowest energy structures and the FEP generated structure was an impressive 2.75 A (Figure 31).

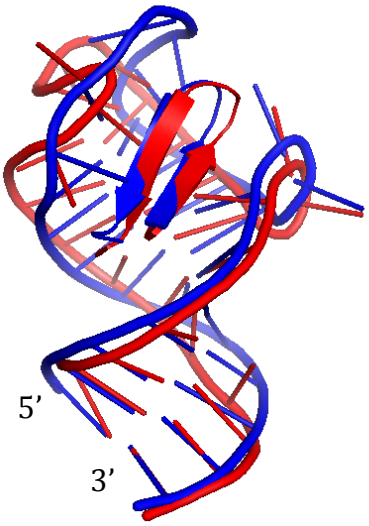


Figure 31: Alignment of FEP JB-181 HIV-1 TAR RNA complex (red) with NMR JB-181 HIV-1 TAR RNA complex (blue).

The peptide scaffold of L-22 and JB-181 differ in the modification of R1Dab and R11NorR. These residues are positioned similarly in both the FEP model and the low energy NMR resolved structures. The guanidinium group of Dab is pointed downward toward the phosphate backbones of G21 and A22 in both structures. The only noteable difference is that in the FEP structure, the guanidimium group is pointed toward the backbone, while in our top ten NMR structures, some reveal similar positioning while in others, the guanidinium group is pointed away. This could be primarily due to the lack of NMR constraints for that group in our NMR structures. The guanidinium group of norarginine is pointed upward in both the FEP and NMR resolved structures. It is positioned parallel to the A35 base.

The apical loop, for which we lacked experimental NOEs, was the most different between the two structures. Our NMR data suggests that nucleotides 30, 31, 32, and 33 are very flexible in solution. The difference in our two structures for this region was thus, not surprising (Figure 31). Interestingly, G34 remains tucked behind the peptide, near Ile10, similar to its positioning in the NMR structure of JB-181 complexed with TAR. The base triple is still present in the FEP complex structure, supporting previous data of it being formed by peptide interaction.

Section 6: Discussion

Utilizing traditional energy minimization simulations is a method that has been widely used as a starting point for the design of protein ligands for RNA. However this method is limited to very large changes in free energy. For example, the peptide JB-58 has an *in vitro* binding affinity with HIV-1 TAR RNA at 3 nM. An energy minimization simulation revealed a free energy of ~ -1,800 kcal/mol. Another peptide, with a sequence similar to JB-58 but containing the K1R and R11W mutation, JB-79, has an *in vitro* binding affinity of > 1 μM, and a corresponding free energy of ~ -20 kcal/mol. The relative difference in free energy expected for a peptide binding more than 25-fold weaker was observed. However in an attempt establish a similar trend with peptides that bind with similar *in vitro* binding affinities with JB-58 (between 1 and 5 nM), we were not able to discriminate between peptides (Table 1). Peptides that bound at 1, 3 and 5 nM had very similar free energies of ~ -2,000 kcal/mol. Thus, we hypothesized that the traditional energy minimization approach is more optimal for very large changes in binding activity and that the free energy minimization perturbation method, which accommodates very small changes in free energy, would be more appropriate for screening peptides containing nonstandard amino acids.

Table 7: The *in vitro* binding affinities and corresponding free energies of peptides evaluated by energy minimization simulations.

	Position													
Peptide	1	2	3	4	5	6	7	8	9	10	11	12	K_D (nM)	ΔG (kcal/mol)
JB-79	R	V	R	T	R	K	G	R	R	I	W	I	> 1000	-20.0
JB-58	K	V	R	T	R	K	G	R	R	I	R	I	3	-1774.8
JB-62	R	Y	R	T	R	K	G	R	R	I	R	I	3	-1991.35
JB-63	R	N	R	T	R	K	G	R	R	I	R	I	1	-2037.29
JB-65	R	T	R	T	R	K	G	R	R	I	R	I	1	-1991.84
JB-66	R	A	R	T	R	K	G	R	R	I	R	I	1	-1986.66
JB-75	R	V	R	T	R	K	G	R	R	A	R	I	5	-1989.86
JB-76	R	V	R	T	R	K	G	R	R	V	R	I	1	-1974.39
JB-84	R	V	R	T	R	K	G	R	R	I	R	V	1	-1993.07
JB-85	R	V	R	T	R	K	G	R	R	I	R	L	5	-1991.65

Screening peptides containing nonstandard amino acids in the JB-181 scaffold would require a method capable of discriminating between very small changes in energy states. The nonstandard amino acids to be introduced, after all, only differ by one to two methylene groups. Thus, we utilized the free energy minimization perturbation approach.

While the FEP predicted structure and free energy of the JB-181 HIV-1 TAR RNA complex agreed well with our experimentally determined results, there were two critical factors to be addressed before evaluating other peptides. First, the RMSD value for the complex should remain constant as and FEP simulation progresses toward the final state [73], however the RMSD for both the peptide and the RNA continuously increased, despite changes in window and step size (Figure 32).

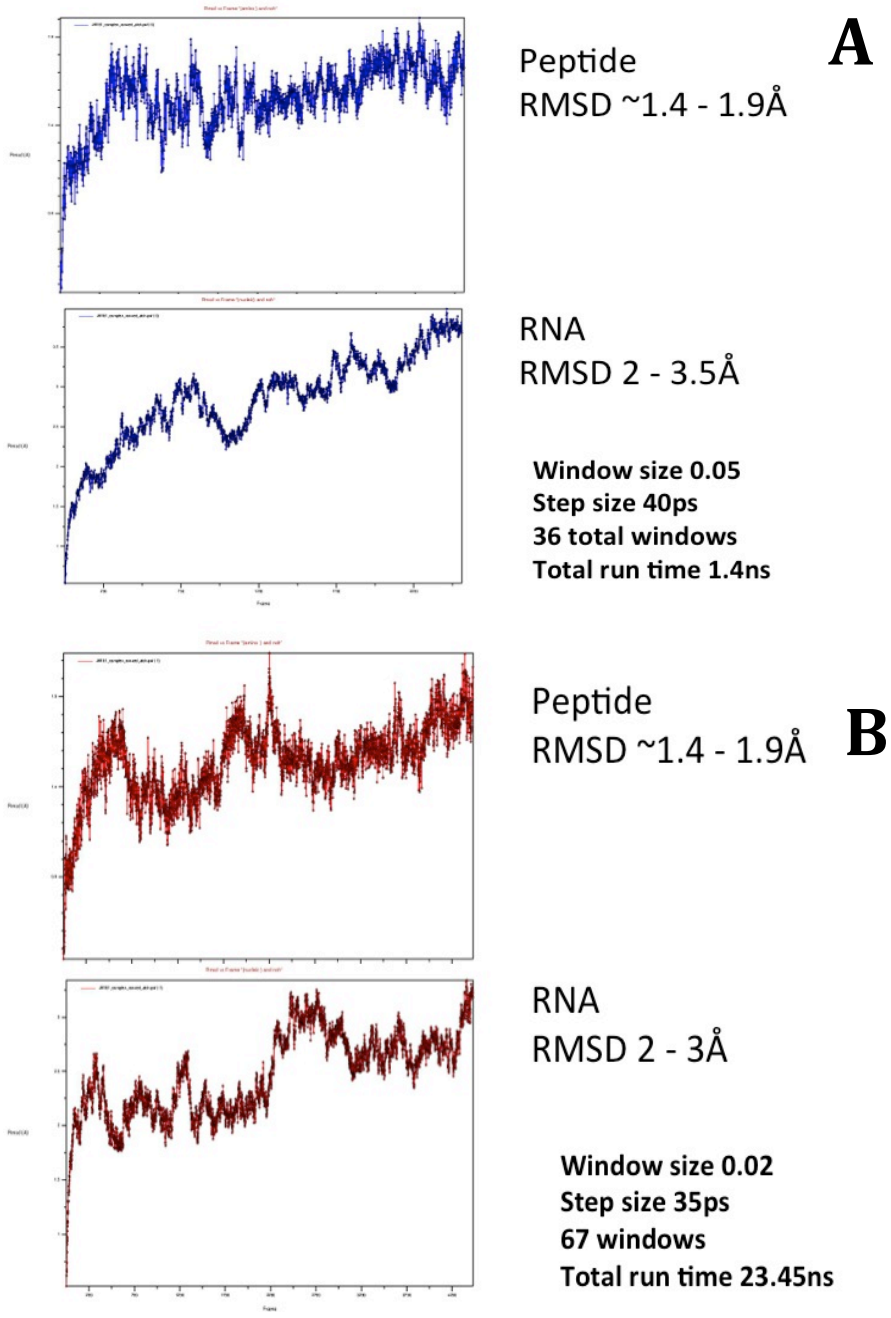


Figure 32: Frame plotted against RMSD values. A) The RMSD value increases as the frame number increases and despite changes in window and step size. B) The RMSD value still continues to increase for the RNA and peptide, despite changes in window and step size.

After personal communication with other members of the NAMD/VMD community, it was brought to our attention that the peptide, the RNA, and the final complex were inadequately equilibrated prior to calculations being run. Varying the equilibration time until an optimal time is reached would be necessary to fix this issue. The second red flag was the stability of the RNA structure as the simulation progressed. Within each frame, the lower portion of the HIV-1 TAR RNA helix appeared to be 'pulled apart', compromising the stability of the structure. This project was pursued in 2012, when calculations were run with the CHARMM27 force field. CHARMM27 is not optimized for RNA molecules, hence the lack of stability in our structures [74]. We attempted to utilize the AMBER force field, which is more suited for RNA, however AMBER (at the time) did not support FEP simulations. In 2014, a new CHARMM field, CHARMM36, was released to the public; this force field was specifically created to optimize the parameters for RNA.

We are currently exploring equilibration times and utilizing the recent advancements in CHARMM36 force field to fix those issues. Once we are able to optimize those parameters, we will perform FEP simulations on the peptides listed in Table 1, followed by further *in silico* optimization of the JB-181 peptide.

Section 7: Conclusion

We were able to independently model the JB-181 complex computationally to match the NMR resolved structure of JB-181 with HIV-1 TAR RNA. We have provided evidence that the FEP method is a viable tool for guiding peptide – RNA interaction optimization. With further development, the FEP method could help unlock predictive power of RNA- peptide interactions.

Chapter 5: Experimental

RNA preparation

Native HIV-1 TAR RNA was transcribed *in vitro* using desalted DNA templates purchased from Integrated DNA Technologies, Inc. (IDT) with in-house prepared T7 RNA polymerase. The following plasmid was ordered:

(5' - mGmGC AGA GAG CTC CCA GGC TCA GAT CTG CCT AA GTG AGT CGT ATT A – 3')

containing methylated guanosines at the 5' end to reduce the addition of n+1 and n+2 nucleotides at the end of the transcript by T7 RNA polymerase.

DNA oligonucleotides were centrifuged at 13,000 rpm for 30 seconds and then dissolved in water to a concentration of 8 μM. The following protocol was used to transcribe 4 mL of the HIV-1 TAR RNA:

RNAse free H_2O 1648 μL

1 M $MgCl_2$ 112 μL

Top strand DNA 400 μL

DNA template 400 μL

ATP 80 μL

CTP 80 μL

GTP 80 μL

UTP 80 μL

20X 200 μL

PEG 40% 800 µL

T7 120 µL

The RNAse free water, $MgCl_2$, Top and template DNA were combined and mixed in a 50 mL falcon tube, then heated for 5 min at 90°C in a heating block. The mixture was then allowed to cool to room temperature. dNTPs, 20X buffer (20x Tris-HCl (800 mM, pH 8.1) Triton X-100 (0.2%), spermidine (20 mM) and DTT (100 mM)), and 40% PEG were then added to the mixture, and it was allowed to incubate at 37°C for 5 min in a water bath. Immediately following the 5 minute incubation, T7 was added and the transcription was allowed to run for 4 hr at 37°C. Following transcription, 250 µL of 500 mM EDTA was added to the reaction for quenching. Evidence of RNA product was visible after 1 hr by the presence of inorganic phosphate precipitate.

Following the addition of EDTA, 1/10 transcription volume of 3 M sodium acetate and 3X the transcription of cold, 100% ethanol were added to the RNA, and mixed vigorously. The RNA was placed at -20.0°C overnight, and centrifuged at 4500 rpm for 45 min at -4.0°C the following day. The supernatant was poured off of the RNA pellet, followed by redissolving the pellet in equal amounts of water and 2X denaturing loading buffer (Novex 2X). The mixture was then loaded onto a 20% denaturing polyacrylamide gel (100 mL of 40% acrylamide: bisacrylamide 19:1, 80 g urea, 20 mL 10X TBE (890 mM Trizma base, 890 mM boric acid, 890 mM EDTA), 2.0 mL of 10% ammonium persulfate, and 200 µL of TEMED (BioRad). The gel was run with 1X TBE running buffer for 8 hr at 35 W. The RNA band was visualized with UV light (254 nm) against fluorescent background plates, and the gel containing the RNA was cut from the band.

The RNA was then removed from the gel via electroelution in 1X TBE buffer for 5 hr. The RNA was collected ever 1 hr, and at the end of the electroelution, the concentration was determined using UV absorption at 260 nm. At the end of the electro elution, 1/10 the total volume of electroeluted RNA of 3 M sodium acetate was added, in addition to 3X the aforementioned volume of cold, 100% ethanol. This was stored overnight at -20.0°C, and centrifuged at 4500 rpm at -4.0°C the following day. The supernatant was poured off, and the RNA pellet was resuspended in 1 mL of water.

The redissolved RNA was then run through a NAP-10 desalting column, and eluted with 1.5 mL of water. The concentration was determined again to ensure that recovery of the RNA from the desalting column was complete. The RNA was then lyophilized overnight, resuspended in 300 µL of water, and then microdialyzed against 10 mM potassium phosphate, pH 6.4 buffered solution.

HIV-1 TAR RNA containing mutations and 7SK RNA were synthesized, PAGE and HPLC purified by IDT.

Peptide Preparation

Peptides were prepared by Fmoc chemistry on an Applied Biosystems 433A peptide synthesizer using MBHA-Rink amide resin by our collaborator. Peptides were also ordered from Anaspec and further purified using a PD-10 column for desalting and buffer exchange.

EMSA

Purified HIV-1 TAR RNA was dephosphorylated with calf intestine phosphorylase (CIP) and then purified on a NAP-10 desalting column. RNA was then concentrated to 3 μM and the following protocol for rephosphorylation with ^{32}P-ATP was used.

HIV-1 TAR RNA 5 μL

PNK buffer 2 μL

Water 6 μL

^{32}P-ATP 6 μL

PNK enzyme 1 μL

The aforementioned reagents were mixed and incubated at 37°C for 30 min. 1 μL of 0.5 M EDTA was added to the mixture, and then incubated at 75°C for 10 min. The RNA was allowed to cool to room temperature.

Dephosphorylated BIV TAR RNA was prepared for EMSA with the same protocol as above.

For each binding assay, complex formation was evaluated in binding mix containing, or without tRNA. Briefly, 5 μL samples of peptide at varying concentrations were incubated with ^{32}ATP TAR RNA, and binding mix A and binding mix B. Both binding mixes contained 2.5 μL 0.05 M Tris-HCl, 0.05 M KCl, 0.2 M DTT, 1% Triton X-100 and 10 mg/mL yeast tRNA. Binding mix B did not contain tRNA.

Samples were allowed to sit on ice for 30 min. 10% glycerol was added to each mixture, then samples were run on a 12% acrylamide, nondenaturing gel for 1 hr

and 15 min at 35 W. The gels were dried for 30 min at 80°C and exposed overnight on a phosphor image plate. The plate was scanned with a phosphor scanner (Typhoon 9000) and autoradiograms were analyzed with ImageJ software.

Fluorescence Binding Assays

2-AP HIV TAR RNA was synthesized and PAGE and HPLC purified by IDT. RNA was dissolved in a binding buffer containing 10 mM phosphate, 50 mM NaCl, 0.1 mM EDTA, 1 mM $MgCl_2$, pH 6.8. For each binding assay, the 2-AP TAR RNA was added to a fluorescence appropriate cuvette and excited at 390 nm with a fluorescent spectrometer (Horiba). Emission was collected over a range of 400 to 600 nm. The wavelength corresponding to the strongest fluorescence intensity was used for the emission collection during the titration. Peptide was added incrementally until fully bound complex was achieved. This was evident when the fluorescence intensity was no longer changing. Fluorescent intensity was plotted against peptide concentration, and the binding curve was fit with Kaliedograph software.

NMR Spectroscopy and Spectral Assignments

NMR experiments were run on Bruker 500-, 600- and 800MHz spectrometers equipped with HCN cryo-probes. 2D ^{1}H ^{1}H total correlation spectroscopy (TOCSY) and ^{1}H ^{1}H nuclear overhauser effect spectroscopy (NOESY) experiments on the complex were collected in both D_2O and H_2O at both 25°C and 4°C, with mixing times ranging from 150 ms to 350 ms to facilitate H6/H8, H1', H2' and adenine H2 proton resonance assignments. Due to spectral overlap, ^{13}C 3D NOESY HMQC experiments were run to facilitate spectral assignment of RNA sugar proton H3', H4', H5' and H5" resonances. Peptide resonances were initially assigned sans RNA from

2D ^{1}H ^{1}H TOCSY and NOESY experiments, to confirm the typical α-sheet Ha to HN (i to i+1) NOES, and to gather Ha-Ha distance restraints. These assignments, along with those observed in the D_2O and H_2O NOESY spectra, were added to the peptide complex assignment tables. Data were processed using NMRPipe and Sparky.

Because of the strong similarity between the L-22 and JB-181 peptides, spectral assignments for both the RNA and peptide were guided by the L-22-TAR structure. Many of the protons for both complexes resonated in the same region in the NOESY spectra. While proton resonances of TAR complexed with JB-181 were independently assigned, backbone and ribose dihedral restraints were adopted from the L-22-TAR structure.

Structural Determination

The structure of JB-181 complexed with HIV-1 TAR was calculated with XPLOR-NIH software with torsion angle dynamics and simulated annealing from a single extended starting structure. Furthermore, we used 943 NOE-derived distance restraints from 2D NOESY spectra, as well as standard hydrogen bonding and base-pair planarity restraints for the positioning of unambiguously established base pairs, to calculate 100 structures and check for convergence until there were no NOE violations greater than 0.5 Å or dihedral violations greater than 5 degrees for the majority of structures. The restraints described in Table A1 (Appendix A) led to an overall root mean square deviation (RMSD) of 1.6 Å from the average structure. The top ten converged structures are shown superimposed in Figure 17.

Gamma amino acid peptoid synthesis (by Cai research group)

Fmoc protected α-amino acids and Knorr resin were obtained from Chem-Impex International, Inc. All other reagents and solvents were provided by either Sigma-Aldrich or Fisher Scientific. NMR spectra of intermediates and γ-AApeptide building blocks were obtained on a Varian Inova 400. γ-AApeptide sequences were prepared on a Knorr resin in peptide synthesis vessels on a Burrell Wrist-Action shaker. The γ-AApeptides were analyzed and purified on a Waters HPLC with both analytical and preparative modules, respectively, and the desired fractions were lyophilized using a Labconco lyophilizer. Molecular weights of γ-AApeptides were identified on a Bruker AutoFlex MALDI-TOF mass spectrometer.

Synthesis of γ-AApeptide building blocks

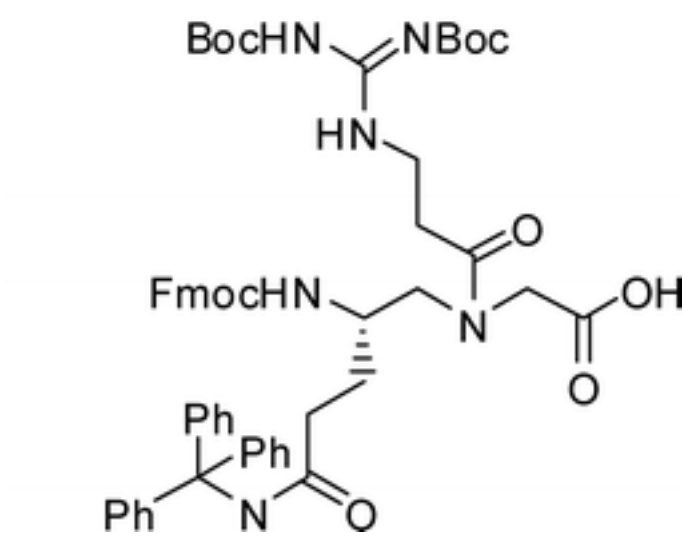
General procedure. Typical synthesis of **2**. To glycine benzyl ester hydrochloride in 20 mL methanol in a 100 mL round bottom flask was added 1.2 equiv. of triethylamine and stirred at 0°C for 15 min. Stoichiometric amount of a Fmoc protected amino acid aldehyde was added and the solution mixture was stirred for another 30 min. Catalytic amount of acetic acid was then added, followed by 2 equivalents of $NaBH_3CN$. The solution was allowed to stir at 0°C for 1 h and continue at room temperature overnight. The solvent was evaporated and 100 mL ethyl acetate and 100 mL saturated sodium bicarbonate solution were added to the residue. The organic layer was separated and washed with 100 mL brine, dried over anhydrous sodium sulfate, and removed in vacuo. Flash chromatography using ethyl acetate/hexane 1 : 1 gave **2** as a colorless oil.

Typical synthesis of **3**. Compound **2**, 1.2 equiv. of DIC, Oxohydroxybenzotriazole, and R_2COOH were stirred in 20 ml DMF overnight. The solution was then partitioned

in 100 ml ethyl acetate and 100 ml water. The organic layer was separated and washed with water (3 × 100 mL) and Brine (2 × 100 mL), dried over anhydrous sodium sulfate, and then concentrated in vacuo. Flash chromatography using ethyl acetate/hexane 1 : 3 gave **3** as a colorless oil.

Typical synthesis of **4**. **3** in 20 ml ethyl acetate was added to 10% Pd/C and hydrogenated at atmospheric pressure and room temperature overnight. The solution was evaporated and the residue was purified by flash chromatography 5–7% MeOH/CH_2Cl_2 to give **4** as a white foam solid.

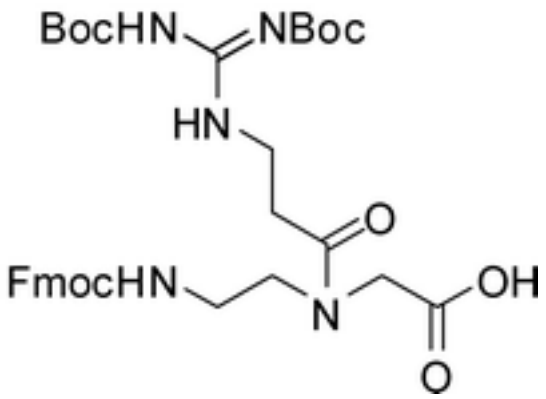
4a



Yield was 41% in three steps. ^{1}H NMR (400 MHz, DMSO-d_6) δ (two rotamers) 11.39 (m, 1H), 10.71 (d, J = 4.0 Hz, 1H), 8.66–8.60 (m, 1H), 8.53–8.50 (m, 1H), 8.29–8.23 (m, 1H), 7.85 (d, J = 8.0 Hz, 2H), 7.65–7.62 (m, 2H), 7.41–7.35 (m, 2H), 7.29–7.09 (m, 15H), 4.30–4.14 (m, 3H), 4.05–4.03 (m, 1H), 3.98–3.75 (m, 2H) 3.46–3.40 (m, 2H), 3.30–3.17 (m, 2H), 3.02–2.56 (m, 2H), 2.35–2.22 (m, 2H), 1.65–1.19 (m, 20H). ^{13}C NMR (100 MHz, DMSO-d_6) δ (two rotamers) 171.8, 170.8,

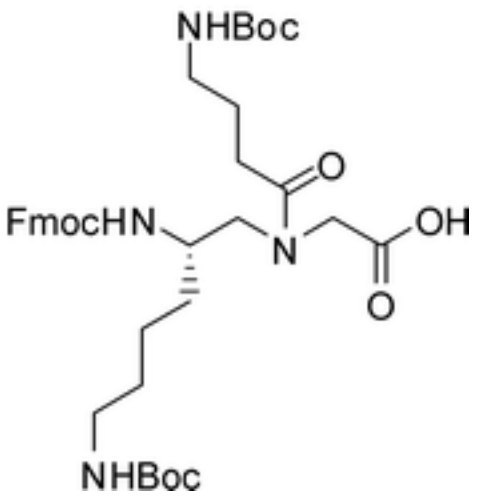
156.5, 155.3, 153.5, 152.1, 151.9, 145.3, 144.3, 144.2, 141.2, 128.9, 128.0, 127.8, 127.4, 126.7, 126.7, 125.5, 125.4, 120.6, 120.5, 84.0, 83.3, 69.6, 65.8, 56.4, 52.3, 49.9, 47.3, 47.2, 37.1, 33.0, 31.7, 28.4, 28.0, 27.97, 27.90, 18.97. HRMS for $[M+Na]^+$ Calc: 989.4420, found: 989.4428.

4b



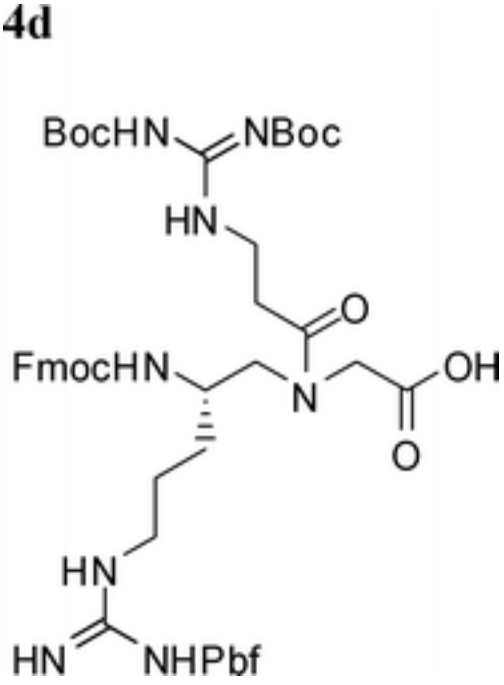
Yield was 38% in three steps. ^{1}H NMR (400 MHz, DMSO-d_6) δ (two rotamers) 11.43 (s, 1H), 8.58–8.50 (m, 1H), 7.86–7.80 (m, 2H), 7.64–7.61 (m, 2H), 7.40–7.25 (m, 4H), 4.25 (m, 3H), 2.75–3.67 (m, 2H), 3.55–3.42 (m, 5H), 3.15–3.10 (m, 2H), 2.55–2.54 (m, 1H), 2.45–2.42 (m, 1H), 1.42&1.40&1.34&1.33 (4 s, 18H) ^{13}C NMR (100 MHz, DMSO-d_6) δ (two rotamers) 172.3, 163.6, 156.52, 156.5, 152.3, 144.3, 143.0, 141.1, 139.9, 137.9, 129.4, 127.7, 127.3, 125.7, 129.4, 127.7, 127.5, 125.7, 120.5, 83.2, 79.63, 78.5, 65.9, 56.5, 53.4, 47.2, 38.6, 36.8, 32.5, 28.4, 28.04. HRMS for $[M+H]^+$ Calc: 654.3137, found: 654.3138.

4c



Yield was 69% in three steps. ^{1}H NMR (400 MHz, DMSO-d_6) δ (two rotamers) 7.84 (d, J = 8.0 Hz, 2H), 7.64–7.61 (dd, J = 4.0, 8.0 Hz, 2H), 7.37 (t, J = 8.0 Hz, 2H), 7.28 (t, J = 8.0 Hz, 2H), 7.14 & 6.98 (2d, J = 8.0 Hz, 1H), 6.74–6.69 (m, 1H), 4.29–4.14 (m, 3H), 3.40–3.81 (m, 2H), 3.62–3.52 (m, 2H), 3.28–3.16 (m, 1H), 2.95–2.69 (m, 4H), 2.36–2.22 (m, 1H), 2.12–2.01 (m, 1H), 1.70–1.13 (m, 26H). ^{13}C NMR (100 MHz, CD_3OH) δ (two rotamers) 174.7, 174.4, 171.5, 171.2, 157.32, 157.3, 157.1, 143.9, 143.8, 141.23, 141.20, 127.3, 126.7, 124.7, 124.6, 119.5 78.5, 78.4, 66.1, 65.9, 53.1, 51.2, 50.1, 49.9, 49.8, 39.8, 39.5, 39.3, 31.6, 31.2, 29.7, 29.5, 29.2, 27.4, 25.2, 22.9, 22.8. HRMS for $[M+H]^+$ Calc: 697.3807, found: 697.3796.

4d



Yield was 43% in three steps. ^{1}H NMR (400 MHz, DMSO-d_6) δ (two rotamers) 11.37 (s, 1H), 10.80 (d, J = 8.0 Hz, 1H), 8.83–8.77 (m, 1H), 8.36–8.32 (m, 1H), 7.83 (d, J = 8.0 Hz, 2H), 7.64–7.59 (m, 2H), 7.38–7.34 (m, 2H), 7.30–7.25 (m, 2H), 7.20–7.16 (m, 1H), 7.04 (d, J = 8.0 Hz, 1H), 6.72 (br.s, 1H), 6.40 (br.s, 1H), 4.28–4.11 (m, 4H), 4.10–3.81 (m, 2H), 3.64–3.54 (m, 1H), 3.53–3.15 (m, 3H), 3.04–2.94 (m, 2H), 2.90 (s, 2H), 2.76–2.62 (m, 2H), 2.50–2.42 (m, 4H), 2.39 (s, 3H), 1.96 (s, 3H), 1.49–1.19 (m, 28H). ^{13}C NMR (100 MHz, DMSO-d_6) δ (two rotamers) 172.2, 171.8, 163.5, 157.9, 156.6, 152.3, 144.4, 144.25, 144.23, 144.2, 141.2, 137.7, 134.7, 131.8, 127.98, 127.43, 125.6, 125.5, 124.7, 120.5, 116.7, 86.7, 83.2, 78.5, 65.7, 65.5, 55.3, 52.41, 52.39, 49.9, 47.9, 47.3, 47.2, 42.9, 36.96, 36.8, 32.5, 32.0, 29.8, 29.5, 28.7, 28.4, 28.0, 27.98, 26.2, 19.4, 18.0, 18.0, 12.7. HRMS for $[M+H]^+$ Calc: 1004.4677, found: 1004.4677.

Solid phase synthesis, purification and characterization of γ-AApeptides

The Tat 48–57 peptide **P1** was synthesized and analyzed by the USF peptide facility, and was used without further purification. The two γ-AApeptides were prepared on a Knorr resin in peptide synthesis vessels on a Burrell Wrist-Action shaker following the standard Fmoc chemistry of solid phase peptide synthesis protocol. Each coupling cycle included an Fmoc deprotection using 20% Piperidine in DMF, and 4 h coupling of 1.5 equiv of γ-AApeptide building blocks onto resin in the presence 2 equiv of DIC (diisopropylcarbodiimide)/Oxohydroxybenzotriazole in DMF. After the desired sequences were assembled, they were transferred into a 4 mL vial and cleaved from solid support in 48 : 50 : 2 TFA/CH_2Cl_2/triisopropylsilane overnight. Then solvent was evaporated and the residues were analyzed and purified on an analytical (1 mL min^{-1}) and a preparative Waters (20 mL min^{-1}) HPLC systems, respectively. The same methods were used by running 5% to 100% linear gradient of solvent B (0.1% TFA in acetonitrile) in A (0.1% TFA in water) over 40 min, followed by 100% solvent B over 10 min. The desired fractions were >70% in crude (as determined by HPLC) and eluted as single peaks at > 95% purity. They were collected and lyophilized. The molecular weights of γ-AApeptides and Tat peptide were obtained on Bruker AutoFlex MALDI-TOF mass spectrometer using α–cyano-4-hydroxy-cinnamic acid as the matrix.

Bibliography:

1. Gallego, J. and G. Varani, *Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges.* Acc Chem Res, 2001. **34**(10): p. 836-43.
2. Hermann, T. and E. Westhof, *RNA as a drug target: chemical, modelling, and evolutionary tools.* Curr Opin Biotechnol, 1998. **9**(1): p. 66-73.
3. Ivanova, G., et al., *Anti-HIV activity of steric block oligonucleotides.* Ann N Y Acad Sci, 2006. **1082**: p. 103-15.
4. Murchie, A.I., et al., *Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA.* J Mol Biol, 2004. **336**(3): p. 625-38.
5. Davis, B., et al., *Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic "hot spots".* J Mol Biol, 2004. **336**(2): p. 343-56.
6. Gadek, T.R. and R.S. McDowell, *Discovery of small molecule leads in a biotechnology datastream.* Drug Discov Today, 2003. **8**(12): p. 545-50.
7. Craik, D.J., et al., *The future of peptide-based drugs.* Chem Biol Drug Des, 2013. **81**(1): p. 136-47.
8. Selby, M.J., et al., *Structure, sequence, and position of the stem-loop in tar determine transcriptional elongation by tat through the HIV-1 long terminal repeat.* Genes Dev, 1989. **3**(4): p. 547-58.
9. Peterlin, B.M. and D.H. Price, *Controlling the elongation phase of transcription with P-TEFb.* Mol Cell, 2006. **23**(3): p. 297-305.
10. Karn, J., *Tackling Tat.* J Mol Biol, 1999. **293**(2): p. 235-54.
11. Jain, R., et al., *Limitations of current antiretroviral agents and opportunities for development.* Curr Pharm Des, 2006. **12**(9): p. 1065-74.
12. Le Douce, V., et al., *Achieving a cure for HIV infection: do we have reasons to be optimistic?* J Antimicrob Chemother, 2012. **67**(5): p. 1063-74.
13. Stevens, M., E. De Clercq, and J. Balzarini, *The regulation of HIV-1 transcription: molecular targets for chemotherapeutic intervention.* Med Res Rev, 2006. **26**(5): p. 595-625.
14. Mischiati, C., et al., *Aromatic polyamidines inhibiting the Tat-induced HIV-1 transcription recognize structured TAR-RNA.* Antisense Nucleic Acid Drug Dev, 2001. **11**(4): p. 209-17.
15. Hamma, T., et al., *Inhibition of HIV TAR-Tat interactions by an antisense oligo-2'-O-methylribonucleoside methylphosphonate.* Bioorg Med Chem Lett, 2003. **13**(11): p. 1845-8.
16. Arzumanov, A., et al., *Inhibition of HIV-1 Tat-dependent trans activation by steric block chimeric 2'-O-methyl/LNA oligoribonucleotides.* Biochemistry, 2001. **40**(48): p. 14645-54.
17. Kaushik, N., et al., *Anti-TAR polyamide nucleotide analog conjugated with a membrane-permeating peptide inhibits human immunodeficiency virus type 1 production.* J Virol, 2002. **76**(8): p. 3881-91.
18. Alper, P.B., et al., *Probing the specificity of aminoglycoside ribosomal RNA interactions with designed synthetic analogs.* Journal of the American Chemical Society, 1998. **120**(9): p. 1965-1978.
19. Tok, J.B.H. and R.R. Rando, *Simple aminols as aminoglycoside surrogates.* Journal of the American Chemical Society, 1998. **120**(32): p. 8279-8280.
20. Zapp, M.L., et al., *Modulation of the Rev-RRE interaction by aromatic heterocyclic compounds.* Bioorg Med Chem, 1997. **5**(6): p. 1149-55.

21. Wang, T.M., et al., *Synthesis of novel polyazadipyridinocyclophane scaffolds and their application for the generation of libraries.* Tetrahedron, 1998. **54**(28): p. 7955-7976.
22. Hamy, F., et al., *A new class of HIV-1 Tat antagonist acting through TAR-Tat inhibition.* Biochemistry, 1998. **37**(15): p. 5086-95.
23. Mei, H.Y., et al., *Inhibitors of protein-RNA complexation that target the RNA: specific recognition of human immunodeficiency virus type 1 TAR RNA by small organic molecules.* Biochemistry, 1998. **37**(40): p. 14204-12.
24. Lind, K.E., et al., *Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA.* Chem Biol, 2002. **9**(2): p. 185-93.
25. Stelzer, A.C., et al., *Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble.* Nat Chem Biol, 2011. **7**(8): p. 553-9.
26. Davidson, A., et al., *Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein.* Proceedings of the National Academy of Sciences, 2009. **106**(29): p. 11931-11936.
27. Athanassiou, Z., et al., *Structure-guided peptidomimetic design leads to nanomolar beta-hairpin inhibitors of the TAR-Tat interaction of bovine immunodeficiency virus.* Biochemistry, 2007. **46**(3): p. 741-51.
28. Bardaro, M.F., Jr., et al., *How binding of small molecule and peptide ligands to HIV-1 TAR alters the RNA motional landscape.* Nucleic Acids Res, 2009. **37**(5): p. 1529-40.
29. Lalonde, M.S., et al., *Inhibition of both HIV-1 reverse transcription and gene expression by a cyclic peptide that binds the Tat-transactivating response element (TAR) RNA.* PLoS Pathog, 2011. **7**(5): p. e1002038.
30. Moehle, K., et al., *Design of beta-hairpin peptidomimetics that inhibit binding of alpha-helical HIV-1 Rev protein to the rev response element RNA.* Angew Chem Int Ed Engl, 2007. **46**(47): p. 9101-4.
31. Nathans, R., et al., *Small-molecule inhibition of HIV-1 Vif.* Nat Biotechnol, 2008. **26**(10): p. 1187-92.
32. Huq, I., X. Wang, and T.M. Rana, *Specific recognition of HIV-1 TAR RNA by a D-Tat peptide.* Nat Struct Biol, 1997. **4**(11): p. 881-2.
33. Wang, X.L., I. Huq, and T.M. Rana, *Hiv-1 Tar Rna Recognition by an Unnatural Biopolymer.* Journal of the American Chemical Society, 1997. **119**(27): p. 6444-6445.
34. Huq, I., et al., *Controlling human immunodeficiency virus type 1 gene expression by unnatural peptides.* Biochemistry, 1999. **38**(16): p. 5172-7.
35. Gelman, M.A., et al., *Selective binding of TAR RNA by a Tat-derived beta-peptide.* Org Lett, 2003. **5**(20): p. 3563-5.
36. Hamy, F., et al., *An inhibitor of the Tat/TAR RNA interaction that effectively suppresses HIV-1 replication.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3548-53.
37. Huang, W., G. Varani, and G.P. Drobny, *13C/15N-19F intermolecular REDOR NMR study of the interaction of TAR RNA with Tat peptides.* J Am Chem Soc, 2010. **132**(50): p. 17643-5.
38. Niu, Y.H., et al., *gamma-AApeptides: design, synthesis and evaluation.* New Journal of Chemistry, 2011. **35**(3): p. 542-545.
39. Rapireddy, S., et al., *Strand invasion of mixed-sequence B-DNA by acridine-linked, gamma-peptide nucleic acid (gamma-PNA).* J Am Chem Soc, 2007. **129**(50): p. 15596-600.
40. Davidson, A., et al., *Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein.* Nucleic Acids Res, 2011. **39**(1): p. 248-56.

41. Tamilarasu, N., I. Huq, and T.M. Rana, *Targeting RNA with peptidomimetic oligomers in human cells.* Bioorg Med Chem Lett, 2001. **11**(4): p. 505-7.
42. Gregoire, C., et al., *Homonuclear (1)H-NMR assignment and structural characterization of human immunodeficiency virus type 1 Tat Mal protein.* Biopolymers, 2001. **62**(6): p. 324-35.
43. Hu, Y., et al., *Design and synthesis of AApeptides: a new class of peptide mimics.* Bioorg Med Chem Lett, 2011. **21**(5): p. 1469-71.
44. Aboulela, F., J. Karn, and G. Varani, *The Structure of the Human-Immunodeficiency-Virus Type-1 Tar Rna Reveals Principles of Rna Recognition by Tat Protein.* Journal of Molecular Biology, 1995. **253**(2): p. 313-332.
45. Aboulela, G., J. Karn, and G. Varani, *Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge (vol 24, pg 3974, 1996).* Nucleic Acids Research, 1996. **24**(22): p. 4598-4598.
46. Niu, Y.H., et al., *gamma-AApeptides bind to RNA by mimicking RNA-binding proteins.* Organic & Biomolecular Chemistry, 2011. **9**(19): p. 6604-6609.
47. Emery, F., et al., *A template for the solid-phase synthesis of conformationally restricted protein loop mimetics.* Chemical Communications, 1996(18): p. 2155-2156.
48. Favre, M., et al., *Structural mimicry of canonical conformations in antibody hypervariable loops using cyclic peptides containing a heterochiral diproline template.* Journal of the American Chemical Society, 1999. **121**(12): p. 2679-2685.
49. Athanassiou, Z., et al., *Structural mimicry of retroviral tat proteins by constrained beta-hairpin peptidomimetics: ligands with high affinity and selectivity for viral TAR RNA regulatory elements.* J Am Chem Soc, 2004. **126**(22): p. 6906-13.
50. Leeper, T.C., et al., *TAR RNA recognition by a cyclic peptidomimetic of Tat protein.* Biochemistry, 2005. **44**(37): p. 12362-72.
51. Bradrick, T.D. and J.P. Marino, *Ligand-induced changes in 2-aminopurine fluorescence as a probe for small molecule binding to HIV-1 TAR RNA.* RNA, 2004. **10**(9): p. 1459-68.
52. Sobhian, B., et al., *HIV-1 Tat Assembles a Multifunctional Transcription Elongation Complex and Stably Associates with the 7SK snRNP.* Molecular Cell, 2010. **38**(3): p. 439-451.
53. Durney, M.A. and V.M. D'Souza, *Preformed protein-binding motifs in 7SK snRNA: structural and thermodynamic comparisons with retroviral TAR.* J Mol Biol, 2010. **404**(4): p. 555-67.
54. Arien, K.K., G. Vanham, and E.J. Arts, *Is HIV-1 evolving to a less virulent form in humans?* Nat Rev Microbiol, 2007. **5**(2): p. 141-51.
55. Delling, U., et al., *Conserved nucleotides in the TAR RNA stem of human immunodeficiency virus type 1 are critical for Tat binding and trans activation: model for TAR RNA tertiary structure.* J Virol, 1992. **66**(5): p. 3018-25.
56. Michael, N.L., et al., *Naturally occurring genotypes of the human immunodeficiency virus type 1 long terminal repeat display a wide range of basal and Tat-induced transcriptional activities.* J Virol, 1994. **68**(5): p. 3163-74.
57. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
58. Zheng, Y.H., N. Lovsin, and B.M. Peterlin, *Newly identified host factors modulate HIV replication.* Immunol Lett, 2005. **97**(2): p. 225-34.
59. DeJong, E.S., B. Luy, and J.P. Marino, *RNA and RNA-protein complexes as targets for therapeutic intervention.* Curr Top Med Chem, 2002. **2**(3): p. 289-302.

60. Gelus, N., et al., *Inhibition of HIV-1 TAR-Tat interaction by diphenylfuran derivatives: effects of the terminal basic side chains.* Bioorg Med Chem, 1999. **7**(6): p. 1089-96.
61. Srinivas, N., et al., *Peptidomimetic antibiotics target outer-membrane biogenesis in Pseudomonas aeruginosa.* Science, 2010. **327**(5968): p. 1010-3.
62. Reyes, C.M. and P.A. Kollman, *Molecular dynamics studies of U1A-RNA complexes.* RNA, 1999. **5**(2): p. 235-44.
63. Wang, J., et al., *Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA.* J Am Chem Soc, 2001. **123**(22): p. 5221-30.
64. Gouda, H., et al., *Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods.* Biopolymers, 2003. **68**(1): p. 16-34.
65. Miyamoto, S. and P.A. Kollman, *Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches.* Proteins, 1993. **16**(3): p. 226-45.
66. Bash, P.A., et al., *Calculation of the relative change in binding free energy of a protein-inhibitor complex.* Science, 1987. **235**(4788): p. 574-6.
67. Kollman, P., *Free energy calculations: applications to chemical and biochemical phenomena.* Chem. Rev., 1993. **93**: p. 2395 - 2417.
68. McCammon, T.P.S.a.J.A., *Computational Alchemy.* Annu. Rev. Phys. Chem., 1992(43): p. 407 - 435.
69. Liu, P.D., F.; Cai, W.; Chipot, C., *A toolkit for the analysis of free-energy perturbation calcuations.* J. Chem. Theory Comput., 2012. **8**(8): p. 2606 - 2616.
70. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J Comput Chem, 2005. **26**(16): p. 1781-802.
71. Dixit, S.C., C., *A tutorial to set up alchemical free energy perturbation calculations in NAMD*, 2002.
72. Gapsys, V., et al., *Calculation of binding free energies.* Methods Mol Biol, 2015. **1215**: p. 173-209.
73. Phillips, J., *NAMD Tutorial: Unix/MacOSX Version*, 2012: University of Illinois at Urbana-Champaign.
74. Malolepsza, E.S., B.; Khalili, M.; Trygubenko, S.; Fejer, S.; Wales, D., *Symmetrization of the AMBER and CHARMM force fields.* J Comput Chem, 2009(31): p. 1402 - 1409.

Appendix A

Table A1: Experimental restraints and structural statistics for the JB-181 complex HIV-1 TAR RNA structure.

Total number of restraints	2157
NOE-derived restraints	943
Intermolecular NOEs	168
RNA (intramolecular)	486
Peptide (intramolecular)	323
Dihedral restraints	173
Hydrogen-bonding restraints	53
Planarity restraints	11
Average rmsd values from experimental restraints	
Distance, Å	0.09
Dihedral,°	1.5
Average rmsd values from ideal geometries	
Bonds, Å	0.007
Angles, °	1.1
Improper, °	1.0
Heavy atom rmsd from mean structure, Å	
JB-181 (all backbone atoms)	0.29
JB-181 (all heavy atoms)	1.5
TAR RNA (all heavy atoms)	1.6
TAR RNA core (G18–U23, G26–C29, G36–C44)	0.83
JB-181 and TAR (heavy atoms)	0.86
Entire structure (heavy atoms)	1.6

Dual topology files for FEP minimizations:

Topology for diamino butyric acid

```
RESI DAB         1.00
GROUP
ATOM N    NH1    -0.47
ATOM HN   H       0.31
ATOM CA   CT1     0.07
ATOM HA   HB      0.09
GROUP
ATOM CB   CT2    -0.18
ATOM HB1  HA      0.09
ATOM HB2  HA      0.09
GROUP
ATOM CG   CT2     0.21
ATOM HG1  HA      0.05
ATOM HG2  HA      0.05
ATOM ND   NH3    -0.30
ATOM HD1  HC      0.33
```

```
ATOM HD2  HC      0.33
ATOM HD3  HC      0.33
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB CA   CG CB   ND CG
BOND N  HN   N  CA   C  CA
BOND C  +N   CA HA   CB HB1  CB HB2  CG HG1
BOND CG HG2
DOUBLE   O  C
BOND ND HD1  ND HD2  ND HD3
IMPR N -C CA HN  C CA +N O
CMAP -C  N  CA  C   N  CA  C  +N
DONOR HN N
DONOR HZ1 NZ
DONOR HZ2 NZ
DONOR HZ3 NZ
ACCEPTOR O C
IC -C   CA   *N   HN    1.3482 123.5700  180.0000 115.1100  0.9988
IC -C   N    CA   C     1.3482 123.5700  180.0000 107.2900  1.5187
IC N    CA   C    +N    1.4504 107.2900  180.0000 117.2700  1.3478
IC +N   CA   *C   O     1.3478 117.2700  180.0000 120.7900  1.2277
IC CA   C    +N   +CA   1.5187 117.2700  180.0000 124.9100  1.4487
IC N    C    *CA  CB    1.4504 107.2900  122.2300 111.3600  1.5568
IC N    C    *CA  HA    1.4504 107.2900 -116.8800 107.3600  1.0833
IC N    CA   CB   CG    1.4504 111.4700  180.0000 115.7600  1.5435
IC CG   CA   *CB  HB1   1.5435 115.7600  120.9000 107.1100  1.1146
IC CG   CA   *CB  HB2   1.5435 115.7600 -124.4800 108.9900  1.1131
IC CA   CB   CG   ND    1.5397 112.3300  180.0000 110.4600  1.4604
IC ND   CB   *CG  HG1   1.4604 110.4600  119.9100 110.5100  1.1128
IC ND   CB   *CG  HG2   1.4604 110.4600 -120.0200 110.5700  1.1123
IC CB   CG   ND   HD1   1.5350 110.4600  179.9200 110.0200  1.0404
IC HD1  CG   *ND  HD2   1.0404 110.0200  120.2700 109.5000  1.0402
IC HD1  CG   *ND  HD3   1.0404 110.0200 -120.1300 109.4000  1.0401

# Topology for norarginine

RESI NOR         1.00
GROUP
ATOM N    NH1    -0.47
ATOM HN   H       0.31
ATOM CA   CT1     0.07
ATOM HA   HB      0.09
GROUP
ATOM CB   CT2    -0.18
ATOM HB1  HA      0.09
ATOM HB2  HA      0.09
GROUP
ATOM CG   CT2     0.20
ATOM HG1  HA      0.09
ATOM HG2  HA      0.09
ATOM ND   NC2    -0.70
ATOM HD   HC      0.44
```

```
ATOM CE   C       0.64
ATOM NH1  NC2    -0.80
ATOM HH11 HC      0.46
ATOM HH12 HC      0.46
ATOM NH2  NC2    -0.80
ATOM HH21 HC      0.46
ATOM HH22 HC      0.46
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB CA  CG CB  ND CG  CE ND
BOND NH2 CE  N  HN  N  CA
BOND C   CA  C  +N  CA HA  CB HB1
BOND CB HB2  CG HG1  CG HG2  ND HD
BOND ND HD  NH1 HH11  NH1 HH12  NH2 HH21  NH2 HH22
DOUBLE O  C    CE  NH1
IMPR N  -C  CA  HN   C CA +N O
CMAP -C  N  CA  C   N  CA  C  +N
IMPR CE NH1 NH2 ND
DONOR HN N
DONOR HD ND
DONOR HH11 NH1
DONOR HH12 NH1
DONOR HH21 NH2
DONOR HH22 NH2
ACCEPTOR O C
IC -C   CA   *N   HN    1.3496 122.4500  180.0000 116.6700  0.9973
IC -C   N    CA   C     1.3496 122.4500  180.0000 109.8600  1.5227
IC N    CA   C    +N    1.4544 109.8600  180.0000 117.1200  1.3511
IC +N   CA   *C   O     1.3511 117.1200  180.0000 121.4000  1.2271
IC CA   C    +N   +CA   1.5227 117.1200  180.0000 124.6700  1.4565
IC N    C    *CA  CB    1.4544 109.8600  123.6400 112.2600  1.5552
IC N    C    *CA  HA    1.4544 109.8600 -117.9300 106.6100  1.0836
IC N    CA   CB   CG    1.4544 110.7000  180.0000 115.9500  1.5475
IC CG   CA   *CB  HB1   1.5475 115.9500  120.0500 106.4000  1.1163
IC CG   CA   *CB  HB2   1.5475 115.9500 -125.8100 109.5500  1.1124
IC CA   CB   CG   ND    1.5552 115.9500  180.0000 114.0100  1.5384
IC ND   CB   *CG  HG1   1.5384 114.0100  125.2000 108.5500  1.1121
IC ND   CB   *CG  HG2   1.5384 114.0100 -120.3000 108.9600  1.114
IC CB   CG   ND   CE    1.5475 114.0100  180.0000 107.0900  1.5034
IC CE   CG   *ND  HD    1.5034 107.0900  120.6900 109.4100  1.1143
IC CG   ND   CE   NH1   1.5034 123.0500  180.0000 118.0600  1.3311
IC ND   CE   NH1  HH11  1.3401 118.0600 -178.2800 120.6100  0.9903
IC HH11 CE   *NH1 HH12  0.9903 120.6100  171.1900 116.2900  1.0023
IC NH1  ND   *CE  NH2   1.3311 118.0600  178.6400 122.1400  1.3292
IC ND   CE   NH2  HH21  1.3401 122.1400 -174.1400 119.9100  0.9899
IC HH21 CE   *NH2 HH22  0.9899 119.9100  166.1600 116.8800  0.9914

# Topology for D-proline

RESI DPR         0.00
GROUP               !      HD1 HD2
ATOM N    N      -0.29
```

```
ATOM CD   CP3     0.00
ATOM HD1  HA      0.09
ATOM HD2  HA      0.09
ATOM CA   CP1     0.02
ATOM HA   HB      0.09
GROUP
ATOM CB   CP2    -0.18
ATOM HB1  HA      0.09
ATOM HB2  HA      0.09
GROUP
ATOM CG   CP2    -0.18
ATOM HG1  HA      0.09
ATOM HG2  HA      0.09
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND C  CA  C   +N
BOND N  CA  CA  CB  CB  CG  CG  CD  N   CD
BOND HA CA  HG1 CG  HG2 CG  HD1 CD  HD2 CD  HB1 CB  HB2 CB
DOUBLE   O  C
IMPR N -C CA CD
IMPR C CA +N O
CMAP -C  N  CA  C   N  CA  C  +N
ACCEPTOR O C
IC -C   CA   *N   CD    1.3366 122.9400  178.5100 112.7500  1.4624
IC -C   N    CA   C     1.3366 122.9400  -76.1200 110.8600  1.5399
IC N    CA   C    +N    1.4585 110.8600  180.0000 114.7500  1.3569
IC +N   CA   *C   O     1.3569 114.7500  177.1500 120.4600  1.2316
IC CA   C    +N   +CA   1.5399 116.1200  180.0000 124.8900  1.4517
IC N    C    *CA  CB    1.4585 110.8600 -113.7400 111.7400  1.5399
IC N    C    *CA  HA    1.4585 110.8600  122.4000 109.0900  1.0837
IC N    CA   CB   CG    1.4585 102.5600   31.6100 104.3900  1.5322
IC CA   CB   CG   CD    1.5399 104.3900  -34.5900 103.2100  1.5317
IC N    CA   CB   HB1   1.4585 102.5600  -84.9400 109.0200  1.1131
IC N    CA   CB   HB2   1.4585 102.5600  153.9300 112.7400  1.1088
IC CA   CB   CG   HG1   1.5399 104.3900 -156.7200 112.9500  1.1077
IC CA   CB   CG   HG2   1.5399 104.3900   81.2600 109.2200  1.1143
IC CB   CG   CD   HD1   1.5322 103.2100  -93.5500 110.0300  1.1137
IC CB   CG   CD   HD2   1.5322 103.2100  144.5200 110.0000  1.1144
PATCHING FIRS PROP


# Topology for arginine to norarginine transformation.

RESI R2N         1.00
GROUP
ATOM N    NH1    -0.47
ATOM HN   H       0.31
ATOM CA   CT1     0.07
ATOM HA   HB      0.09
GROUP
ATOM CB   CT2    -0.18
ATOM HB1  HA      0.09
```

```
ATOM HB2  HA      0.09
GROUP
ATOM CGA  CT2    -0.18
ATOM HG1A HA      0.09
ATOM HG2A HA      0.09
GROUP
ATOM CDA  CT2     0.20
ATOM HD1A HA      0.09
ATOM HD2A HA      0.09
GROUP
ATOM NEA  NC2    -0.70
ATOM HEA  HC      0.44
GROUP
ATOM CZA  C       0.64
ATOM NH1  NC2    -0.80
ATOM HH11 HC      0.46
ATOM HH12 HC      0.46
ATOM NH2  NC2    -0.80
ATOM HH21 HC      0.46
ATOM HH22 HC      0.46
GROUP
ATOM CGB  CT2     0.20
ATOM HG1B HA      0.09
ATOM HG2B HA      0.09
GROUP
ATOM NDB  NC2    -0.70
ATOM HDB  HC      0.44
GROUP
ATOM CEB  C       0.64
ATOM NZ1  NC2    -0.80
ATOM HZ11 HC      0.46
ATOM HZ12 HC      0.46
ATOM NZ2  NC2    -0.80
ATOM HZ21 HC      0.46
ATOM HZ22 HC      0.46
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND N  HN
BOND N  CA
BOND C  CA
BOND C  +N
BOND CA  HA
BOND CB  CA
BOND CGA  CB
BOND CDA  CGA
BOND NEA  CDA
BOND CZA  NEA
BOND NH2  CZA
BOND CB  HB1
BOND CB  HB2
BOND CGA  HG1A
BOND CGA  HG2A
```

```
BOND CDA  HD1A
BOND CDA  HD2A
BOND NEA  HEA
BOND NH1  HH11
BOND NH1  HH12
BOND NH2  HH21
BOND NH2  HH22
BOND CGB  CB
BOND NDB  CGB
BOND CEB  NDB
BOND NZ2  CEB
BOND CGB  HG1B
BOND CGB  HG2B
BOND NDB  HDB
BOND NZ1  HZ11
BOND NZ1  HZ12
BOND NZ2  HZ21
BOND NZ2  HZ22
DOUBLE O  C
DOUBLE CZA  NH1
DOUBLE CEB  NZ1
IMPR N  -C  CA  HN  C  CA  +N  O
CMAP -C  N  CA  C   N  CA  C  +N
IMPR CZA NH1 NH2 NEA
IMPR CEB NZ1 NZ2 NDB
DONOR HN N
DONOR HEA NEA
DONOR HH11 NH1
DONOR HH12 NH1
DONOR HH21 NH2
DONOR HH22 NH2
DONOR HZ11 NZ1
DONOR HZ12 NZ1
DONOR HZ21 NZ2
DONOR HZ22 NZ2
ACCEPTOR O C
END


# Topology for arginine to diamond butyric acid transformation

RESI R2B         1.00
GROUP
ATOM N    NH1    -0.47
ATOM HN   H       0.31
ATOM CA   CT1     0.07
ATOM HA   HB      0.09
GROUP
ATOM CB   CT2    -0.18
ATOM HB1  HA      0.09
ATOM HB2  HA      0.09
GROUP
ATOM CGA  CT2    -0.18
```

```
ATOM HG1A HA      0.09
ATOM HG2A HA      0.09
GROUP
ATOM CDA  CT2     0.20
ATOM HD1A HA      0.09
ATOM HD2A HA      0.09
GROUP
ATOM NEA  NC2    -0.70
ATOM HEA  HC      0.44
GROUP
ATOM CZA  C       0.64
ATOM NH1  NC2    -0.80
ATOM HH11 HC      0.46
ATOM HH12 HC      0.46
ATOM NH2  NC2    -0.80
ATOM HH21 HC      0.46
ATOM HH22 HC      0.46
GROUP
ATOM CGB  CT2     0.21
ATOM HG1B HA      0.05
ATOM HG2B HA      0.05
GROUP
ATOM NDB  NH3    -0.30
ATOM HD1B HC      0.33
ATOM HD2B HC      0.33
ATOM HD3B HC      0.33
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB  CA
BOND CGA CB
BOND CDA CGA
BOND NEA CDA
BOND CZA NEA
BOND NH2 CZA
BOND N  HN
BOND N  CA
BOND C  CA
BOND C  +N
BOND CA HA
BOND CB HB1
BOND CB HB2
BOND CGA HG1A
BOND CGA HG2A
BOND CDA HD1A
BOND CDA HD2A
BOND NEA  HEA
BOND NH1 HH11
BOND NH1 HH12
BOND NH2 HH21
BOND NH2 HH22
BOND CGB CB
BOND NDB  CGB
```

```
BOND CGB HG1B
BOND CGB HG2B
BOND NDB HD1B
BOND NDB HD2B
BOND NDB HD3B
DOUBLE O  C
DOUBLE CZA  NH1
IMPR N  -C  CA  HN   C CA +N O
CMAP -C  N  CA  C   N  CA  C  +N
IMPR CZA NH1 NH2 NEA
DONOR HN N
DONOR HEA NEA
DONOR HH11 NH1
DONOR HH12 NH1
DONOR HH21 NH2
DONOR HH22 NH2
DONOR HD1 ND
DONOR HD2 ND
DONOR HD3 ND
ACCEPTOR O C
```

Part 2: Structure of the Conserved Core Region of lincRNA Cyrano

Chapter 1: Long Intervening Noncoding RNAs

The central dogma of molecular biology states that ribonucleic acid (RNA) functions as an intermediate between genomic DNA sequence and the proteins it encodes. Three classes of RNAs, messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), are key in the process of protein expression. While mRNA encodes for protein, tRNA and rRNA are noncoding functional transcripts. These noncoding RNAs were once considered to be outliers in the transcriptome, as it was primarily believed that most RNAs were protein-encoding [1, 2]. Over the past few decades, however, it has been determined that <1.5% of mammalian genomes encode proteins, while 80% of RNA are estimated to be noncoding, functional transcripts [3-10]. Many types of noncoding RNAs have been identified over the past several decades and can be categorized according to size. For example, miRNAs and siRNAs are between 18 and 25 nucleotides in length [11-13], snoRNAs and piRNAs are between 20 and 200 nucleotides [14, 15], and lncRNAs are greater than 200 nucleotides in length [16-20]. Of these, my specific interest is long noncoding RNAs (lncRNAs) (Figure 1).

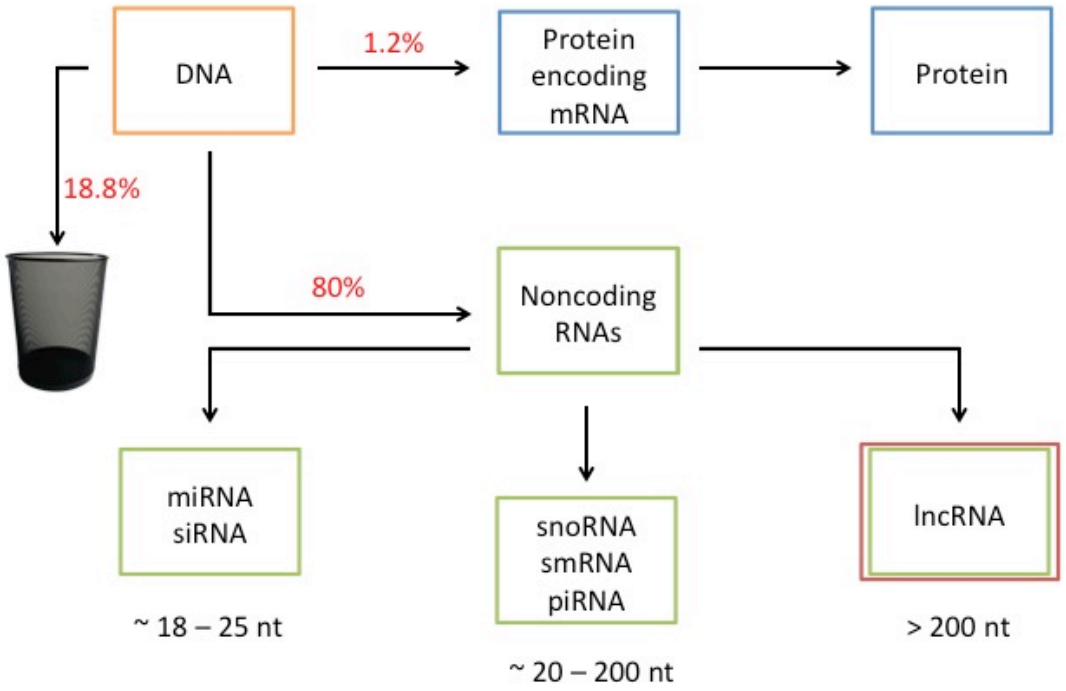


Figure 1: An estimated 1.2% of mammalian RNA encodes for protein, while 80% are functional, noncoding RNAs. 18.8% of RNAs are estimated to be genomic 'noise'.

LncRNAs are non-protein encoding transcripts longer than 200 nucleotides in length. These RNAs are primarily found in the nucleus and are transcribed by RNA polymerase II, like mRNAs [21-23]. They have marginal sequence conservation and often overlap with protein-encoding genes [23]. The abundance level of lncRNAs is relatively low: recent reports have suggested that, for some lncRNAs, copy number is as low as 0.0006 copies per cell [24, 25]. Despite their low transcriptional abundance, a number of lncRNAs have demonstrated significant biological function. For example, XIST, a 19 kb noncoding RNA transcript, engenders heterochromatization of one of the two X chromosomes in females [26-29]. A more recently discovered noncoding RNA is HOTAIR, a 2 kb noncoding RNA, involved in physiological epidermal development, and represses tumor and metastasis

suppression genes in cancer development [18, 30-33]. Many other lncRNAs have demonstrated a wide repertoire of activity; they often function epigenetically, target chromatin-remodeling enzymes whose activity is non-specific to specific genomic loci [1, 34-52].

Identification of lincRNAs

Despite the vast range of functionality that has been discovered among lncRNAs, the identification of lncRNAs is very challenging; their overlap with protein-encoding genes and a lack of clearly defined open reading frames (ORFs) make it difficult to identify them throughout the genome [21, 22]. However, a subclass of stand-alone lncRNAs, known as long intervening noncoding RNAs (lincRNAs), are more easily identified because they are capped and polyadenlyated, like mRNAs. These lincRNAs do not overlap with protein-encoding regions, usually contain multiple exons, and are subject to alternative splicing [21, 24, 53].

Many large-scale efforts have been utilized to accurately identify lincRNA transcripts through the genomic positions of the start-, splice-, and polyadenylation sites (Figure 2).

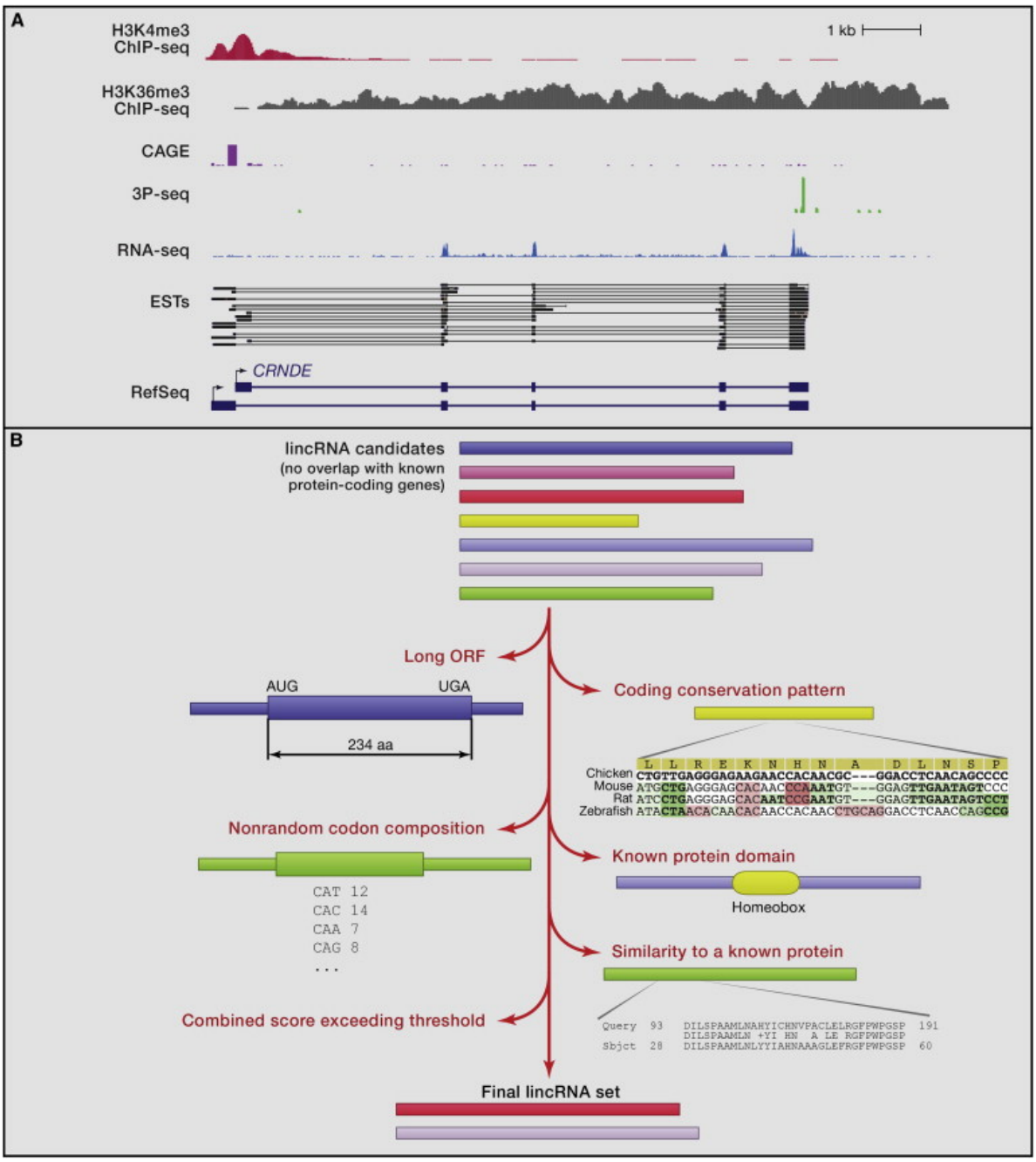


Figure 2: General overview for assembling lincRNA collections. A) The data sourcess useful for constructing lincRNA transcript models. B) A generic lincRNA annotation pipeline [21].

Chromatin immunoprecipitation, followed by high-throughput sequencing (ChIP-Seq), allows for the generation of genome wide chromatin-state maps that can be

used to identify those sites [53]. These include histone H3 lysine 4 trimethylation maps (H3K4me3), which identifies promoters of genes transcribed by RNA polymerase II, and histone H3 lysine 36 trimethylation maps (H3K36me3), which mark the bodies of the genes transcribed by RNA polymerase II. 3P-sequencing defines poly(A) tail positioning, a characteristic of lincRNAs. Implementation of these methods, integrated with transcriptome data sets (RNA-Seq reads, annotated expressed sequence tags (ESTs), and full-length cDNAs) has led to the discovery of approximately 38,000 lincRNAs such as H19, MALAT1, and of specific interest to me, Cyrano [21, 22, 54-64].

Chapter 2: The lincRNA Cyrano

Cyrano is a 4.5kb lincRNA transcript discovered in *Danio* rerio, more commonly known as zebrafish. It is found within chromosome 13 and is convergent with *oip5* gene transcription. While Cyrano is in a synteny block containing the protein encoding genes *nusap-1, ndufaf-1* and *rtf-1*, ribosome profiling in HeLa cells showed that Cyrano is not translated [65]. After splicing, Cyrano lincRNA is comprised of 3 exons, and is expressed in the central nervous system. At 24 hr post fertilization (hpf) it is localized in the brain and notochord, and at 72 hpf, it is localized in the spinal cord [53]. Like many lincRNAs, Cyrano is not highly abundant; in the CNS it is estimated that 20 – 40 copies of the lincRNA are present per cell, which is very abundant for a lincRNA (unpublished data).

Cyrano lincRNA is conserved in a wide range of species, spanning from human to lamprey [21, 53]. While there is little sequence conservation across the entire transcript, there are a few regions of high conservation, including a 300-nt region containing a core of 26 nearly perfectly conserved bases (sequence 5' – AACAACAAAAUCACCAAUGUCUUCCA -3') (only two nucleotides are not conserved) across at least 52 vertebrates (Figure 3). A BLAST search of this 26 nucleotide conserved region against the drosophila genome provided by the NCBI database, revealed that conservation is maintained in this species as well (Figure 4).

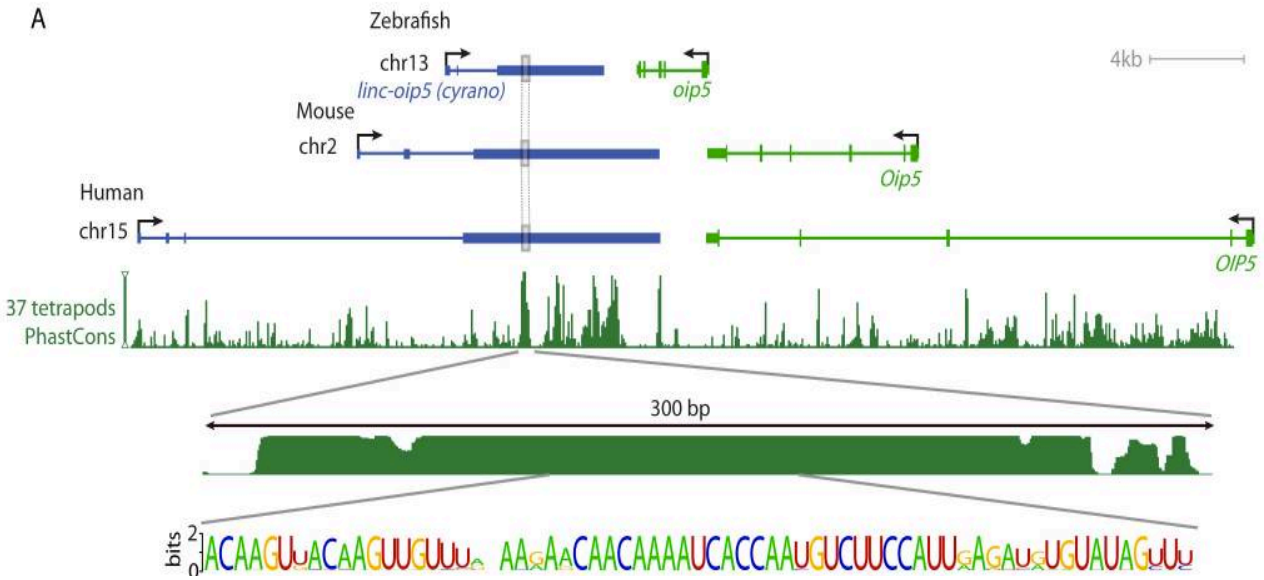


Figure 3: Cyrano is located in chr13, chr2 and chr15 of zebrafish, mouse, and human, respectively. In an alignment of 37 tetrapods, a highly conserved 300-nucleotide region was found. Within this 300-nucleotide region, a sub-region containing 26 nucleotides is very highly conserved in every tetrapod, with all but two nucleotides perfectly conserved [53].

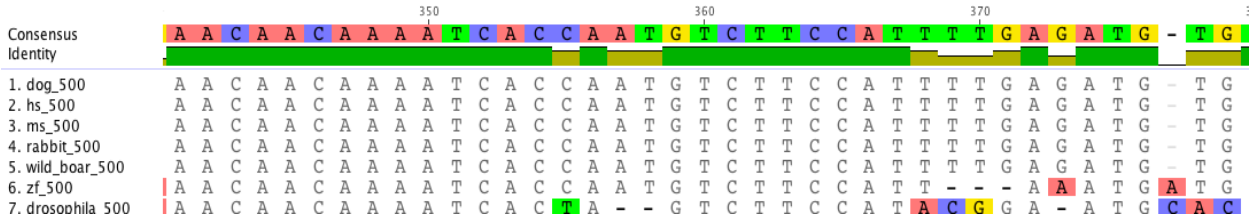


Figure 4: An alignment of *Drosophila melanogaster* (drosophila_500) with other mammalian species reveals nearly perfect sequence conservation within the 26 nucleotide core region.

Interestingly, with the exception of 4 nucleotides, the core 26 nucleotide region is complementary with mature miR-7, an RNA that is present in zebrafish brain and perfectly conserved from annelids to humans [66]. While it is not known how and if mir-7 RNA interacts with Cyrano, it has been hypothesized that either mir-7 regulates Cyrano, Cyrano regulates mir-7, or Cyrano and mir-7 RNA collaborate in some unknown downstream function [53].

Effects of perturbing bases within the 300-nucleotide conserved region

Using morpholino antisense oligonucleotides (MOs) targeting both Cyrano's splice site and the 26 nucleotide conserved region (Figure 5), it was determined that Cyrano is required for proper morphogenesis and neurogenesis in zebrafish. When the splice site was blocked by an MO, Cyrano transcript levels were reduced by 70% (Figure 5C). Injecting embryos with either splice- or conserved-site MOs caused zebrafish phenotypes to develop with several obvious deformities. These deformities included: smaller than usual heads and eyes, short curly tails, defects in the neural tube opening, loss of NeuroD-positive neurons in the retina and tectum, and enlarged nasal placodes (Figure 5D, E, F). Embryos injected with an MO containing five mismatch pairs complementary to the conserved region, along with MOs complementary to non-conserved regions of Cyrano lincRNA, did not affect the embryonic development of the zebrafish [53].

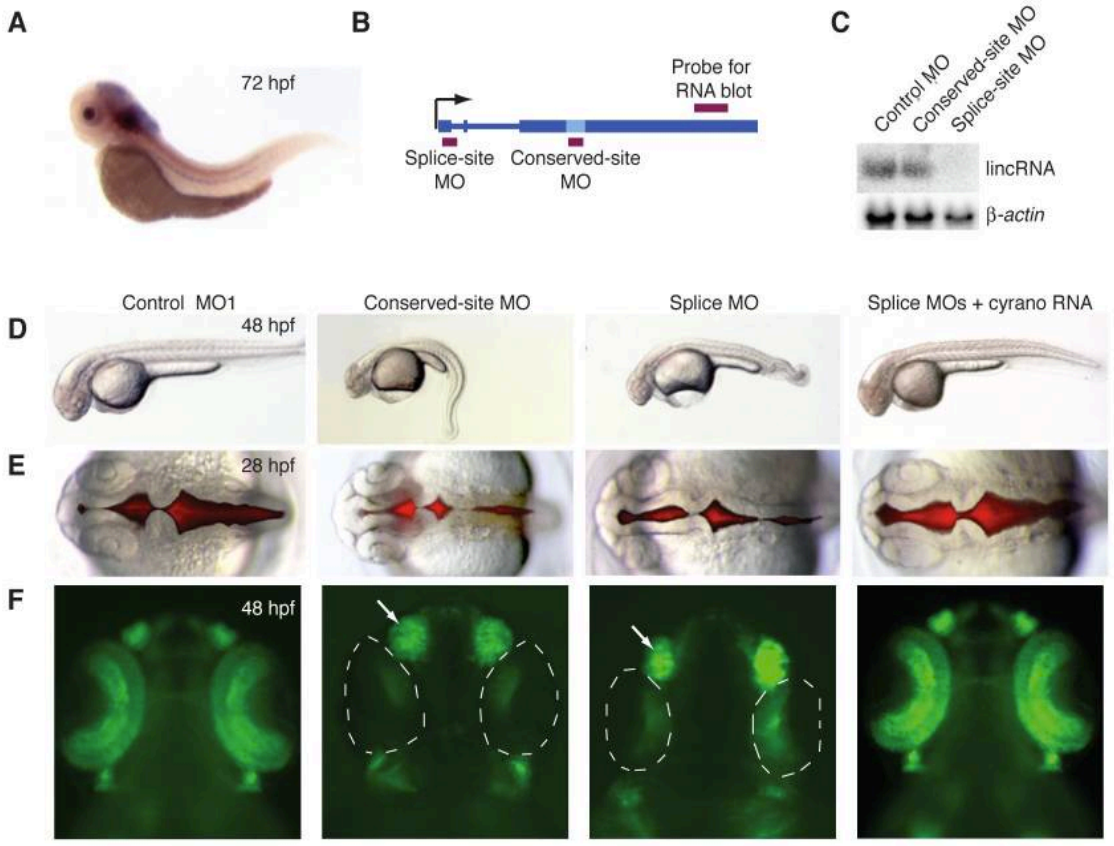


Figure 5: A) A healthy zebrafish embryo at 72 hr post fertilization (hpf). B) The location of the morpholino antisense oligonucleotides designed for the studying the function of Cyrano. C) demonstrates the effect of Cyrano transcript levels when the splice site is blocked; no transcript is present when the splice site is blocked. D- F) The effects of the various MOs on zebrafish embryos involving the tail, notochord, and nasal placodes [53].

Co-injection of full-length spliced Cyrano RNA, from zebrafish, human, or mouse, in MO-treated zebrafish rescued morphant phenotypes to varying, but substantial, degrees (60%, 60%, and 35%, respectively) [53]. Co-injection of RNAs containing mutations throughout the conserved region significantly reduced the rescuing potential of the RNAs. However, co-injection of just the 300-nucleotide conserved region (zebrafish, human, or mouse) (personal communication) did not result in a

rescue of the morphant phenotypes. Co-injection of the 67-nucleotide conserved region, containing 32 flanking bases on each side, were also not able to rescue MO treated zebrafish [53]. Only when fully spliced, full-length Cyrano RNA was introduced to MO treated zebrafish were the morphants able to be rescued. Based on these results, it was concluded that the conserved region is not sufficient for Cyrano function, but it is necessary and functions in conjunction with other regions of the lincRNA.

Since it is well known that lincRNA sequences exhibit poor conservation across divergent species [21], the conservation of the 300-nucleotide region of Cyrano, alongside the ability of human and mouse Cyrano spliced RNA to rescue morphant zebrafish phenotypes, was of considerable interest. I hypothesized that, although the 300-nucleotide RNA was not able to rescue the morphant zebrafish phenotypes, the 300-nucleotide conserved region of Cyrano, with the assistance of other local nucleotides, may fold into a functional, evolutionarily well-conserved secondary structure that is required for its function. Determination of structural motifs within the 300-nucleotide conserved region of Cyrano would provide an avenue for further understanding this RNA's role in embryogenesis.

Using thermodynamics, selective 2' hydroxyl acylation and primer extension (SHAPE), and comparative sequence analysis, I established the secondary structure of Cyrano in zebrafish, mouse (*Mus musculus,* ~8.2 kb*)*, and human (*Homo sapiens*, ~8.8 kb*)*. I discovered that the 300-nucleotide conserved region is only properly folded when 100 flanking nucleotides are added to each end of the 300-nucleotide conserved region. Inspection of the 300-nucleotide conserved region revealed multiple motifs conserved between the three species, supported by evolutionary

covariation of base pairs. In the next few chapters, I will describe the work supporting these conclusions.

Chapter 3: Thermodynamic folding of RNA: The Free Energy Method

As described by Tinoco and Bustamante [67], "an RNA molecule can be thought of as possessing a hierarchical structure in which the primary sequence determines the secondary structure, which, in turn, determines the tertiary folding". This simplification is possible because RNA contains only four building blocks: the purines, adenine and guanine; and the pyrimidines, cytosine and uracil. When these bases pair together, they form four basic secondary structural elements: helices, loops, bulges, and junctions. Because these structures are usually more stable than tertiary structures and their energetic quantities can be predicted, use of a theoretical folding algorithm to determine the secondary structure of the RNA is possible.

Given the high conservation of the 26-nucleotide core region in Cyrano across divergent species, I initially hypothesized that this conserved region may have a conserved secondary structure as well. However, upon folding this RNA sequence with the multiple fold (Mfold) webserver (a program that predicts secondary structure of single-stranded nucleic acids), very little structure was observed. As seen in Figure 6, the lowest energy structure (2 kcal/mol) is limited to a small 3 base-paired helix, with a 10 nucleotide apical loop.

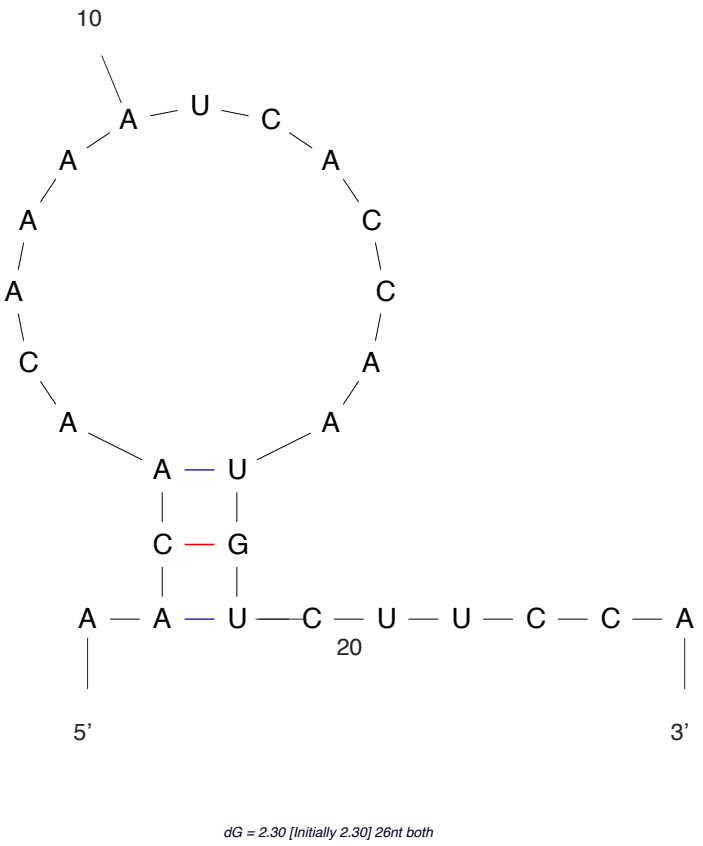


Figure 6: The predicted 26-nucleotide conserved core lacks secondary structure in the absence of neighboring nucleotides.

Thus, I hypothesized that other regions of the RNA were necessary for proper folding. In mice and humans, a region of 67 nucleotides is >98% conserved; this region encompasses the 26 nucleotide conserved region. Despite some evolutionary divergence, zebrafish has a high 82.4% pair wise identity with this region (Figure 7), a very high level of conservation over 200 million years of evolution [68].

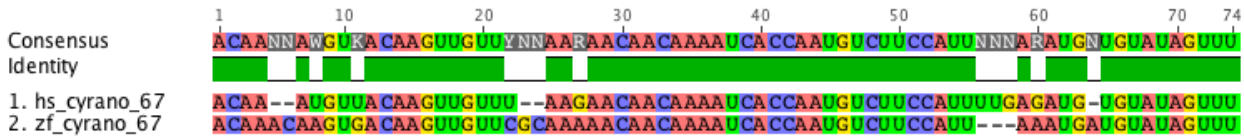


Figure 7: Sequence logo showing conservation between the 67-nucleotide conserved region between human and zebrafish RNA. Zebrafish RNA is 82.4% pairwise conserved with human.

The bases in the 67-nucleotide region corresponding to both human and zebrafish were added to the sequences previously folded, and the RNAs were refolded. As expected, the structure predicted for the 26-nucleotide conserved region was lost, and a new local conformation was observed. Three low energy structures were predicted for human Cyrano, and two were predicted for zebrafish. (Since mouse and human Cyrano lincRNAs share nearly identical sequences, only human Cyrano was folded.) Interestingly, in all models, a conserved 8 base paired helix, containing an apical loop of 4 and 6 nucleotides, for human and zebrafish respectively, was observed. This hairpin (H1_hs and H1_zf) is inclusive of the first 6 nucleotides of the core 26 nucleotide region. They pair with 6 nucleotides upstream of this region (Figure 8).

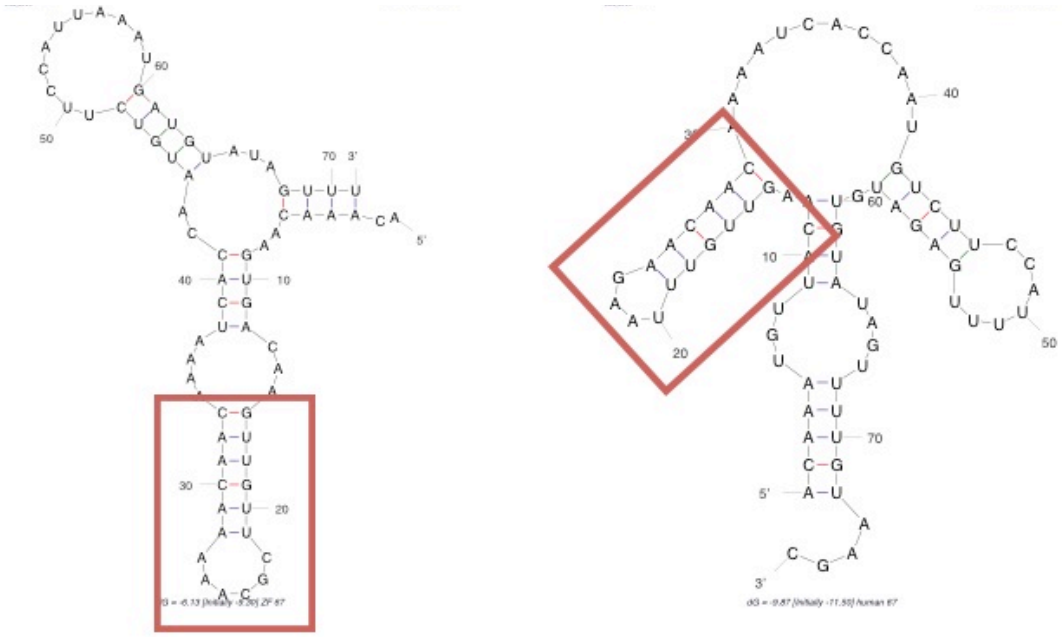


Figure 8: A six base pair helix, capped by an apical loop was observed in human, mouse, and zebrafish secondary structures encompassing the 67-nucleotide conserved region.

After inspecting and comparing the rest of the predicted structures, no other significant similarities were found between the RNAs. Because H1 was only present when the primary sequences of the RNAs were expanded, I further expanded the sequences of human, mouse, and zebrafish RNA to contain the full approximate 300-nucleotide conserved region of Cyrano RNA, as well as an approximate 500-nucleotide region (100 nucleotide flanking on each end of the 300-nucleotide region). These sequences were then folded and the results are listed and discussed in the next sections.

RNA structures of human 300-nucleotide (hs_300) and human 500-nucleotide (hs_500)

Folding of hs_300 and hs_500 yielded 15 and 8 predicted low energy structures, respectively. In the majority of structures (11/15 and 8/9 for hs_300 and hs_500, respectively) H1 was conserved. Careful inspection of the structures also revealed 6 other potential structural motifs, within the 300-nucleotide conserved region (Figure 9, Appendix Figure B1).

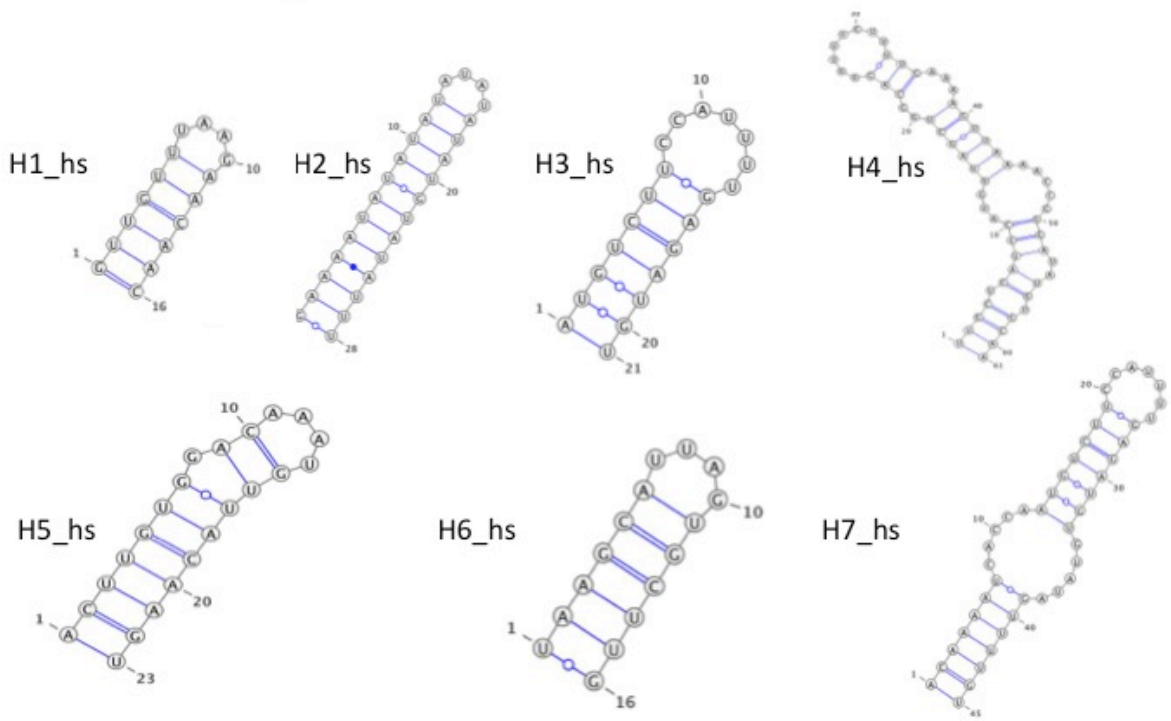


Figure 9: Folding of hs_300 and hs_500 revealed 7 hairpins common among the majority of the predicted structures.

The second motif is a hairpin (H2_hs) consisting of a 12 base pair helix and a 4 nucleotide apical loop. Another hairpin, H3_hs, is composed of a 7 base pair helix and a 7 nucleotide apical loop. H4_hs is more complex than the previously listed motifs; it contains three internal loops, an internal bulge, and an apical loop. H5_hs

contains a single nucleotide bulge and an apical loop. H6_hs is composed of a 6 base pair helix and a 4 nucleotide apical loop. H7_hs is a structure that appears in both hs_300 and hs_500, and overlaps with H3_hs, but also contains a secondary helical region and an internal loop.

RNA structures of mouse 300-nucleotide (mm_300) and mouse 500-nucleotide (mm_500)

Folding of mm_300 and mm_500 yielded 3 and 8 predicted low energy structures, respectively. H1 was present in all 11 structures generated in the 300 and 500 nucleotide predictions. Five other potential motifs were found within the 300-nucleotide region (Figure 10, Appendix Figure B2).

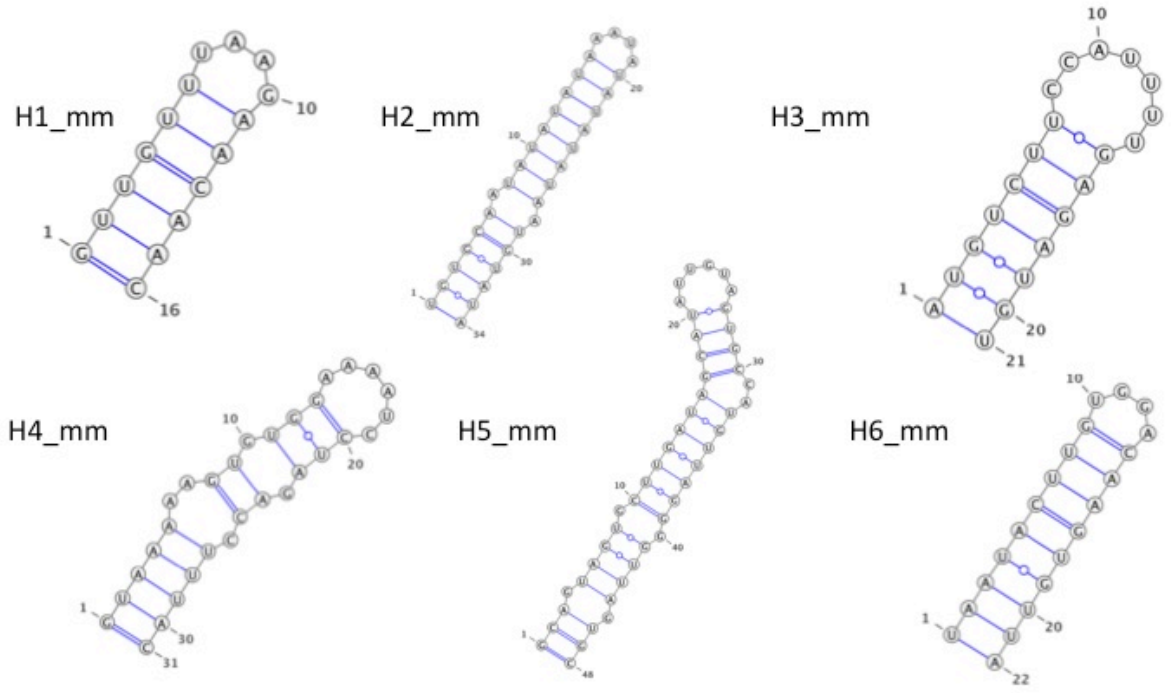


Figure 10: Folding of mm_300 and mm_500 revealed six common motifs present in the majority of structures.

The second (H2_mm), third (H3_mm) and sixth (H6_mm) motifs are all simple hairpins, with H2 containing a small, 2-nucleotide internal loop near the center of the helical domain. H4_mm and H5_mm are hairpins spotted with bulges and loops throughout the helical domain, and a 6-nucleotide apical loop at the end of the hairpin.

RNA structures of zebrafish 300-nucleotide (zf_300) and zebrafish 500-nucleotide (zf_500)

Folding of zf_300 and zf_500 yielded 17 and 23 predicted low energy structures, respectively. H1 was present in 15/17 structures from zf_300 and 23/23 times in zf_500. Five other potential motifs were found within the 300-nucleotide region (Figure 11, Appendix Figure B3).

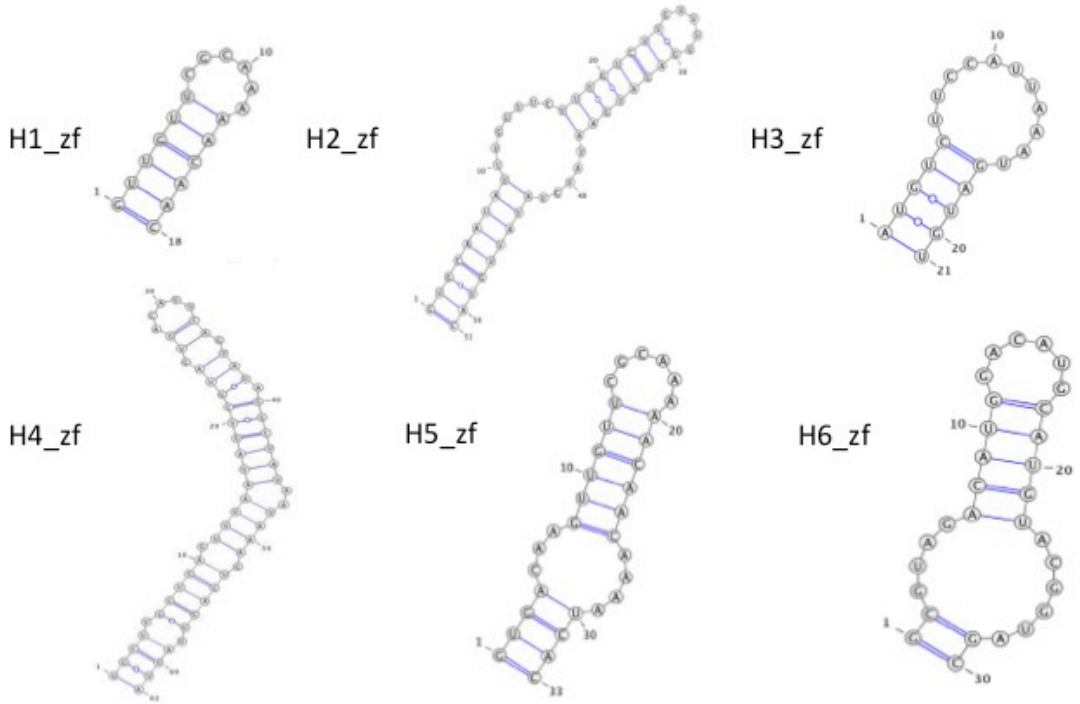


Figure 11: Folding of zf_300 and zf_500 revealed 6 hairpins common among the majority of the predicted structures.

The second (H2_zf), fifth (H5_zf) and sixth (H6_zf) motifs are hairpins containing an apical loop, and two helical regions separated by an internal loop. The third motif, H3_zf, contains a short, 5 base pair helix, with an 11 nucleotide apical loop. The fourth motif, H4_zf, is a lengthy helix, inclusive of numerous bulges and small, 2-nucleotide internal loops.

Predicted motifs shared in mouse, human, and zebrafish

Folding the 300 and 500 nucleotide fragments of Cyrano revealed that expanding the number of bases allowed for the formation of additional motifs not present based on the 300-nucleotide folding. There were three motifs common to all three species (Figure 12).

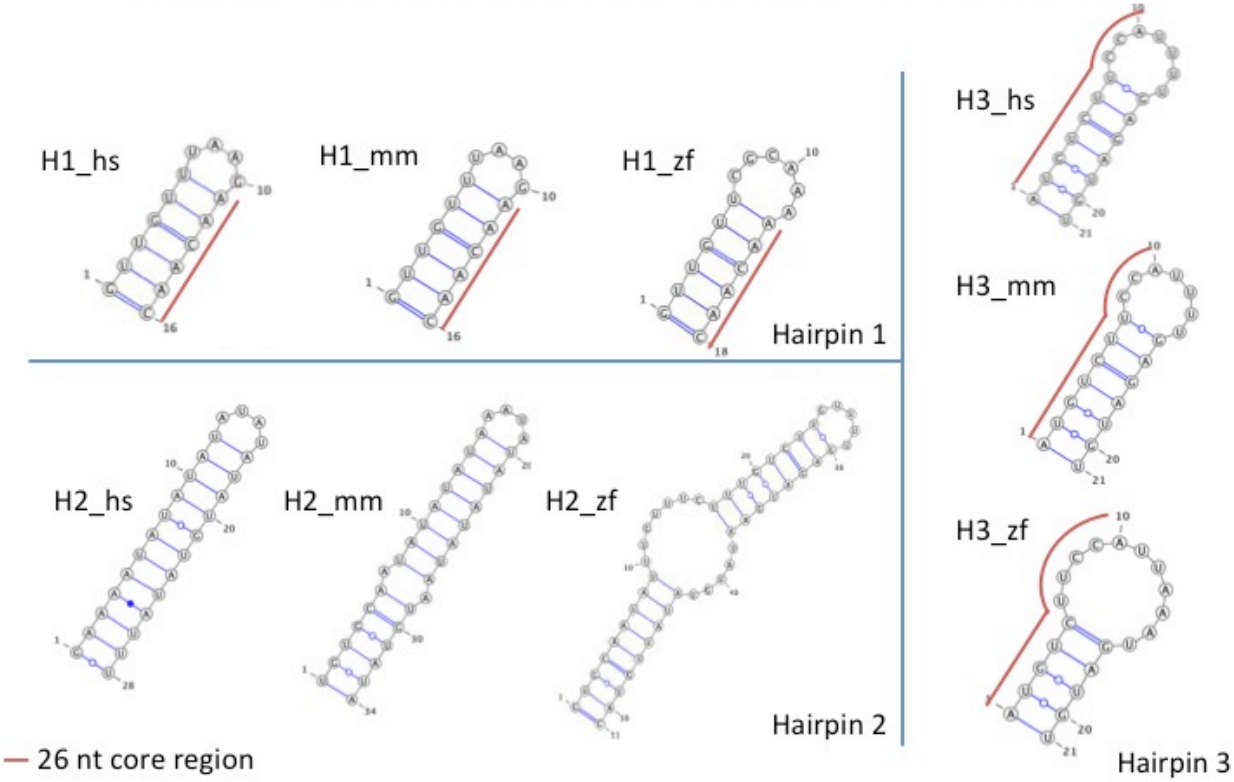


Figure 12: Three motifs, Hairpin 1, Hairpin 2 and Hairpin 3, were conserved in all three species. The red line indicates nucleotides that fall within the 26-nucleotide core conserved region.

I hypothesized these motifs may be present in the actual structure of each corresponding lincRNA because they were predicted independently and consistently for each species, despite slight divergence in the sequence alignment.

Folding of full-length hs_cyrano

Because expansion of the primary sequences from 67 nucleotides to 500 nucleotides revealed additional potential secondary structure motifs in zebrafish, human and mouse, I hypothesized that folding the full-length RNA may lead to the identification of other significant structural features. While it is well known that thermodynamic folding can provide useful information about the architecture of RNAs, unfortunately it is only able to accurately predict more than 70% of an RNA's secondary structure correctly [69]. This is because, as the number of nucleotides increases, so does the number of possible secondary structures; for an RNA of nucleotide length *n*, there are 1.8^n possible secondary conformations [70]. Additionally, folding algorithms have run times that grow as the cube of the sequence length, yielding a search that is very computationally expensive [71].

Despite the impracticality of folding a 4.5 kb long lincRNA transcript, the results revealed that human Cyrano possessed a dumbbell-like structure containing two large 2 kb domains separated by an approximate 250 nucleotide base paired helical region (Figure 13).

This 200 base paired helical region was an interesting robust result. A BLAST search of this region in the NCBI database revealed an alignment with the SVA (SINE (short interspersed elements) VNTR (variable number tandem repeat) Alu) retrotransposon element. A retrotransposon is a genetic element that mobilizes as an RNA intermediate, which is subsequently reverse transcribed to a cDNA copy via the Target Primed Reverse Transcription (TPRT) mechanism [72, 73].

This allows for the spread of retrotransposons throughout the genome. Mobile genetic elements account for approximately 50% of the genome. SVA is a nonautonomous composite repetitive element that requires the enzymatic machinery elements for retrotransposition [74]. While its function is unknown, it is believed that SVA may function in the origins of replication, play a role as a chromosomal band-aid, or be a mediator of translational activation [73, 75]. The 250 base paired region found within human Cyrano is a portion of the RNA intermediate of the SVA retrotransposon, but further evaluation is necessary to understand its role in Cyrano's function. Zebrafish and mouse lincRNA predicted structures did not contain a long helical region separating two domains, as inspected in human Cyrano. I hypothesize that the absence of this retrotransposon in zebrafish is due to an evolutionary time scale difference; SVAs have only existed for an estimated 11 million years [74] whereas zebrafish have existed for approximately 13 million years. Additionally the alignment provided by Ulitsky et al. [53] reflected a lack of conservation with this region compared to zebrafish. Further studies are underway to determine if, and where, SVA resides in mouse Cyrano lincRNA.

Figure 13: Folding of the 4.5 kb human Cyrano lincRNA reveals two 2 kb domains separated by a 250 base paired helical region (predicted by RNAstructure folding algorithm).

Establishment of the 250 base paired domain allowed me to employ the ‘sliding window’ approach for future folding predictions. With this strategy, a sequence is split into multiple, overlapping, fixed segments. Each segment is then processed separately to reduce the computational cost. A drawback of this method is that one risks arbitrarily truncating the RNA such that it folds improperly. However, knowing, instead of guessing, the boundaries for splitting the RNA (by means of the apparent 250 base paired helical region), allowed for human Cyrano lincRNA to be split into 2 smaller, 2 kb segments to reduce the folding time without consequence (folding of

the 4.5 kb human Cyrano took 17 hr). This was especially useful when performing chemical probing experiments, as described in Chapter 5.

To further validate the existence of the three RNA motifs common between human, mouse, and zebrafish determined by free energy minimization (H1, H2, and H3), I utilized comparative sequence analysis, as described in the next chapter.

Chapter 4: Comparative Sequence Analysis

The ability of an RNA to maintain its structure and function, despite primary sequence evolution, is due to positional covariance. Covariance occurs when base pairs vary at the same time during evolution, allowing regions of an RNA that are structurally critical to maintain structure, despite primary sequence divergence [76]. One example of this is tRNA, an RNA highly conserved in both prokaryotes and eukaryotes, which decodes a mRNA sequence into a protein, a critical step in translation [77]. tRNA is able to maintain its clover-like structure through co-varying base pairs across divergent species (Figure 14).

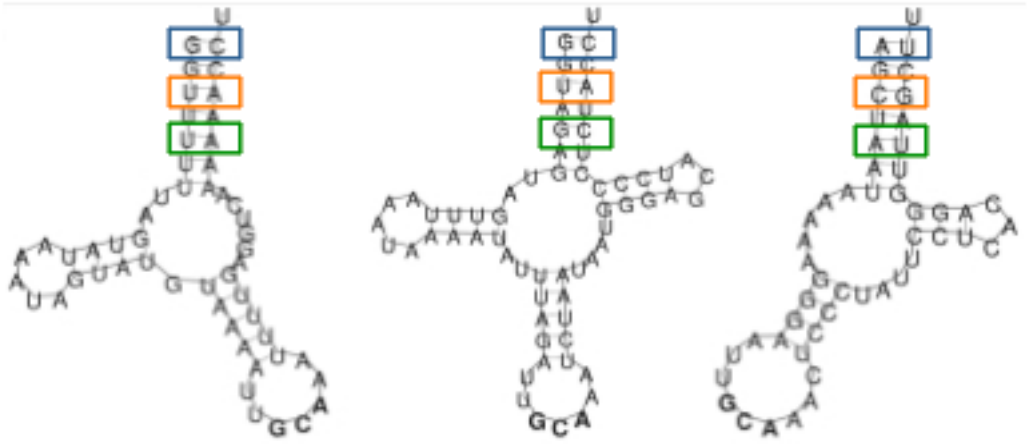


Figure 14: tRNA is conserved in function because covariation allows for conservation of the secondary structure. (left: Tigriopus japonicus, middle: Daphnia pulex, right: Lepeophtheirus salmonis). The blue, orange and green boxes show areas of co-varying base pairs [78].

Because morphant zebrafish embryos were able to be rescued with co-injection of spliced Cyrano RNA from not only zebrafish (60%) but also mouse (60%), and human (35%), I hypothesized that structural conservation, via covariation, may exist within the 300-nucleotide conserved region of Cyrano. I used CMfinder (Covariance

Model Finder) and WAR (Webserver for Aligning Structural RNAs) as comparative sequence analysis tools for identifying motifs and, predicting the secondary structure of the 300-nucleotide conserved region of Cyrano lincRNA. Secondary structure prediction by computational algorithms is not always perfect [79], so I inspected, and when necessary, adjusted the alignments to further optimize the predicted structures.

CMfinder is one of many comparative sequence analysis tools used for the prediction of secondary RNA structures, exploiting evolutionary conservation of RNAs through covariation, thermodynamic stability, and phylogeny [80]. This program uses unaligned sequences to predict RNA motifs (Figure 15).

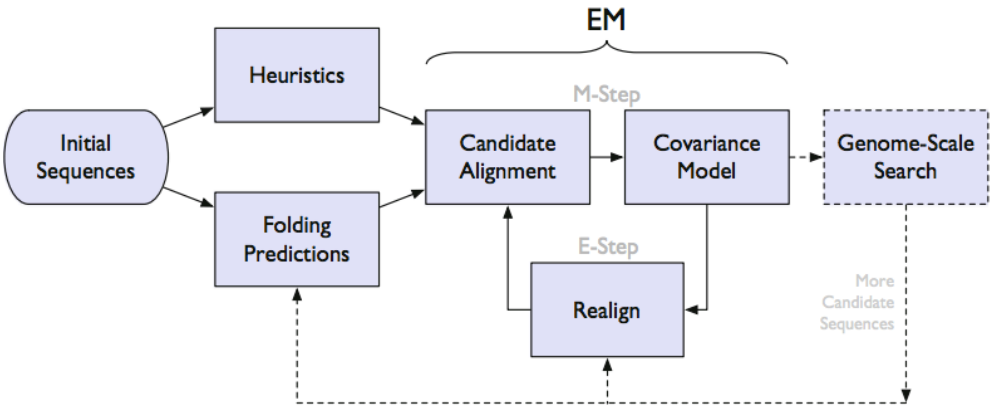


Figure 15: Basic steps CMfinder uses for generating a consensus secondary structure [71].

The unaligned sequences of interest are grouped together and a candidate alignment is generated via both sequence conservation and single-sequence structure prediction. From this alignment, both a consensus structure and a covariance model are built. The covariance model is then refined by an Expectation-Maximization-like

iteration, where input sequences are aligned to the covariance model, and the final model is rebuilt from the refined alignment [71].

CMfinder is robust in that it allows a variety of features to be modified to guide and optimize motif predictions. The allowances include the number of stem-loops, the number of motifs, the minimum and maximum length of a motif, the number of candidates, the fraction of sequences containing the motif, and the option to use BLAST for identification of candidates. Multiple rounds of structural predictions are run, varying each of those options, to generate potential RNA motifs.

WAR is another comparative sequence analysis tool that simultaneously uses a number of comparative sequence analysis programs to perform multiple alignment and secondary structure predictions for an RNA (Table 1) [81].

Table 1: A list of the comparative sequence analysis programs simultaneously run by WAR.

CMfinder
FoldalignM
LaRA
MASTR
RNAalifod + ClustalW
RNAforester _ RNAcast
RNASampler

The benefit of using an ensemble of methods is that the final output is a combined view of the predictions. This makes it easier to identify which predicted structures, or

components of a structure, are most correct. Each program represented by the WAR webserver utilizes a different algorithm. While it may be difficult to choose which structure is correct from independently run programs, it is easier to identify correctly predicted structures from a combined consensus prediction.

Methods:

CMfinder is available for webserver use. However because the webserver has RNA size limitations (500 base pairs), the software was downloaded and installed from http://bio.cs.washington.edu/CMfinderWeb/CMfinderInput.pl. The unaligned sequences evaluated by CMfinder were gathered from NCBI's Basic Local Alignment Search Tool (BLAST) [82]. An approximate 500 nucleotide region of human Cyrano (~300-nucleotide conserved region with 100 flanking nucleotides on both the 5' and 3' ends of the RNA) was BLASTed, yielding 51 sequences from a variety of species as listed in Appendix A (Table A1). Because covariation between aligned positions is a crucial requirement for the prediction of RNA structure, multiple fasta files with varying sequence identities were created from the original 51 sequences returned by the BLAST search (ranging from 70% to 95% conservation). Alignment methods generally fail when sequence conservation falls below 60% [83], while sequences that are too closely related, yielding high percentages of conservation, lack sufficient covariation necessary for determination of secondary structures.

The sequences were then formatted to a .fasta file and run by the cmfinder.pl perl script, initially under default conditions (Table 2).

Table 2: Parameters set for initially identifying motifs in CMfinder; these are default values.

Number of stem loops	1
Number of motifs	3
Minimum length of a motif	30
Maximum length of a motif	100
Number of candidates	40
Fraction of sequences containing the motif	0.8

An output motif and its corresponding covariance model were generated. Alignment options listed in Table 2 were further adjusted over multiple rounds of modeling to identify motifs.

Of the motifs predicted by CMfinder (Figure 16), two motifs were consistently predicted despite the .fasta input file, and despite the adjustment of run parameters. These motifs correspond to the H1 and H3 (predicted by free energy minimization, Chapter 3) for all three species.

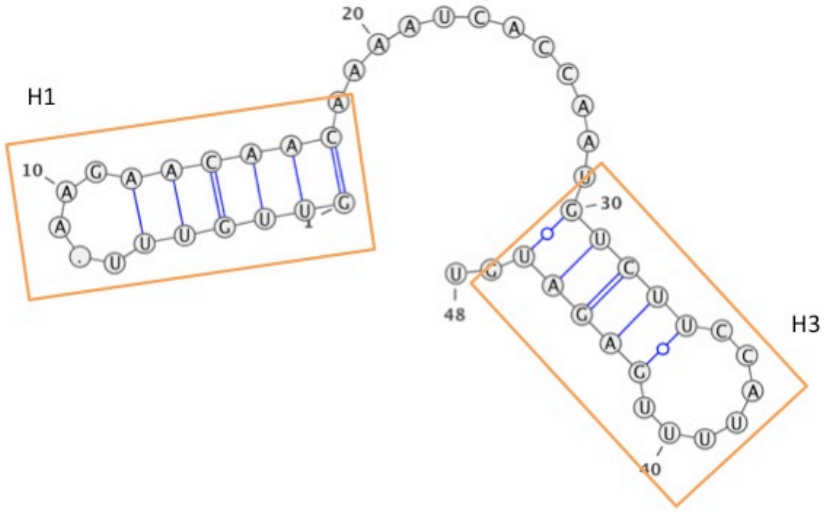


Figure 16: Two motifs predicted by comparative sequence analysis were consistently generated and also matched motifs predicted by free energy minimization.

The same .fasta files run with CMfinder were input into the WAR webserver, and local alignment was selected (this was because of sequence length limitations set by some of the programs, which rely on global alignments). The consensus structure generated (Figure 17) revealed four motifs, two of which were consistent with those predicted by CMfinder (orange box). Another motif was consistent with motif H2 (predicted by free energy minimization) for all three species (blue box in Figure 17).

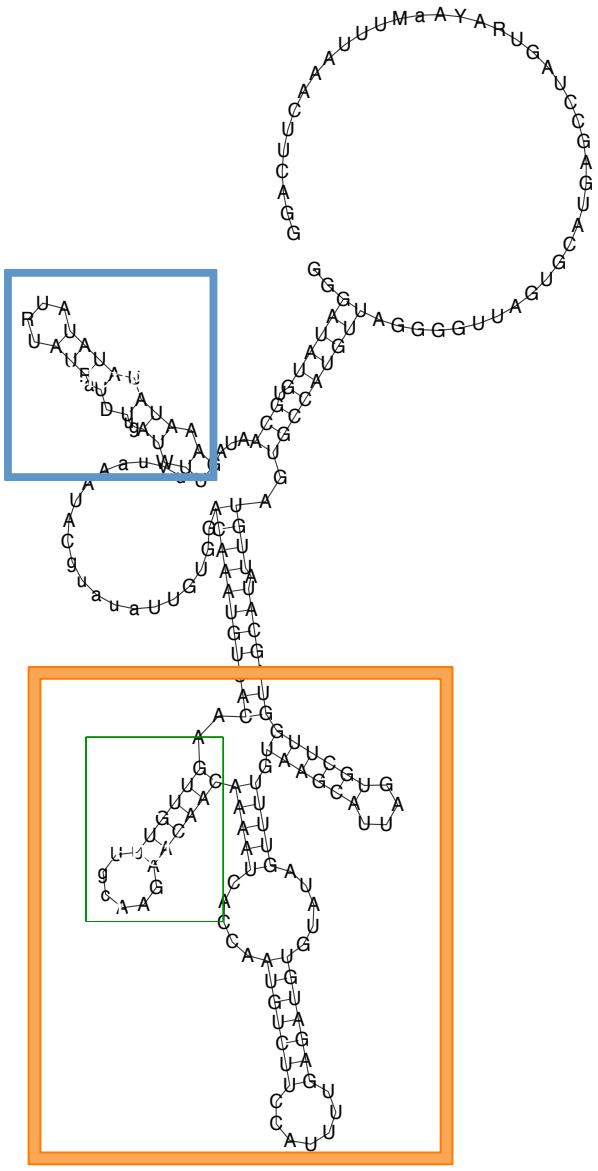


Figure 17: The consensus structure predicted by CMfinder contains four motifs. The motif outlined in blue corresponds to H2, and the motif in green corresponds to H1. In orange, two other motifs are predicted, which are inclusive of the 26-nucleotide core region.

WAR also revealed one new motif not observed in the free energy predicted structures, as seen in the orange box on the far right, in Figure 17. This motif causes the 26-nucleotide conserved core to take on a Y-like structure.

Inspection of the alignment files generated by both CMfinder and WAR, alongside toggling of the structures also revealed another potential motif, maintained by covariation (Figure 18).

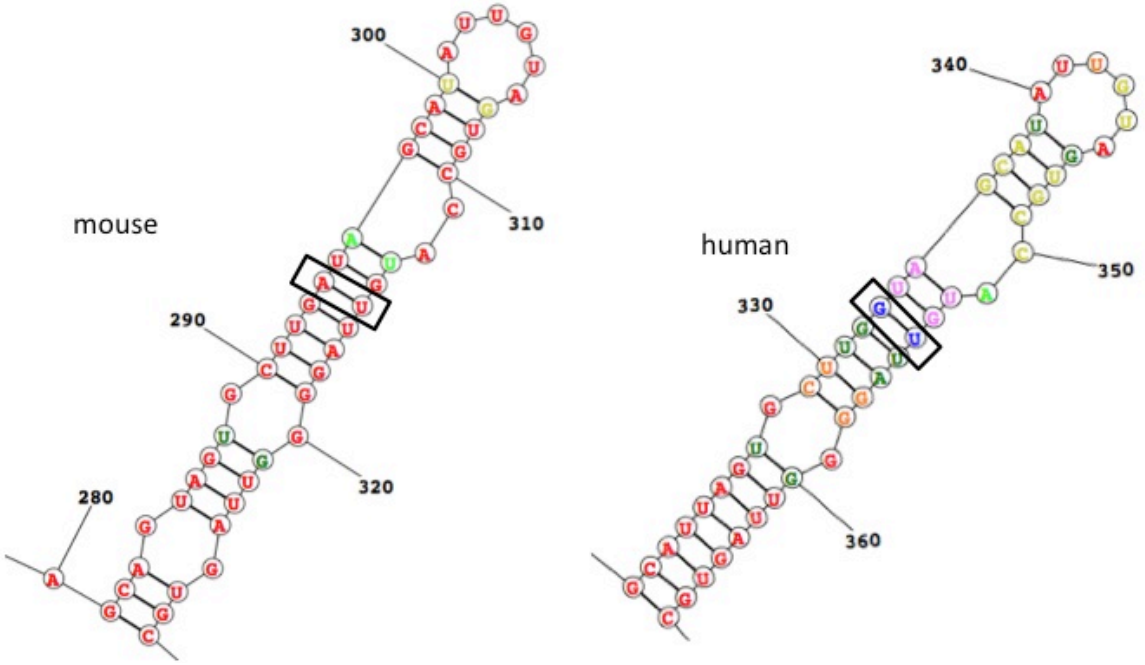


Figure 18: Inspection of alignment led to the observation of co-variance in this helical domain of human and mouse Cyrano. Boxed in black is a base pair maintained by positional co-variance.

Despite the apparent areas of covariation in the structures, and consistency with free energy minimization predictions, experimental validation was pursued. I checked the validity of these predicted motifs and structures with selective 2'-hydroxyl acylation and primer extension (SHAPE) as described in the next chapter.

Chapter 5: Selective 2'-Hydroxyl Acylation and Primer Extension Analysis

Free energy minimization revealed a long 250 nucleotide helical region separating two ~2 kb domains in human Cyrano, and three common hairpins, H1, H2, and H3 in mouse, human, and zebrafish. Comparative sequence analysis revealed four motifs conserved due to covariation. I checked the validity of these predicted motifs and structures with selective 2'-hydroxyl acylation and primer extension (SHAPE).

SHAPE analysis is a cDNA fragment analysis-based method utilized for determining the secondary structure of RNA [84-91]. The 2'-hydroxyl group of the RNA ribose is subject to nucleophilic attack by an electrophilic SHAPE reagent (Figure 19, 20) (1M7, 1-methyl-7-nitroisatoic anhydride, or NMIA, N-methylisatoic anhydride, dissolved in DMSO in varying concentrations) to form a 2'-*O*-adduct [92].

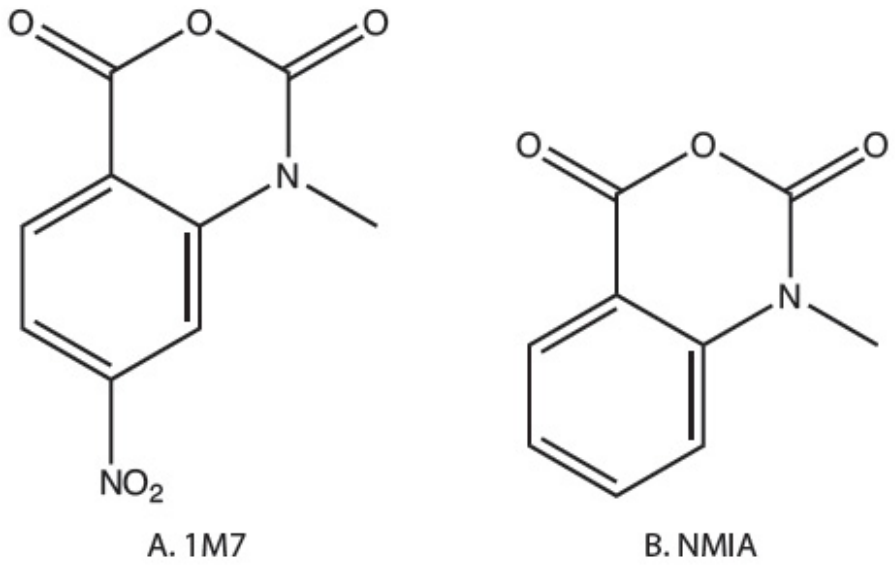


Figure 19: Two SHAPE reagents, a) 1M7 and b) NMIA which form an adduct with non-base paired RNA nucleotides.

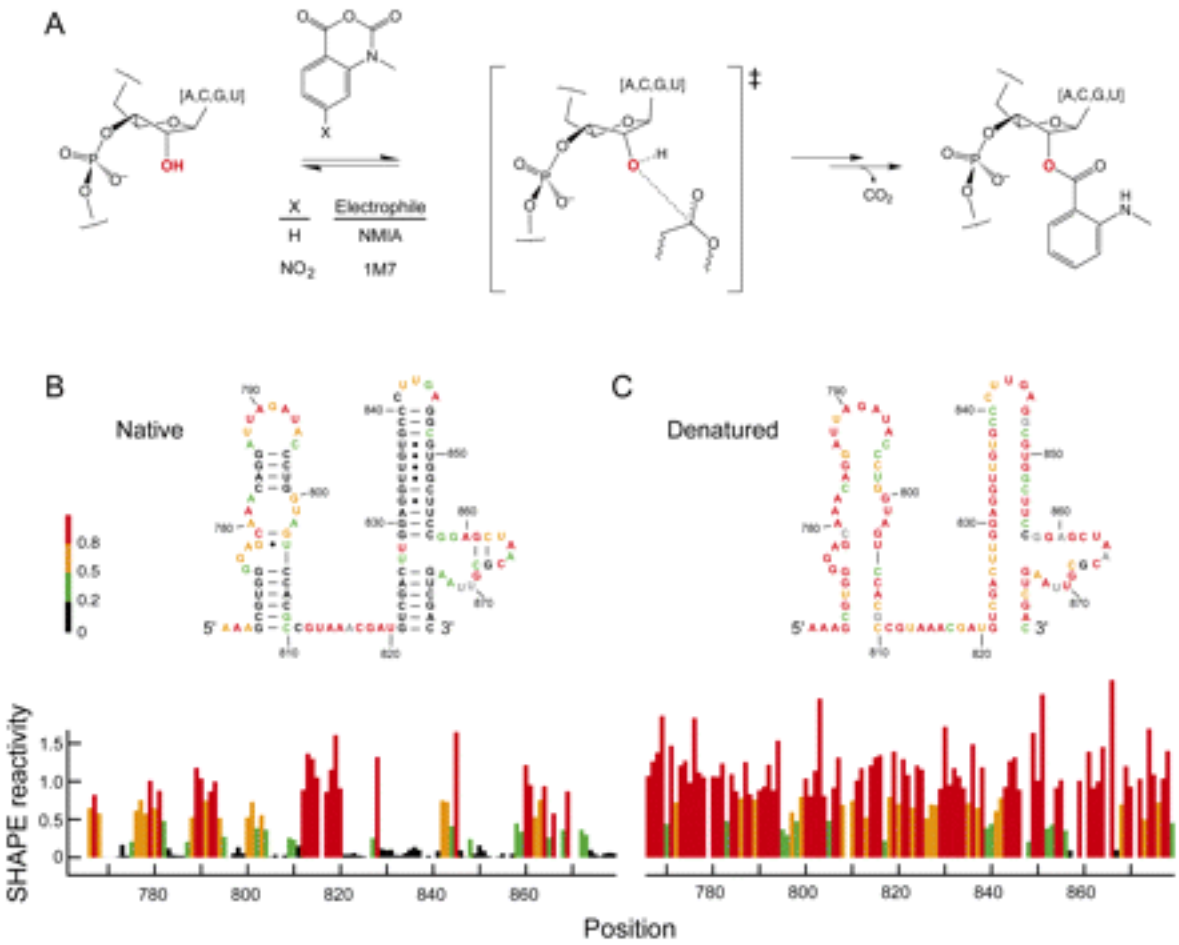


Figure 20: SHAPE reaction. A) NMIA or 1M7 is reacted with RNA. B) SHAPE reactivity is mapped to the primary sequence to generate a secondary structure for native and C) denatured RNA [92].

SHAPE reagents more readily react with flexible nucleotides, such as those in bulges or loops [88], because they are less conformationally constrained and thus more conducive to nucleophilic attack. More structured regions are not as reactive with SHAPE reagents because their geometry is constrained.

Once RNA has been modified with a SHAPE reagent, it is then used as an RNA template for a reverse transcription reaction. Modified nucleotides are identified by primer extension with fluorescently or radioactively labeled primers. Because the reverse transcriptase enzyme is not able to proceed with elongation once it

encounters a base modified with a 2'-*O*-adduct, a cDNA fragment corresponding to a length equivalent to the primer and the number of nucleotides to the modified base is formed. Four sequencing reverse transcription reactions containing unmodified RNA, with a 1:1 ratio of a dideoxy-nucleotides (ddA, ddT, ddG, or ddC) and deoxy-nucleotides, allow for the identification of each SHAPE reagent dependent cDNA fragment.

A control containing DMSO-treated RNA is reverse transcribed alongside the modified RNA and sequencing reactions. DMSO is used as a control because NMIA is dissolved in DMSO due to its lack of stability in water. Because the reverse transcription reaction is imperfect, a natural level of cDNA aborts occur at every nucleotide. Bases labeled with NMIA cause a higher level of aborts for non-structured bases. If fluorescently labeled primers are used, capillary electrophoresis is performed on the cDNA fragments after reverse transcription, and the areas under the DMSO and NMIA curves are compared to determine the reactivity of each base. If radioactively labeled primers are used, the cDNA fragments are run on a gel after the reverse transcription, and the intensities of the DMSO and NMIA bands are compared to determine the reactivity of each base.

Part 1: SHAPE analysis of full-length human Cyrano

I used SHAPE analysis with NMIA and $\gamma^{32}P$ labeled primers to determine the secondary structure of the full-length Cyrano RNA in human. First, I validated the 250 bp SVA retrotransposon element (Figure 21).

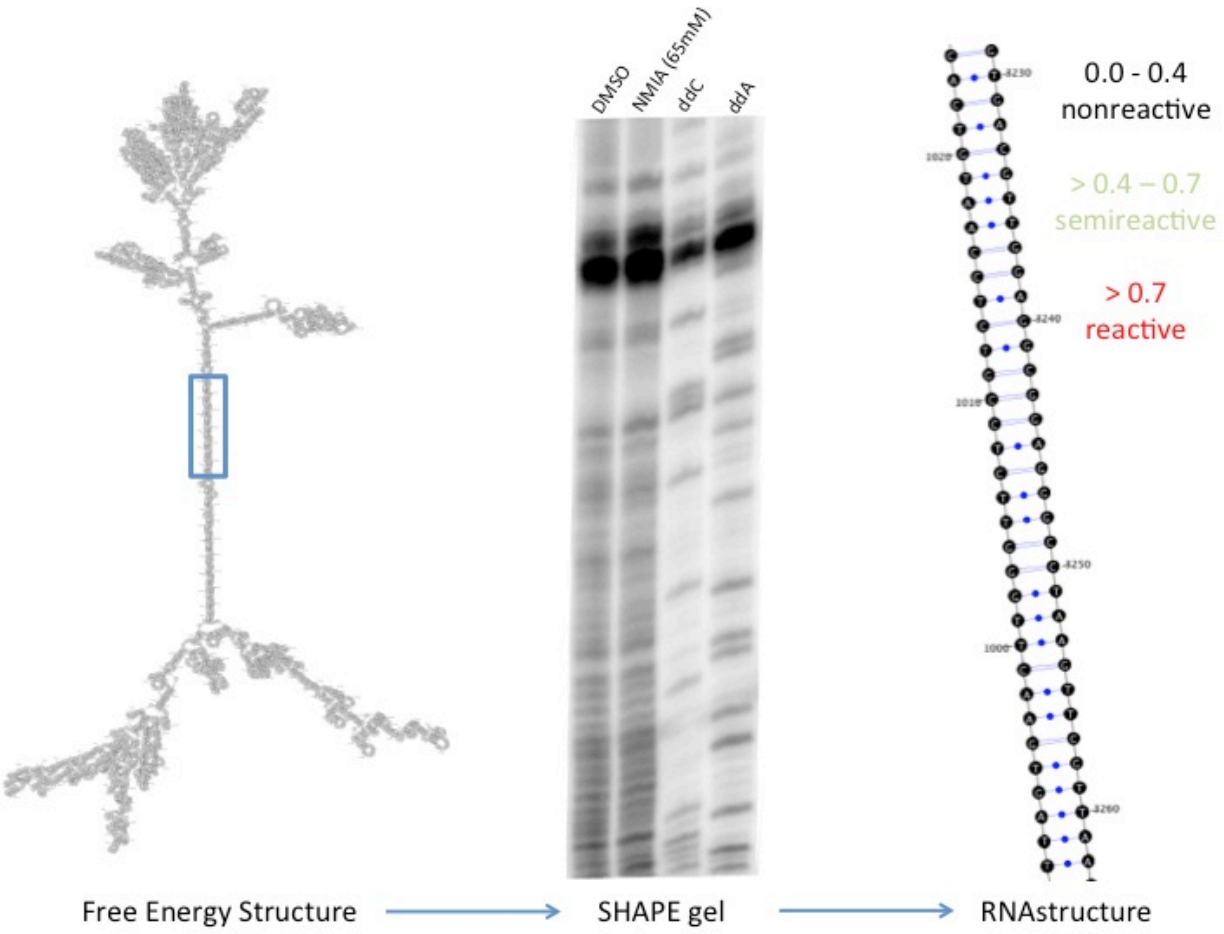


Figure 21: The 250 nucleotide helical region was confirmed by SHAPE analysis. The SHAPE gel (left to right: DMSO treated RNA, 65 mM NMIA treated RNA, ddC, and ddA sequencing) corresponds to the region enclosed in the blue box. The reactivity is mapped on the secondary structure (right).

After validation of the helical domain, the lincRNA sequence was split in half to contain only the 2256 nucleotide end of the dumbbell structure that contained the 300-nucleotide conserved region. This was done to reduce the folding space and time of the folding. The structure was refolded inclusive of the SHAPE reactivity, and the 300-nucleotide conserved region was inspected. The predicted structures, guided by SHAPE reactivity, suggested that the 300-nucleotide region folded locally in some regions, but that it base paired with nucleotides many hundreds of positions away in others (Figure 22).

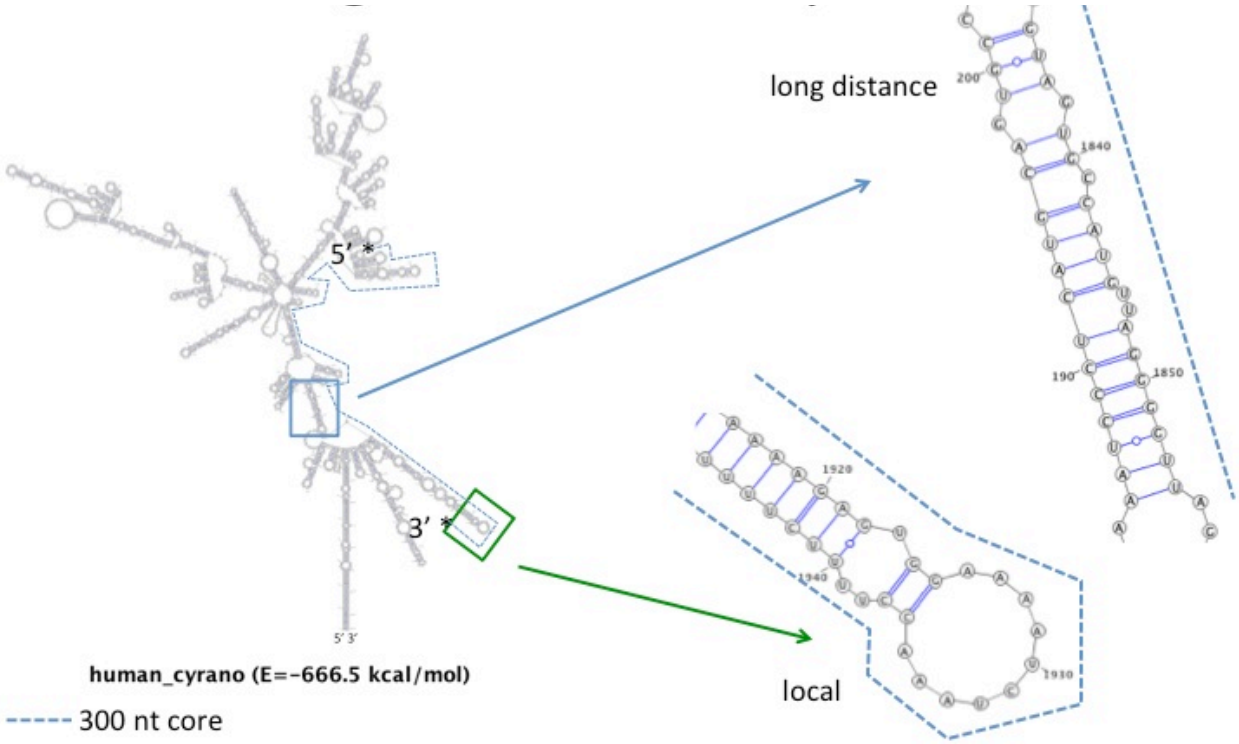


Figure 22: The conserved region spans nucleotides 1600 to 1900 (dashed blue line). The blue box corresponds to a region where the conserved region (nucleotides 1830 to 1860) base pairs with nucleotides hundreds of nucleotides away, while the green box corresponds to an area where the 300-nucleotide conserved region folds locally.

One of the drawbacks of SHAPE is that while it provides information about the structural characteristic of each base, it fails to provide information about exact base pairing. Additionally, folding 2256 bases, even with SHAPE data, leaves a lot of room for incorrectly predicted structures [93]. I hypothesized that it was possible the conserved region did not base pair with such long-range contacts at all. After looking at an alignment of 47 species, no conservation was noted in the region base paired with the corresponding conserved region (as seen in the blue box in Figure 22). With this information, I experimentally isolated the 300-nucleotide conserved region (with 100 flanking nucleotides on each end) and performed SHAPE on the fragmented 500 nucleotide long lincRNA with capillary electrophoresis.

Part 2: SHAPE analysis of 500 nucleotide fragment of zebrafish, mouse and human Cyrano

After performing SHAPE on the 500 nucleotide fragment of human Cyrano lincRNA, I compared the SHAPE reactivity of each base with the full-length human Cyrano SHAPE reactivities. This is known as the 'shotgun' method [34]. Capillary electrophoresis results revealed that the reactivity of the bases in the full-length Cyrano were very similar to the reactivities of the bases in the 500 nucleotide region, suggesting that the structural information posed by the fragmented RNA was correct (Figure 23).

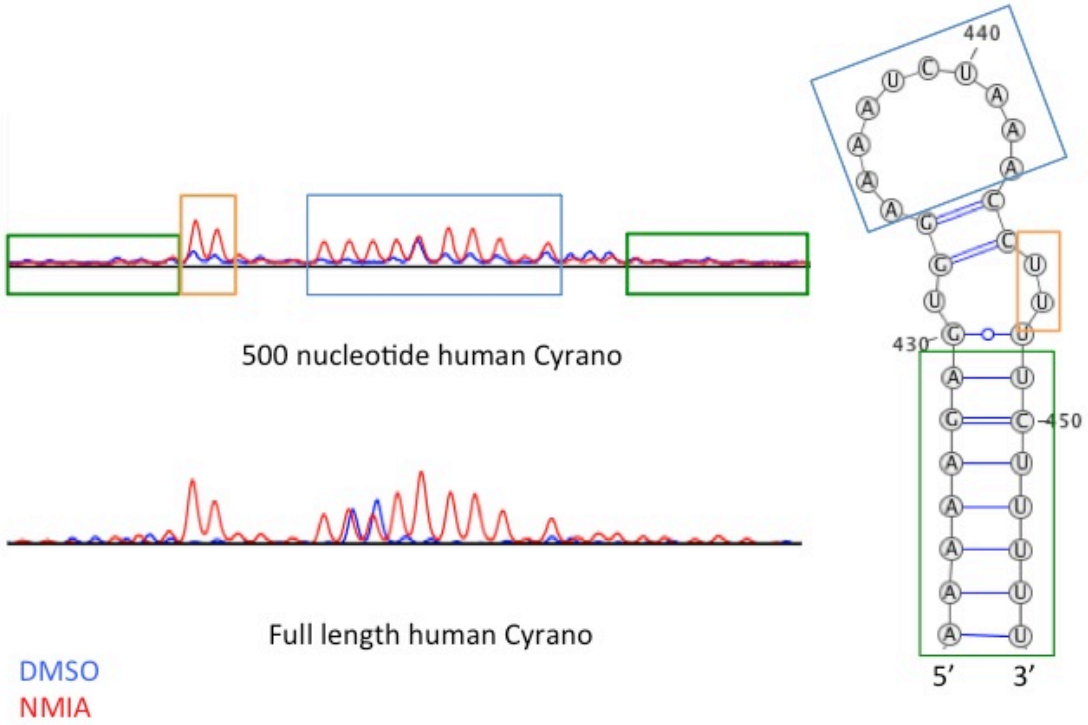


Figure 23: The full-length human lincRNA and the 500 nucleotide human lincRNA had consistent SHAPE reactivity throughout the 300-nucleotide conserved region. The green, orange and blue boxes on the gel images correspond to the helix, the apical loop, and the 2-nucleotide bulge, respectively.

Because the SHAPE data between the full-length and 500 nucleotide long RNAs were consistent, I concluded the 300-nucleotide region folded locally, instead of interacting with bases multiple hundreds of nucleotides away. However, SHAPE results for zebrafish, mouse, and human Cyrano lincRNA revealed that the 300-nucleotide conserved region only folds properly in the presence of the 100 flanking nucleotides on each end of the 300-nucleotide conserved region.

Folding each of the three RNAs revealed intricate and modular structures for each of the species (Figures 24, 25, and 26). The secondary structures for each of the three species appears to have a shamrock-like shape; a long helical domain represents the stem of the shamrock, and hairpins strung out in a chain-like fashion represents the fan. Interestingly, the 300-nucleotide conserved region falls within the fan-like region in each of the three structures.

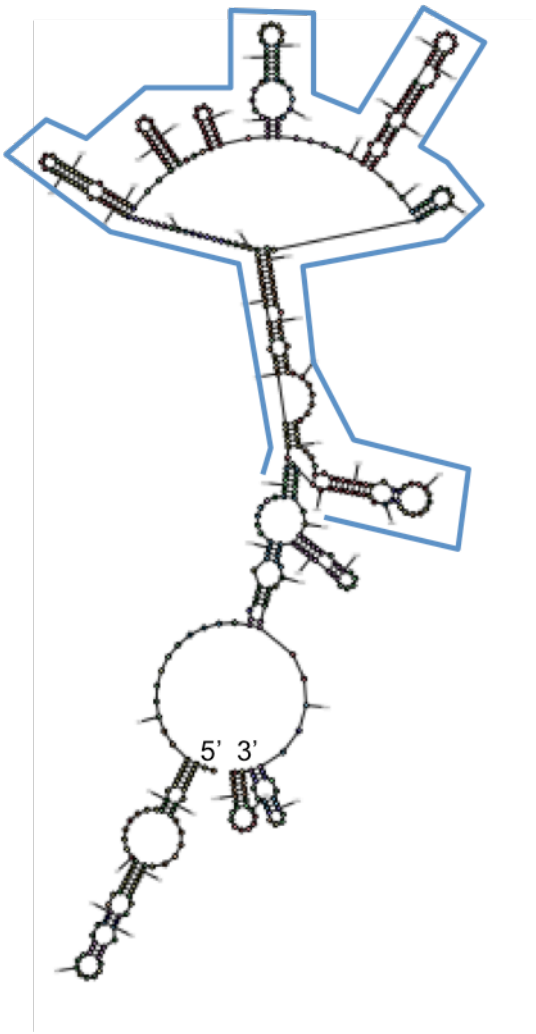


Figure 24: SHAPE predicted structure of mouse Cyrano lincRNA. The blue outline corresponds to the 300-nucleotide conserved region.

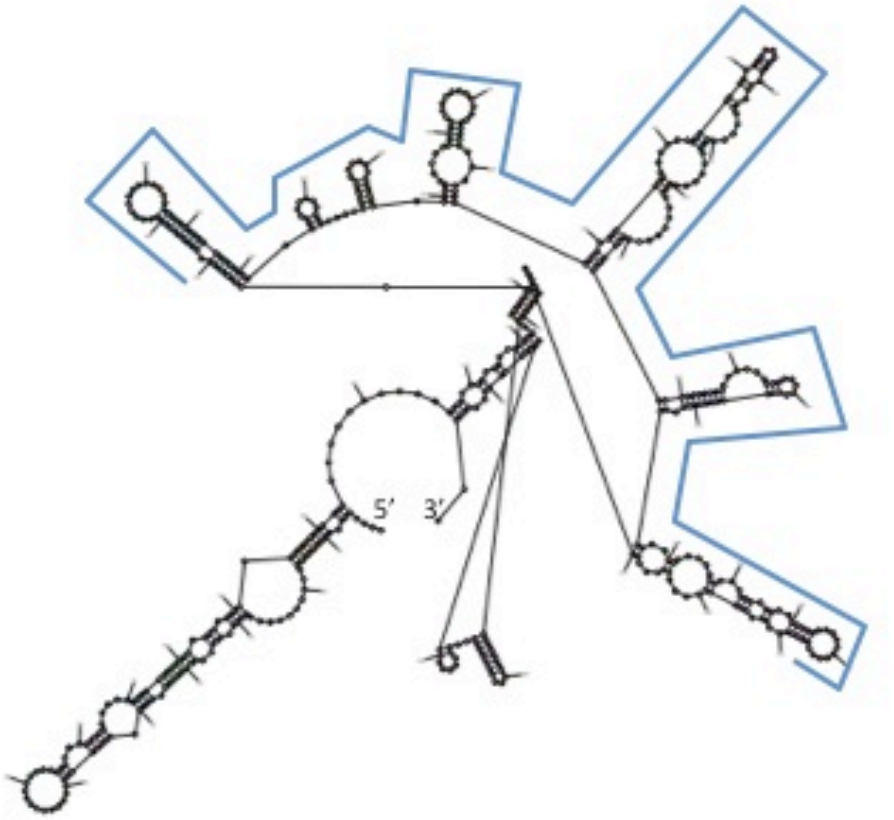


Figure 25: SHAPE predicted structure of zebrafish Cyrano lincRNA. The blue outline corresponds to the 300-nucleotide conserved region.

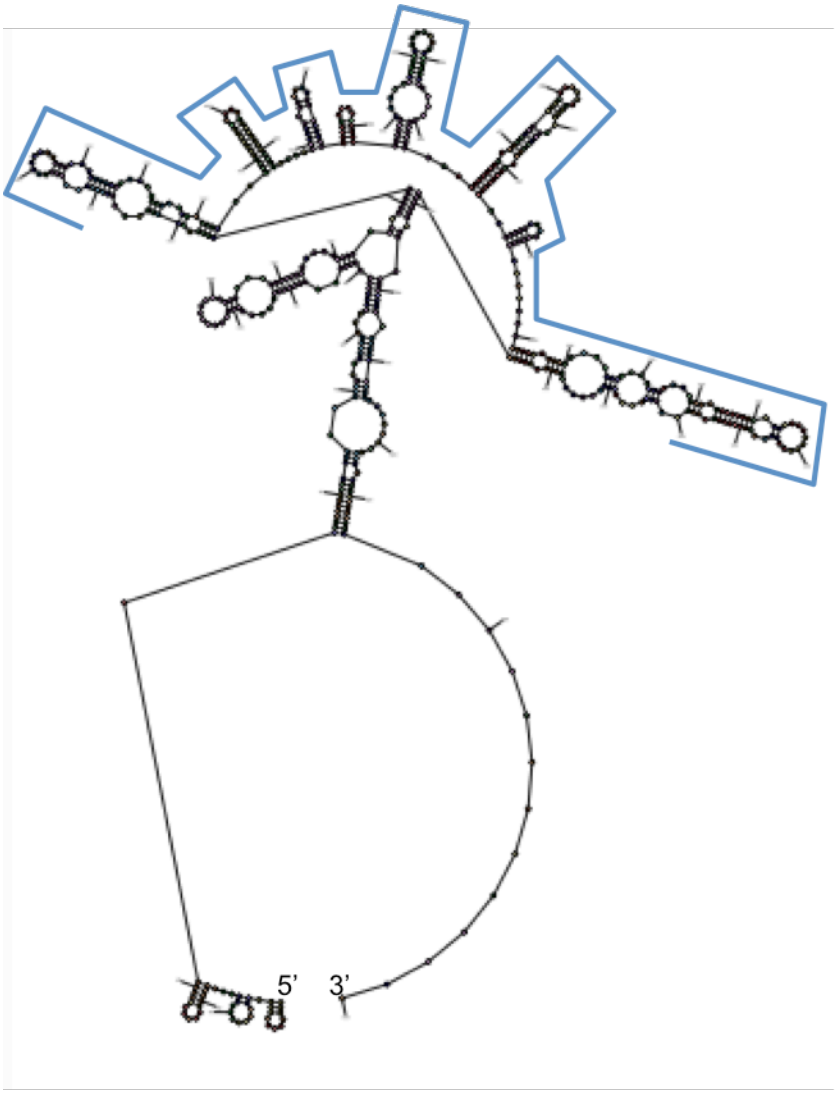


Figure 26: SHAPE predicted structure of human Cyrano lincRNA. The blue outline corresponds to the 300-nucleotide conserved region.

Comparing calculated structures of zebrafish, mouse and human revealed four major domains that were present in all three of the structures (Figure 27).

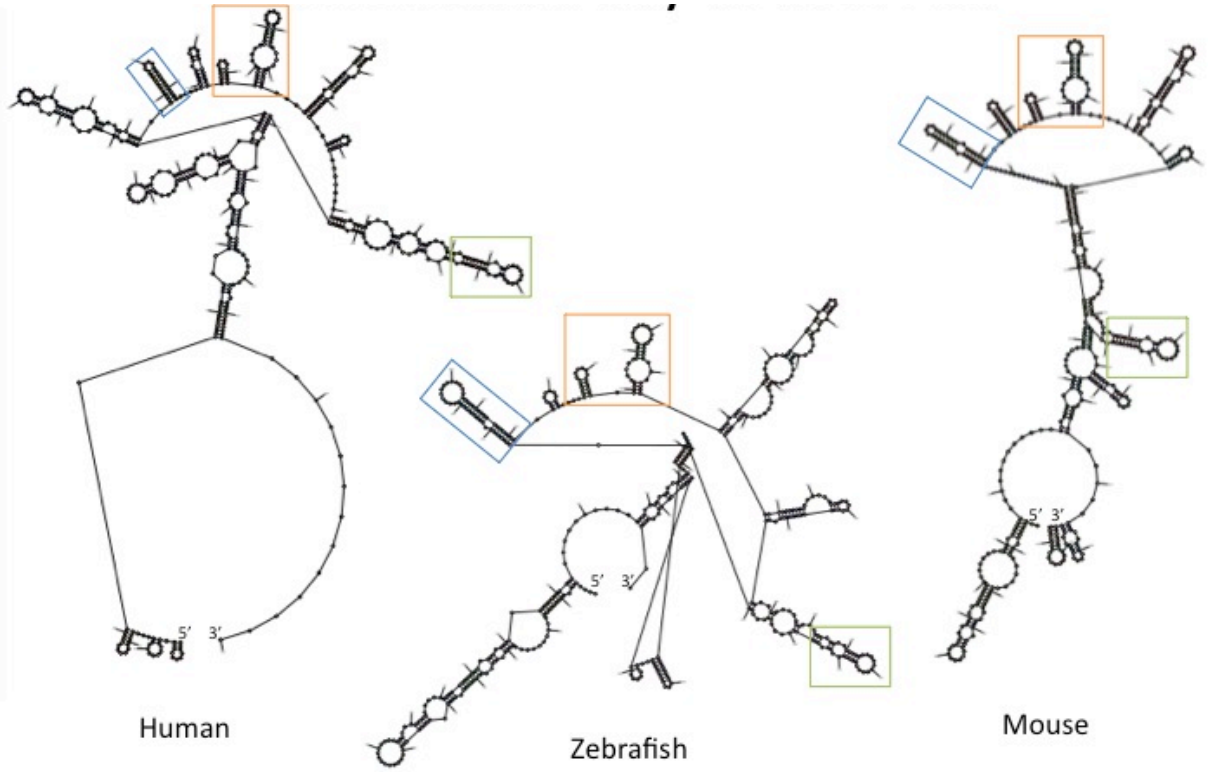


Figure 27: Three domains, highlighted in blue, orange and green, are conserved in each of the three species.

The first domain (Figure 28), is a hairpin composed of an AU-rich repeat region. An insertion of six nucleotides in the zebrafish RNA sequence offers an explanation for the difference in secondary structure relative to human and mouse. This domain corresponds to H2, the hairpin predicted by free energy minimization (Chapter 3).

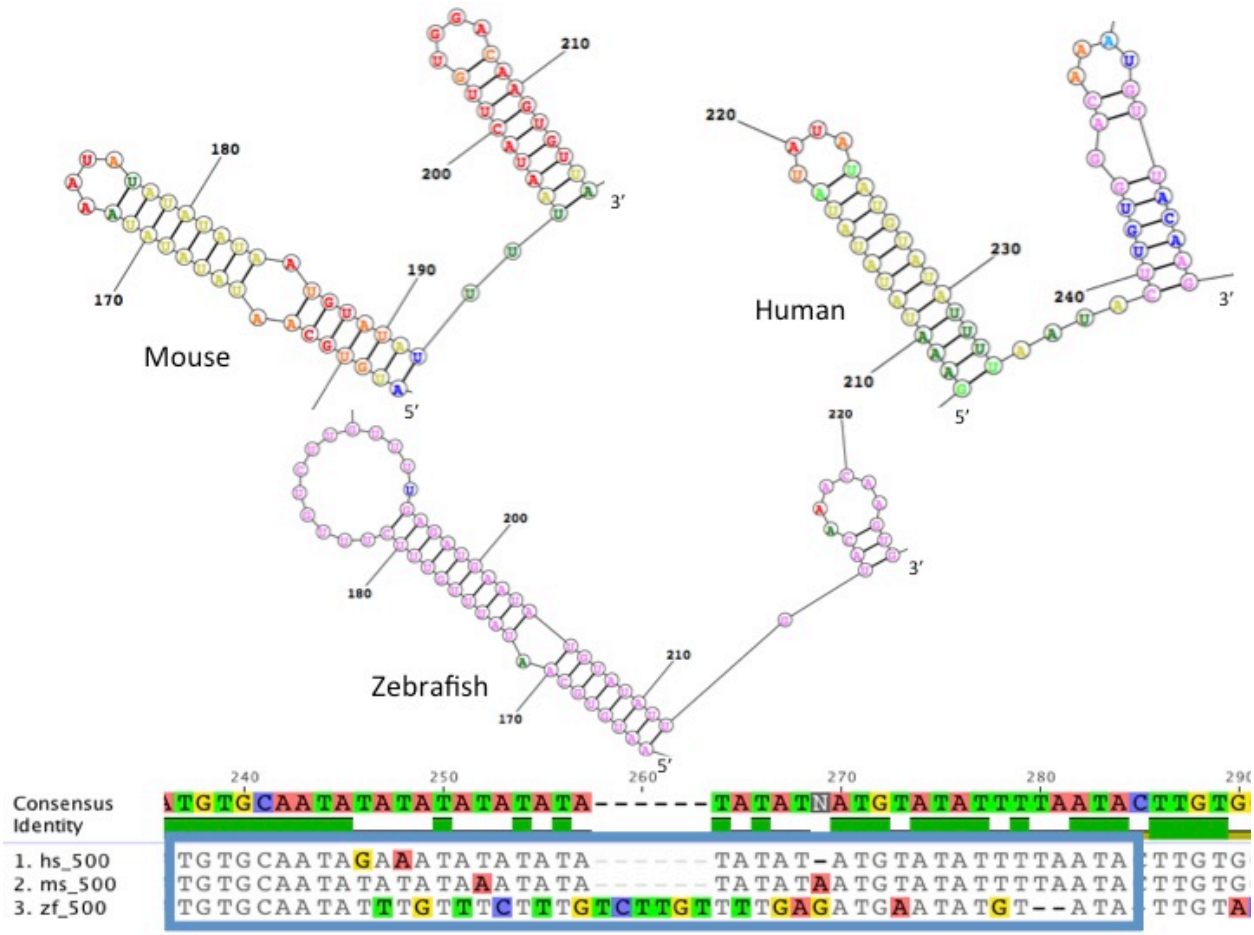


Figure 28: The first domain predicted by SHAPE, is both aligned (blue box) and conserved in structure.

The second and third domains (Figure 29), include the 26-nucleotide core region. The second domain corresponds to H1 (predicted by free energy minimization) and the upper half of the second domain corresponds to H3 (predicted by free energy minimization).

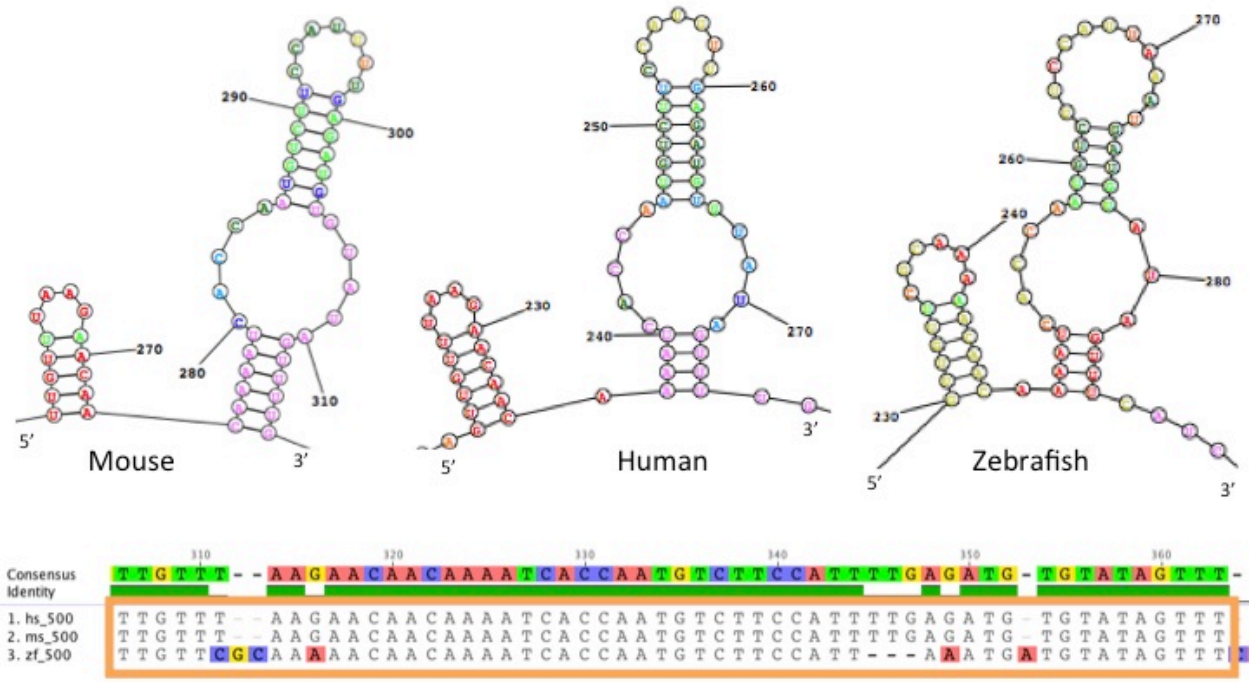


Figure 29: Zebrafish, mouse, and human share two conserved structural domains, as predicted by SHAPE. These regions were also predicted by both free energy minimization and comparative sequence analysis.

The fourth domain (Figure 30), is a hairpin containing two internal loops, and capped by a 10-11 nucleotide apical loop. The helical domain in each of the species contains multiple, consecutive AU base pairs, and the apical loops each contain a 3-nucleotide adenine repeat.

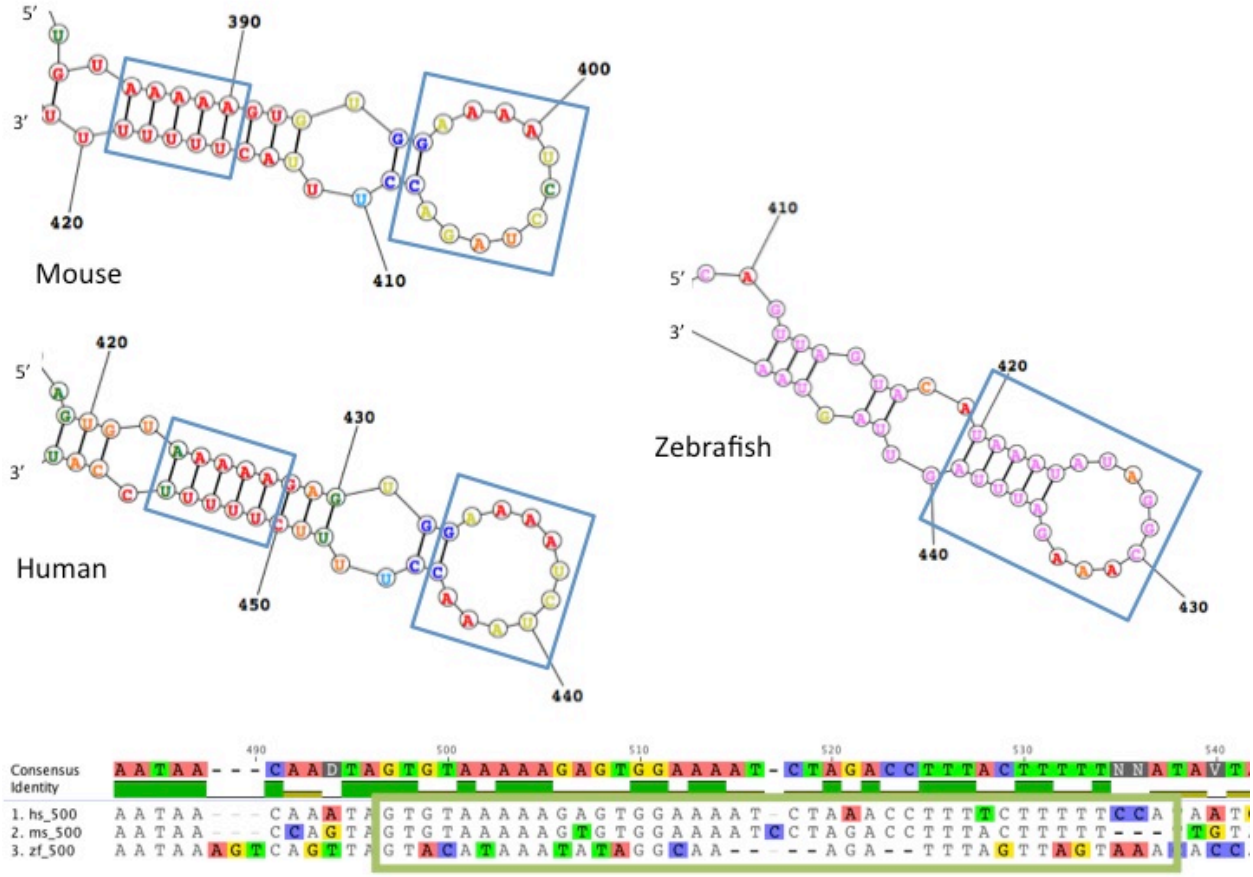


Figure 30: The fourth domain identified by SHAPE contains a series of AU base pairs, along with 3-nucleotide adenine repeats in each of the apical loops.

Given the high sequence similarity between mouse and human, it was not surprising to see motifs that remained structurally conserved between the three species. Further inspection of the domains, such as in domains 1 and 4 for example, revealed co-varying bases (Figure 31, 32).

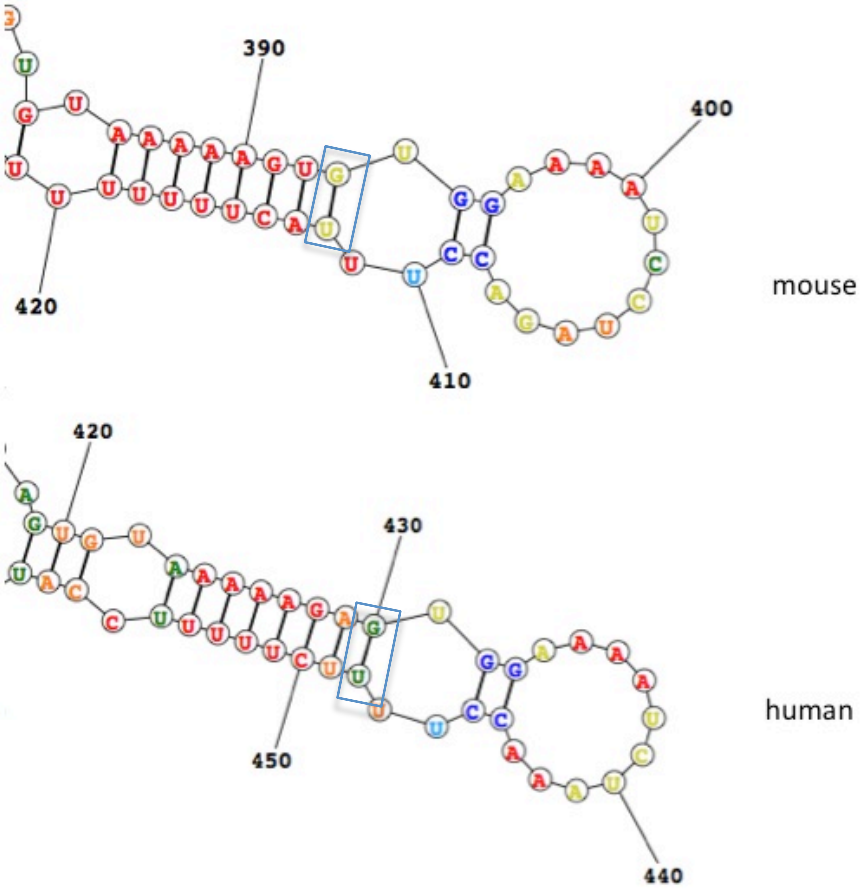


Figure 31: This motif (domain 4) is conserved in both human and mouse. The blue boxes represent an area of covariation.

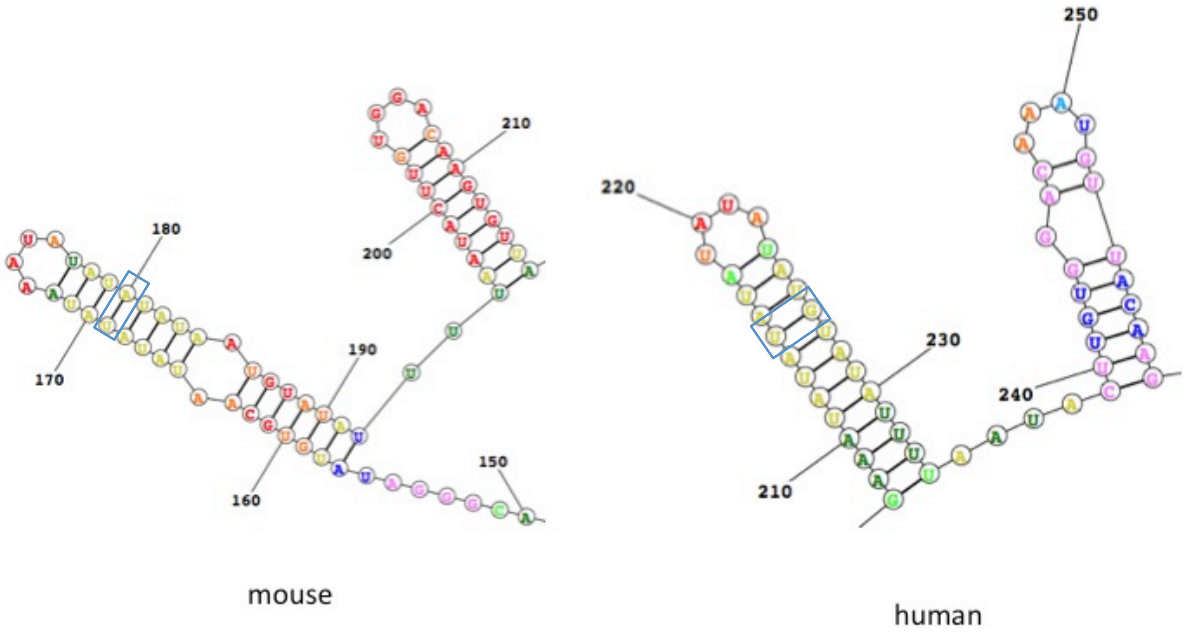


Figure 32: Domain 1 is structurally conserved in both mouse and human due to positional covariation (marked by blue boxes).

Discussion

Performing SHAPE on the 500 nucleotide region revealed that Cyrano lincRNA is folded with the assistance of 100 flanking nucleotides around the 300-nucleotide conserved region. Comparing the SHAPE reactivity of the full-length to the 500-nucleotide fragment yielded similar reactivities, confirming the structure. Both free energy minimization folding and covariation analysis confirmed each of the motifs present in the SHAPE structures.

Our collaborators reported that when the 300-nucleotide conserved region was reintroduced in MO treated zebrafish, morphant phenotypes were unable to be rescued. This could be due to the RNA not being able to fold properly. Inspection of the overall structure revealed that some of the motifs were only present when 100-flanking nucleotides were added to each end of the 300-nucleotide region (Figure 33). I hypothesize that if a 500-nucleotide region were introduced in zebrafish treated with MOs inclusive of the 300-nucleotide conserved region, the morphant phenotypes might be rescued. Alternatively, only a fraction of the motifs may be necessary for proper embryogenesis.

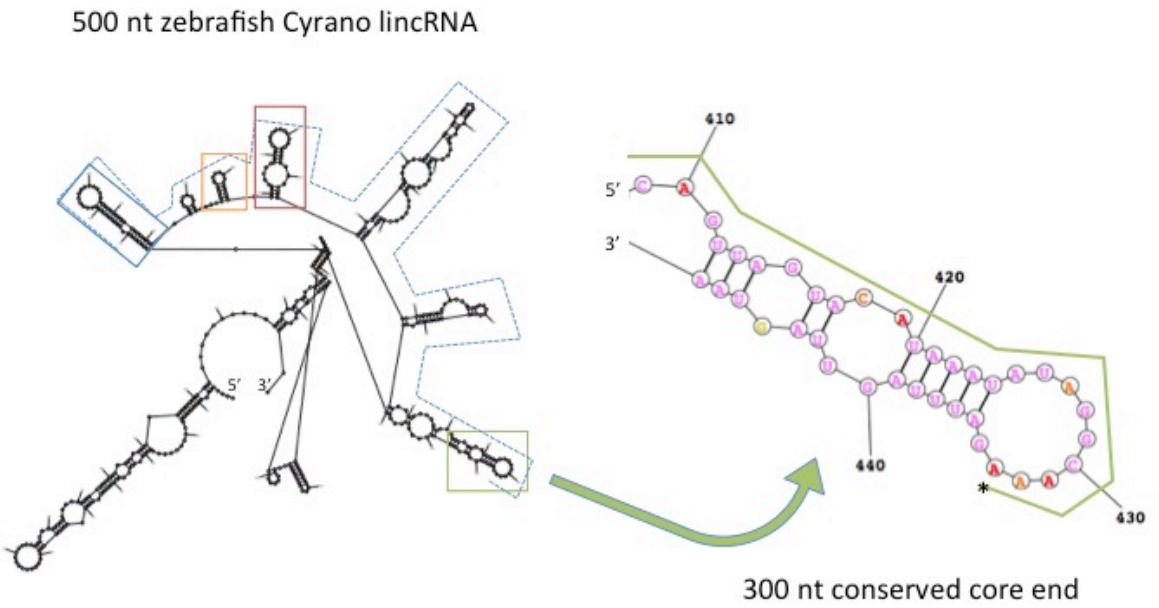


Figure 33: The secondary structure of zebrafish lincRNA revealed that motifs identified by SHAPE are only formed when flanking nucletides are added to each end of the 300-nucleotide core region.

Effects of mutations on secondary structure of zebrafish Cyrano

To investigate the functional importance of the conserved site, Ulitsky et al. introduced point substitutions throughout the conserved site (Figure 34), and tested the potency of the mutated RNAs in rescuing morphant zebrafish.

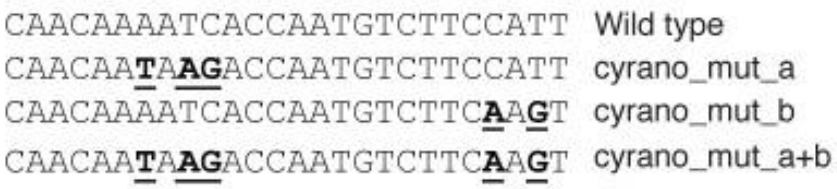


Figure 34. Rescue RNAs were subjected to point mutations throughout the conserved region. Three mutants, Cyrano_mut_a, Cyrano_mut_b, and Cyrano_mut_a+b, were evaluated.

When Cyrano_mut_a and Cyrano_mut_b were injected in morphant zebrafish embryos, the rescuing potential of the zebrafish was reduced, compared to the wildtype, and the rescuing potential of Cyrano_mut_a+b was completely abolished. Inspection of the secondary structure of Cyrano zebrafish lincRNA revealed the positions of the mutations (Figure 35).

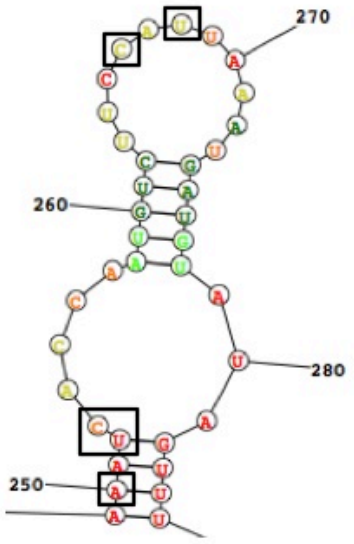


Figure 35: Secondary structure (zebrafish) of the region containing nucleotides subject to point mutations (black boxes).

Because this motif is conserved in all three species that I evaluated, I hypothesize that it serves as a binding partner with either a protein, another RNA, or perhaps (in regard to the apical loop) it serves as part of a pseudoknot interaction. A pseudoknot is an RNA structure that results when a single-stranded region in a hairpin base pairs with a complementary sequence outside of the hairpin. Mutations corresponding to Cyrano_mut_a occur in the lower portion of the structure, at nucleotides 250, 252 and 253. Inspection of the structure suggests that the lower portion of the helix is pulled apart by the A250T and U252A mutations. When this helix was disrupted, it may have weakened this potential binding region, but since the rest of the RNA was still intact, the RNA remained partially functional. Mutations corresponding to Cyrano_mut_b correspond to nucleotides 266 and 268, in the apical loop. I hypothesize that this may be involved in a protein RNA interaction. Despite changes in the apical loop, the rest of the RNA is intact, allowing the RNA to rescue morphant phenotypes. Mutations occurring in one region (but not the other) reduce the ability

of the RNA to rescue morphant phenotypes because the binding domain of the RNA has been slightly altered. Mutations occurring in both regions pose a significant change in the binding region(s), causing the RNA to lose all function.

Further validation of this hypothesis involves mutating the bases paired with 266 and 268, to impose covariation. If the structure remains intact, it is possible that the RNA will still function properly.

Future Work

While SHAPE analysis provided much information about the secondary structure of Cyrano lincRNA in human, mouse, and zebrafish, further validation of these structures will be confirmed with nuclease digestions. The enzymes RNase T1 (which cleaves single stranded guanines) and RNase A1 (which cleaves single stranded cytosines and uracils), RNAse S1 (which cleaves single stranded regions of RNA) and RNAse V1 (which cleaves double stranded regions of RNA) and their cleaving effects on each of the Cyrano lincRNAs will be evaluated.

While SHAPE provides information about which bases are paired and which are unstructured (soft constraints), it fails to provide hard constraints, which indicate the exact pairing of a particular base, as with covariation. This drawback leaves room for the possibility of multiple conformations (as observed in the 20 lowest free energy structures generated for each species) for which each have correctly mapped SHAPE reactivities. Experimentally, co-varying mutations can be made throughout the primary sequence, and the effects of those changes can be observed. Co-variation analysis can provide hard constraints that can be added to guide the secondary folding (this was not done to predict the structures presented here, because they naturally folded that way), but if this route is taken, these hard constraints must be evaluated experimentally.

Once the secondary structure has been validated, mutations disrupting their structure and their subsequent effects on function (in rescue experiments, for example) can be evaluated. Furthermore, the identification of RNA-RNA interactions and protein-RNA interactions, and solving the tertiary structure of the proposed motifs can be pursued.

Chapter 5: Experimental

Preparation of full-length DNA template

DNA plasmid for zebrafish, mouse, and human Cyrano were generously provided by Dr Alena Shkumataa (Unité de Génétique et Biologie du Développement, U934/UMR3215 Institut Curie - Centre de Recherche 26, rue d'Ulm 75248 PARIS Cedex 05 France). The plasmid for each species was transformed into DH5a cells and grown in 25 mL flasks of E. coli. Plasmids were then mini-prepped using a Qiagen midi-prep kit. Plasmids were linearized with the restriction enzymes (New England Biolabs) listed in the table below:

DNA	Enzyme
Human Cyrano	SalI
Mouse Cyrano	SacI
Zebrafish Cyrano	NotI

Plasmid cuts were performed at 37°C (5 µg of plasmid, 50 units of restriction enzyme, 5 µL of Cutsmart Buffer, and H_2O up to 50 µL) for one hr. The digest reactions were stopped by heating the reaction for 20 min at 65°C.

Cuts were confirmed on 1% agarose gels. Linear plasmid was precipitated with 70 µL H_2O, 5 µL of NaCl, 1 µL of 20 µg/µL glycogen, 2 µL of 100 mM EDTA, 350 µL of 4°C 100% ethanol, stored at -80°C for 30 min, and then centrifuged for 20 min at room temperature at 13,000 rpm. The supernatant was removed from the pellet, and the pellet was allowed to air dry for 1 hr. Pellets were resuspended in H_2O to a concentration of 1 µg/µL.

Preparation of 500 nucleotide DNA templates

In order to transcribe 500 nucleotide long RNA fragments (for sequences, see Appendix A), full-length DNA template was subject to a PCR reaction with the following primers (IDT), containing the T7 promoter region (bold text on the forward primer) and a polyA tail (bold text on the reverse primer):

species	Tm	Forward primer	Tm	Reverse primer
zebrafish	**64.5**	**AAGCTTTAATACGACTCACT ATA** GGG ATCTGCTATAGAGCACTGTGA	**60.5**	**TTT TTT TTT TTT TTT T** CCAGAATCGTGCAGCCCTAC
human	**64.7**	**AAGCTTTAATACGACTCACT ATA** GGG GGATATATTCCAGCTGTAGTT GC	**58.3**	**TTT TTT TTT TTT TTT T** CTCAGATTTTGACCCCACATT T
mouse	**64.8**	**AAGCTTTAATACGACTCACT ATA** GGG GTGGCACATTTCCATTTATAGT CT	**58.5**	**TTT TTT TTT TTT TTT T** AGTGGCTCTCAGTGGGAA

DNA template 200 ng was combined with 1 µL of 10 µM forward and 1 µL of 10 µM reverse primer, 5 µL of 10X Phusion buffer (NEB), 1 µL 10 mM dNTPs, 0.5 µL of Phusion enzyme (NEB) and up to 50 µL of H_2O.

The PCR was run using the standard gradient Phusion protocol:

STEP	TEMP	TIME
Initial Denaturation	98°C	30 seconds
25-35 Cycles	98°C 54-65°C 72°C	10 seconds 30 seconds 30 seconds per kb
Final Extension	72°C	5-10 min
Hold	4°C	

Following PCR, DNA template was diluted to a concentration of 600 ng/µL. DNA purity was confirmed on a 1% agarose gel.

RNA transcription

Full length DNA plasmid was transcribed with the enzymes and protocols listed in the table below:

DNA	Transcriptional Promoter
Human Cyrano	T7 (in house *in vitro* transcription)
Mouse Cyrano	T3 (mMessenger mega kit, Ambion)
Zebrafish Cyrano	Sp6 (mMessenger mega kit, Ambion)

500 nucleotide DNA plasmid was transcribed with in house *in vitro* transcription protocol. The protocol for each RNA was performed as follows:

100m $MgCl_2$ (µL)	H_2O (µL)	10x (µL)	NTPs (µL)	1:20 T7 (µL)	DNA (µL)	H_2O (µL)
2	12	2.5	4	2	1	1.5
4	10	2.5	4	2	1	1.5
6	8	2.5	4	2	1	1.5
8	6	2.5	4	2	1	1.5
10	4	2.5	4	2	1	1.5
12	2	2.5	4	2	1	1.5

where 10X buffer contains 1M Tris (pH 8.0), 1 M DTT, 100 mM Spermine, and 10% Triton X-100. NTPs is an equimiolar mix of 25 mM ATP, GTP, UTP, and CTP.

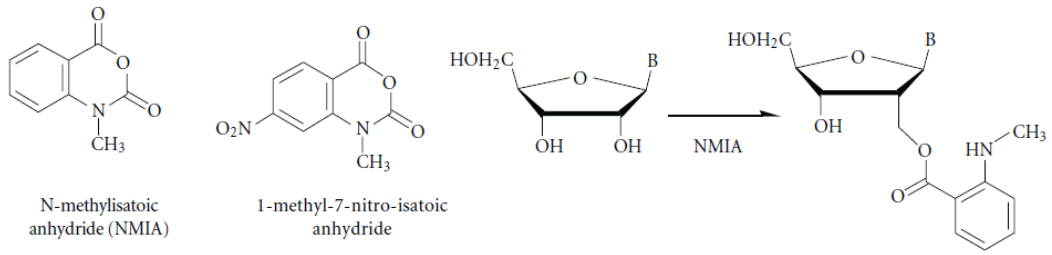
Transcription solutions were incubated at 37°C for 3 hr and the reaction was aborted with 5 µL of 0.5 M EDTA. RNA transcripts were confirmed on a 1% agarose gel alongside a 1kb DNA ruler (Genruler).

Purification of RNA

Each of the RNAs was purified with a phenol:chloroform extraction (350 µL of phenol-chloroform-isoamyl alcohol mixture (Sigma) was added to the transcription reaction, followed by 10 seconds of vortexing, and 1 minute of centrifuging at 13,000 rpm). The RNA was removed from the top aqueous layer and an ethanol glycogen precipitation was performed, as previously described. RNA was then diluted to 1 µM in water.

NMIA labeling

Labeling the RNA with NMIA SHAPE adduct [88]:



SHAPE reagents

was carried out as follows:

36 µL of each 1 µM stock RNA was added to 20 µL of water, followed by the addition of 56 µL of 2x folding buffer (200 mM NaCl, 100 mM HEPES, 0.2 mM EDTA, pH 8.0). The reaction was heated at 95°C for 3 min, and then rapidly cooled on ice for 5 min. The RNA was refolded with the addition of 50 µL of refolding buffer (100 mM NaCl, 50 mM HEPES, 16.5 mM $MgCl_2$, pH 8.0). RNA solution was then distributed in 27 µL aliquots into 6 different tubes. Three microliters of DMSO, 32.5 mM NMIA or 65 mM NMIA (NMIA was dissolved in 100% DMSO) were each added to two of the six tubes.

The reaction was heated at 37°C for 45 min, followed by precipitation with ethanol and glycogen. The RNA pellet was resuspended in 15 µL of RNAse free water.

Shape primer design (gel electrophoresis)

Twenty SHAPE primers spanning the 4.5 kb human Cyrano lincRNA transcript were designed (Appendix B). Four SHAPE primers spanning the 300-nucleotide conserved region were designed for zebrafish (Appendix B). Five SHAPE primers spanning the 300-nucleotide conserved region were designed for mouse (Appendix B).

Dephosphorylated primers were ordered from IDT and diluted to 3 µM and labeled with Υ ^{32}ATP:

5 µL of 3 µM dephosphorylated primer

2 µL PNK Buffer A (New England Biolabs, NEB)

3 µL Υ ^{32}ATP

9 µL H_2O

1 µL PNK enzyme (New England Biolabs)

The reaction was incubated at 37°C for 30 min, followed by the addition of 1 µL of 0.5 mM EDTA, pH 8.0, and incubated at 75°C for 10 min. Primers were purified with a Qiagen nucleotide removal kit. Primers were eluted with 50 µL of elution buffer, bringing the final concentration to 0.3 µM.

Shape primer design (capillary electrophoresis)

Two primer sets (one for both human and mouse), containing four primers, 5' labeled with four of the five dyes from the G5 dye set (NED, VIC, FAM, PET) were ordered from Life Technologies. Primers were diluted to 2 pmol/μL in water.

SHAPE (gel electrophoresis)

Sequencing reaction preparation

1 μL of unmodified 1 μM RNA was combined with 2.5 μL of RNAse free water and 1 pmol of Υ^{32}ATP labeled primer. 1 μL 10 mM ddNTP (either A, G, C, or T) was added to the solution.

SHAPE reaction preparation

5 μL of modified or DMSO treated RNA (~6 pmol) was added to 1 pmol of Υ^{-32}ATP labeled primer.

Reverse transcription

Sequencing and SHAPE reactions were each combined with 6.1 and 8.1 μL, respectively, of enzyme mix (Sigma), and incubated at 52°C for 1 minute; 0.5 μL of SuperScript reverse transcriptase III was then added to each reaction. The reactions were incubated for 10 min at 52°C, followed by the addition of 0.6 μL of 4 M NaOH. The reactions were heated for 5 min at 95°C. Following heating, 14.5 μL of acid stop solution (432 μL 2X Urea Loading Buffer, 64 μL 1 M Tris) was added to each reaction; these were then heated at 95°C for 5 min, then allowed to cool at room temperature.

Gel analysis

Sequencing and SHAPE reactions were analyzed on both 8% acrylamide gels (17.5 mL H_2O, 5 mL 10 X TBE, 10 mL 40% 19:1 bis acrylamide, 24 g urea, 500 µL 10% ammonium persulfate, 50 µL TEMED). Gels were run for either 2 or 5 hr at 70 W. Gels were then dried and exposed to a phosphor plate overnight. The phosphor plates were scanned using a Typhoon 600 imager, followed by analysis with SAFA (Semi Automated Footprinting Analysis software). Reactivity data were then incorporated as a shape constraint file in the RNAstructure folding program, and the 20 lowest energy structures based on those constraints were generated.

SHAPE (capillary electrophoresis)

Sequencing reaction preparation

5 µL of 1 µM unmodified RNA was combined with 2 pmol of each fluorescent labeled primer.

SHAPE reaction preparation

5 µL of modified or DMSO treated RNA (~6 pmol) was added to 2 pmol of each fluorescent labeled primer. The SHAPE and sequencing reactions were placed in a thermocycler and heated to 95°C for 1 minute, followed 5 min incubation at 65° C, 5 min at 35°C. The reaction was then immediately cooled on ice for 1 minute.

Reverse transcription

5 µL of a 10 mM ddNTP was added to sequencing tubes. Then, 8.6 µL of enzyme mix containing Superscript III reverse transcriptase (0.5 units per reaction) (Sigma) was added to each tube. The reactions were placed in a thermocyler and subjected to 42°C incubation for 1 minute, followed by incubation at 52°C for 25 min. Four microliters of 0.5 mM EDTA was added to each tube to quench the reaction. An ethanol glycogen precipitation was performed and pellets were redissolved in 10 µL of formamide.

Capillary electrophoresis

cDNA fragments were analyzed using a 3130xl Genetic Analyzer. SHAPE reactivity was normalized using QuSHAPE [94]. Reactivity data were then incorporated as a shape constraint file in the RNAstructure folding program, and the 20 lowest energy structures based on those constraints were generated.

RNA folding

RNAstructure [95] was downloaded from http://rna.urmc.rochester.edu/ and the 4.5 kb, 2 kb, and 500 nucleotide sequences (Appendix B) of human, mouse, and zebrafish Cyrano RNAs were folded. The secondary structure calculation parameters are listed in the table below:

Temperature (K)	310.15
Maximum loop size	30
Maximum percent energy difference (MFE, MEA)	10
Maximum number of structures (MFE, MEA)	20
Window size (MFE, MEA)	3
Gamma (MEA)	1
Iterations (Pseudoknot Prediction)	1
Minimum helix length (Pseudoknot Prediction)	3

Bibliography

1. Mattick, J.S. and I.V. Makunin, *Non-coding RNA.* Hum Mol Genet, 2006. **15 Spec No 1**: p. R17-29.
2. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
3. Mehler, M.F. and J.S. Mattick, *Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease.* Physiol Rev, 2007. **87**(3): p. 799-823.
4. Mouse Genome Sequencing, C., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
5. Weinberg, R.A. and S. Penman, *Small molecular weight monodisperse nuclear RNA.* J Mol Biol, 1968. **38**(3): p. 289-304.
6. Paul, J. and J.D. Duerksen, *Chromatin-associated RNA content of heterochromatin and euchromatin.* Mol Cell Biochem, 1975. **9**(1): p. 9-16.
7. Salditt-Georgieff, M., et al., *Large heterogeneous nuclear ribonucleic acid has three times as many 5' caps as polyadenylic acid segments, and most caps do not enter polyribosomes.* Mol Cell Biol, 1981. **1**(2): p. 179-87.
8. Salditt-Georgieff, M. and J.E. Darnell, Jr., *Further evidence that the majority of primary nuclear RNA transcripts in mammalian cells do not contribute to mRNA.* Mol Cell Biol, 1982. **2**(6): p. 701-7.
9. Nickerson, J.A., et al., *Chromatin architecture and nuclear RNA.* Proc Natl Acad Sci U S A, 1989. **86**(1): p. 177-81.
10. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
11. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
12. Hamilton, A.J. and D.C. Baulcombe, *A species of small antisense RNA in posttranscriptional gene silencing in plants.* Science, 1999. **286**(5441): p. 950-2.
13. Timmons, L., *The long and short of siRNAs.* Mol Cell, 2002. **10**(3): p. 435-7.
14. Taft, R.J., et al., *Small RNAs derived from snoRNAs.* RNA, 2009. **15**(7): p. 1233-40.
15. Thomson, T. and H. Lin, *The biogenesis and function of PIWI proteins and piRNAs: progress and prospect.* Annu Rev Cell Dev Biol, 2009. **25**: p. 355-76.
16. Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.
17. Navarro, P., et al., *Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program.* Genes Dev, 2006. **20**(20): p. 2787-92.
18. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.
19. Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.
20. Plath, K., et al., *Role of histone H3 lysine 27 methylation in X inactivation.* Science, 2003. **300**(5616): p. 131-5.
21. Ulitsky, I. and D.P. Bartel, *lincRNAs: genomics, evolution, and mechanisms.* Cell, 2013. **154**(1): p. 26-46.

22. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Res, 2012. **22**(9): p. 1775-89.
23. Cabili, M.N., et al., *Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution.* Genome Biol, 2015. **16**: p. 20.
24. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-69.
25. Gezer, U., et al., *Long non-coding RNAs with low expression levels in cells are enriched in secreted exosomes.* Cell Biol Int, 2014. **38**(9): p. 1076-9.
26. Eddy, S.R., *Computational analysis of conserved RNA secondary structure in transcriptomes and genomes.* Annu Rev Biophys, 2014. **43**: p. 433-56.
27. Beletskii, A., et al., *PNA interference mapping demonstrates functional domains in the noncoding RNA Xist.* Proc Natl Acad Sci U S A, 2001. **98**(16): p. 9215-20.
28. Chaumeil, J., et al., *A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced.* Genes Dev, 2006. **20**(16): p. 2223-37.
29. Hall, L.L. and J.B. Lawrence, *XIST RNA and architecture of the inactive X chromosome: implications for the repeat genome.* Cold Spring Harb Symp Quant Biol, 2010. **75**: p. 345-56.
30. Somarowthu, S., et al., *HOTAIR forms an intricate and modular secondary structure.* Mol Cell, 2015. **58**(2): p. 353-61.
31. Xu, Z.Y., et al., *Knockdown of long non-coding RNA HOTAIR suppresses tumor invasion and reverses epithelial-mesenchymal transition in gastric cancer.* Int J Biol Sci, 2013. **9**(6): p. 587-97.
32. Chen, F.J., et al., *Upregulation of the long non-coding RNA HOTAIR promotes esophageal squamous cell carcinoma metastasis and poor prognosis.* Mol Carcinog, 2013. **52**(11): p. 908-15.
33. Liu, X.H., et al., *The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer.* BMC Cancer, 2013. **13**: p. 464.
34. Novikova, I.V., et al., *Rise of the RNA machines: exploring the structure of long non-coding RNAs.* J Mol Biol, 2013. **425**(19): p. 3731-46.
35. Kretz, M., et al., *Control of somatic tissue differentiation by the long non-coding RNA TINCR.* Nature, 2013. **493**(7431): p. 231-5.
36. Kretz, M., et al., *Suppression of progenitor differentiation requires the long noncoding RNA ANCR.* Genes Dev, 2012. **26**(4): p. 338-43.
37. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annu Rev Biochem, 2012. **81**: p. 145-66.
38. Penny, G.D., et al., *Requirement for Xist in X chromosome inactivation.* Nature, 1996. **379**(6561): p. 131-7.
39. Tsai, M.C., et al., *Long noncoding RNA as modular scaffold of histone modification complexes.* Science, 2010. **329**(5992): p. 689-93.
40. Wang, K.C., et al., *A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.* Nature, 2011. **472**(7341): p. 120-4.
41. Bertani, S., et al., *The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin.* Mol Cell, 2011. **43**(6): p. 1040-6.
42. Sleutels, F., R. Zwart, and D.P. Barlow, *The non-coding Air RNA is required for silencing autosomal imprinted genes.* Nature, 2002. **415**(6873): p. 810-3.
43. Pandey, R.R., et al., *Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation.* Mol Cell, 2008. **32**(2): p. 232-46.

44. Yap, K.L., et al., *Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a.* Mol Cell, 2010. **38**(5): p. 662-74.
45. Yao, H., et al., *Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA.* Genes Dev, 2010. **24**(22): p. 2543-55.
46. Jeon, Y. and J.T. Lee, *YY1 tethers Xist RNA to the inactive X nucleation center.* Cell, 2011. **146**(1): p. 119-33.
47. Lai, F., et al., *Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.* Nature, 2013. **494**(7438): p. 497-501.
48. Gomez, J.A., et al., *The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus.* Cell, 2013. **152**(4): p. 743-54.
49. Grote, P., et al., *The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse.* Dev Cell, 2013. **24**(2): p. 206-14.
50. Schoeftner, S., et al., *Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing.* EMBO J, 2006. **25**(13): p. 3110-22.
51. Klattenhoff, C.A., et al., *Braveheart, a long noncoding RNA required for cardiovascular lineage commitment.* Cell, 2013. **152**(3): p. 570-83.
52. Zhao, J., et al., *Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.* Science, 2008. **322**(5902): p. 750-6.
53. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.
54. Mercer, T.R., et al., *Targeted RNA sequencing reveals the deep complexity of the human transcriptome.* Nat Biotechnol, 2012. **30**(1): p. 99-104.
55. Ravasi, T., et al., *Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome.* Genome Res, 2006. **16**(1): p. 11-9.
56. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.
57. Ponjavic, J. and C.P. Ponting, *The long and the short of RNA maps.* Bioessays, 2007. **29**(11): p. 1077-80.
58. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**(5): p. 503-10.
59. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
60. Sigova, A.A., et al., *Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2876-81.
61. Khalil, A.M., et al., *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(28): p. 11667-72.
62. Jia, H., et al., *Genome-wide computational identification and manual annotation of human long noncoding RNA genes.* RNA, 2010. **16**(8): p. 1478-87.
63. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.

64. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
65. Guo, H., et al., *Mammalian microRNAs predominantly act to decrease target mRNA levels.* Nature, 2010. **466**(7308): p. 835-40.
66. Li, X., et al., *A MicroRNA Imparts Robustness against Environmental Fluctuation during Development.* Cell, 2009. **137**(2): p. 273-282.
67. Tinoco, I., Jr. and C. Bustamante, *How RNA folds.* J Mol Biol, 1999. **293**(2): p. 271-81.
68. Ulloa, P.E.I., P.; Neira, R.; Araneda, C., *Zebrafish as a model organism for nutrition and growth: towards comparative studies of nutritional genomics applied to aquacultured fishes.* Rev Fish Bio Fisheries, 2011(21): p. 649 - 666.
69. Mathews, D.H., W.N. Moss, and D.H. Turner, *Folding and finding RNA secondary structure.* Cold Spring Harb Perspect Biol, 2010. **2**(12): p. a003665.
70. Zuker, M. and D. Sankoff, *RNA secondary structures and their predictions.* Bulletin of Mathematical Biology, 1984. **46**(4): p. 591 - 621.
71. Ruzzo, W.L. and J. Gorodkin, *De novo discovery of structured ncRNA motifs in genomic sequences.* Methods Mol Biol, 2014. **1097**: p. 303-18.
72. Lee, J.T., *Epigenetic regulation by long noncoding RNAs.* Science, 2012. **338**(6113): p. 1435-9.
73. Callinan, P.A. and M.A. Batzer, *Retrotransposable elements and human disease.* Genome Dyn, 2006. **1**: p. 104-15.
74. Wang, H., et al., *SVA elements: a hominid-specific retroposon family.* J Mol Biol, 2005. **354**(4): p. 994-1007.
75. Hancks, D.C. and H.H. Kazazian, Jr., *SVA retrotransposons: Evolution and genetic instability.* Semin Cancer Biol, 2010. **20**(4): p. 234-45.
76. Gutell, R.R., N. Larsen, and C.R. Woese, *Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective.* Microbiol Rev, 1994. **58**(1): p. 10-26.
77. Crick, F.H., *The origin of the genetic code.* J Mol Biol, 1968. **38**(3): p. 367-79.
78. Juhling, F., et al., *Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements.* Nucleic Acids Res, 2012. **40**(7): p. 2833-45.
79. Pace, N.R.e.a., *Probing RNA structure, function, and history by comparative analysis.*, in *The RNA World*. 1999, CSHL Press. p. 113 - 141.
80. Yao, Z., Z. Weinberg, and W.L. Ruzzo, *CMfinder--a covariance model based RNA motif finding algorithm.* Bioinformatics, 2006. **22**(4): p. 445-52.
81. Torarinsson, E. and S. Lindgreen, *WAR: Webserver for aligning structural RNAs.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W79-84.
82. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
83. Gardner, P.P., A. Wilm, and S. Washietl, *A benchmark of multiple sequence alignment programs upon structural RNAs.* Nucleic Acids Res, 2005. **33**(8): p. 2433-9.
84. Wilkinson, K.A., E.J. Merino, and K.M. Weeks, *Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution.* Nat Protoc, 2006. **1**(3): p. 1610-6.
85. Mortimer, S.A. and K.M. Weeks, *A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry.* J Am Chem Soc, 2007. **129**(14): p. 4144-5.

86. Watts, J.M., et al., *Architecture and secondary structure of an entire HIV-1 RNA genome.* Nature, 2009. **460**(7256): p. 711-6.
87. Robertson, M.P., et al., *The structure of a rigorously conserved RNA elementwithin the SARS virus genome.* PLoS Biol, 2005. **3**(1): p. e5.
88. Deigan, K.E., et al., *Accurate SHAPE-directed RNA structure determination.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 97-102.
89. Wilkinson, K.A., E.J. Merino, and K.M. Weeks, *RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts.* J Am Chem Soc, 2005. **127**(13): p. 4659-67.
90. Duncan, C.D. and K.M. Weeks, *SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA.* Biochemistry, 2008. **47**(33): p. 8504-13.
91. Hartl, M.J., et al., *Regulation of foamy virus protease activity by viral RNA: a novel and unique mechanism among retroviruses.* J Virol, 2011. **85**(9): p. 4462-9.
92. Wilkinson, K.A., et al., *Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA.* RNA, 2009. **15**(7): p. 1314-21.
93. Kladwang, W., et al., *Understanding the errors of SHAPE-directed RNA structure modeling.* Biochemistry, 2011. **50**(37): p. 8049-56.
94. Karabiber, F., et al., *QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis.* RNA, 2013. **19**(1): p. 63-73.
95. Reuter, J.S. and D.H. Mathews, *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC Bioinformatics, 2010. **11**: p. 129.

Appendix B

The following sequence corresponds to full-length human Cyrano lincRNA. Underlined regions correspond to primers designed for SHAPE analysis:

5′
CUAUAGGGAAGCUGGUACGCCUGCAGGUACCCGGGAAUUCGGCCAUUAUGGCCGGGGGA
GAAGCUGCGAAGAUGGCGGAGUAAGGCGUGCCGCUGCAAACUGGCCUCUGGGCCGGGG
GCGAGCAGCCCCCGGGAGGCCGAGUGCAUCUGUUGGACCGUGCGAGGAGAAGAAAAAAA
AUAUCGGCCAGAGGAGAGGAACUAACCGAACAUUCUUCCUCCCUACCUUAUAGAGGGGAG
UGGUCAUCUACACUAAAGCAAAGUGUUAGGCGCGUCCGUGCGAAGAGACCACCAAACAGG
CUUUGUGUUCCUUAUCACAGGAAGAAAAUUUUCCUUGACCUUUAGGUGCUUUUAUAUUC
AUCUCAAAAACAAAAUUCUGAACUCAGGACUUGGCAAGUGUCUCUAUGUUGUCUCCUAGA
GUGGGUAGUCCUGCUUCUUUUACCCAGUUACUUUCCCGUUUUUGAAAUGUGGCUAUCAC
UUCUCUACACAUUACCUCCAUGAUUUGGAAUGGAAAAGGCCACUUUUCUUUUUGUUCUG
CCUCUCAAAUUCAACACAGAGGAGCUCCUAGGAUUCCAGUUAUCCUGCUAACAUCUCCAG
GAAGAAAGAAGCAACUCACAUGGGUCUUUUGCUGUUUGCUUAAUUAUAAAGACAUCAUUU
UGCAAGCAGAAGGCUGAGUUUCAUUUGAAACAGGUGCUUAGGUGGUGGUAUUUGUGAAU
ACUUUUCAUUCCAAGCAAGAAGACUAAAGAAGUAGCAAGUAUGAAUGACUUCAGGGUUUA
AAAAAAAUGUCUUCCAGUUUCAGCCACUACCAUGAUAAGCACAGUUGAGACUGCAGCAGU
AAAUUCCAAAUAUGUGUUUCUAAUUUGACGUGAAAGAUACUAAAAAUUUAUAUUUGUAUA
UUUAAAUCCUGGCUCAUCCUGUGACAUAGAUUUACUGAAUAGGAACAAAGGCCCAAUUUU
AAACAAAAACCUAGGCCGGGUGUGGUGGCUCACACCUGUAAUCCCAACACUUUGGGAGG
CCAAGGCAGGCAAAUCACCUGAGGUUGGGAGUUUGAGACCAGCCUGACCAACAUGGAGA
AACCCCAUCUCUACUAAAAAUAGAAAAUUAGCCAGGUGUGGUGGUAGAUACGUGUAAUCC
CAGCUACUCGUGAGGCUGAGGCAGGAGAAUUGCUUGAAUCCGGGAGGCGGAGGUUGCA
GUGAACUGAGAUUGCACCACUGCACUCCAGCCUGGGCGACAGAGUGAGACUCCGUCUCA
GAAAAAUAAAAAACAACUUUGGACAUAGAGGAGUUGGGGGCAGAGGGGGGAGGAUGAGA
GAAUUUUCCAUGUAACUCCUUUUCCUUAAGAAAAAGAAGCAAAAAUCCCUCAUGCAGUGC
CAUCUGACUUUAUGGUGUUUCACAUUAUAUAGUUACUUUUUUUAAUAAACGCAUAAGAUU
AACUCUUCCUGCAACCCAAGGUGGAUACUUGGAUGUUUGAUUUUUAUUUAUUUAUUUAU
UUAUUUUUUGAGGCAGUCUCUCUCUGUUGCCCAGGCUGGAGUGAUGCGUGAUUUCUGCU
CACUGCAGUCUCUCCCACCCGGGUUCAGGCGAUUCUCCUGCCUCAGCCUCCCAAGUAGCU
AGGAUUACAGGCGCGCACCACCACGCUCAGCCUGAUUUCUAGGCUGAAAAGGCUGGGAC
UCAGGUACUUUCCCCAGUUGGGACUGAACUUUCUCAUCAGAGAAUGGGCCUCAGAAGCA
AUGUUCCCAAACUUGUGGUUGGGUCAUCGUGGAAAAGAAACCUCAAAUAAGAAAAAUAAU
UACAAUAAGAAAUGGAUUUUCUUUUUUCCCACAACAGUUAUACAUUAUAAAGAACAGACU
GUCGUAGAAAACUGUCUUUGCUUCCAAAUCAGCAGAGGACCAUUGUAUGUAUUGUCAGG
UCUUUAUAUAAGAGUGAAACCUUUAUUUAUGCUUCUUGUGAGUAGAGAAUAAAUUUUAAA
ACUAAUUAGAUGAAAUUAAGAACAGAGUAUUGGAACAUCUCUGUGUUUUCCGGAAUGUU
UACUCAGGUCCAUGAAUGCUGUGAUGCUGGGAACUUAGGGAAGAUUCACCAAAAUUUGA
GAGUGAUAAAAAUGGCCUAAAAUGGGAAAUGGUGAGCACUCUUAUUAUUAGUGAGAUUG
GUCAAUCAGCCCAUUUGAAAUGGGUAAAAAAUACUUUCAGUAAAAUAAUUACAUAUUAUA
UAUAAAAUAUUUCUUUUUUUGAGGGCAGCUACUGAGUGAUAAAAGCAUAGUAGAAAUCAC
UUUGGUUAAAAGUAACUCAAAUUUUUCUCUUAAGUAAUCUACCAACUUCCUGAUGUUUUU
CUUUAAUAUUUUGGAACUGUUCAAGACAGAAUGGAGCCUACAUAUAUGGUGUUCCACUA
GUAUGUUGAAAAUGCCAUAUCAUGGAGAAUGGAGACACCUUCCAGGUGUCUGUUAAACC
CAUCUUCUCUGUGUACUUCUGGCAUCUUUUUUGGUAGGAUCAUUUGGCAGGGGGUAGAG
UACCUGUACUUUUGGCACCAUUGAAACCAGCUCUGGCCCACUUGUUUGAAUAGCUAUCAG
ACUCAGCGUCUCUAUAUGCUUUAUAUACAGUUGAGAGAACCAGUGUUUAAUGCUCUUGA

AAUUGUAUCUUCCAUGCUUCAGACCAGAUGCAUUUAACAAAAUGGAUAUAUUCCAGCUGU
AGUUGCCCAGUGUUUACUUAACACAUCUACAUUUUUUUCUUGUCUAUUUUGGUCCCCUU
GAUAGGAAAAGCUAUAAUUUUAGGCAGGACUAUACGUCGAUUUGUAGCCAUGCUUCCUU
CCUUUCCCUUGCUCAUCCAUGUUAGCUGGCAGUUUUUCUUUUGAAAAGUUAAAACCGGG
AUAUGUGCAAUAGAAAUAUAUAUAUAUAUGUAUAUUUUAAUACUUGUGGACAAAUGUU
ACAAGUUGUUUAAGAACAACAAAAUCACCAAUGUCUUCCAUUUUGAGAUGUGUAUAGUUU
UGUAAGCAUUAGUGCUUGGUAGCAUAUUGUAGUGCCAUGUUAGGGGUUAGUGCAUGAGU
CUAGUGAUUUUAAACUUCAGGAUGAAUUAUUGAUAAUAACAAAUAGUGUAAAAAGAGUGG
AAAAUCUAAACCUUUUCUUUUUCCAUAAUGUCUAAAUCUGUUAUAUUCUUCCUGGGGAAA
AGAGAUUAAGGCCCAAAAGACUCAUUUAUGAAUAGAAAUGUGGGGUCAAAAUCUGAGAUA
CUAUAUUUGAGGACAUUUCAGCUUUCCAUAAGGUAUCUGGAAACCAGCUGUCUUUGUGU
CUUAUGAAACCUCAAGUCAAAAUGAGACCGCAUUUAAUUAUUCUGCUUUGCUCUUUUUUU
GGGGUGGGGGGGAGAUAGAGUUUCACUCUGUCACCCAGGCUGGAGUGCAGUGGCACAA
UCUUUGCUCACUGUAACCUCUGCCUCUUGGGUUCAAGUGAUUCUCGUGCCCCAGCCUCC
CGAGUAGCUGGGACUACAGGCACGUGCCACCACGCCCAGCUAAUUUUUAUAUUUUUAGU
AGAGACGGGGUUUUGCUGUGUUGGCCAGGCUGGUCUCAAACUCCUGACCUCAAGUAAUC
CACCUGGCCUGCUCUUUUCAUGUCUUAACAUGGCAUGUCUUUUAGUUUCAUUAUUUUCC
UACUCCUUGUAUGUCAAGAAAUUACAUUUUGCAUGUCUUAUGGAGAUGCUGUUAAUUGC
UUCAGUGAGUGCUUUUCUAAUCUGCAGACCAUUUACAUUUCCUGUUUGCAGCAUGCUGU
GUGCAAACACUCAGUAAUUUGGAGUAUUCAAUUAUUUGUUAGGGCUCUUCCUAUUUCCA
AAUGUGCUGAAUUGUCUAUUGAUGGGAUUUUCAGAUCUUUUCAUGAGAACUGGAAAUGU
AGCUGGGUGGCACCUACCUAGGUUGCUACGUAGUGAGUAGACUUUCUCUUGGGUAUAGU
AAGCCUCAGACAGCUUUCACUUUUAUCUACUUUACUUGUGGAAAUAAAACAGUCAUUUUG
UUCUGAAAGAAUAAGAUAGCUUUCUGUAGAGAAGGAAUUCCUACCUCUAAAAGCUGCCUU
GAGAACUCAGAACUGGCAGUUUUCUGAGGUGAUUUUUAAAUUUCAGUAUUAGGGAGAGU
CCAGCAUUUGCUGACACAGAUUCUACAUAACUAAUGUAUGAUAGCAAAUGCAAAACUAUU
AUAAUGUGGUGUAUCUUGCGCAUACACAGGUUAGAACAAGUAGACUCUGGCAGCAGAUC
UCCAGAGACCCAAGUUUAGGUUCUCAUAGUGUAUUUGAAGUAGUUAUACUCCUGGCUUA
AGUAGUUUAGUGCCUGGGAGAAUCCAUUACUGAAAAGCAUUUAACUUAAGAAAAAAAAAA
AAAAAAAAAACAUGUCGGCCGCCUCGGCCCAGUCGACUCUAGACUCGAGCAAGCUUACGU
ACGCGUGCAUGCGACGUCAUAGCU 3’

The following sequences correspond to the approximate 500-nucleotide long RNAs transcribed and analyzed with SHAPE and free energy minimization folding. Underlined regions correspond to the SHAPE primers used for probing both full-length and fragmented RNAs:

Zebrafish RNA:

5’-
AUCUGCUAUAGAGCACUGUGAAAAUAGGAAUUUCUCUUCAUCUGUUGUGACACUACAUCU
UUUUUGUUGUUGUUGGUAUAGUAGUUCCUAUCAUACGGAUGGUAGAUAUUCAAGUGUUC
UGAGUUUUCAUGAGAGGAAGAGUAGCGAGGCCUCAGUGUGGGAAAUGUGCAAUAUUUGU
UUUCUUUGUCUUGUUUUGAGAUGAAUAUGUAUAUUGUACAAACAAGUGACAAGUUGUUCG
CAAAAACAACAAAAUCACCAAUGUCUUCCAUUAAAUGAUGUAUAGUUUCAUCUGCUUGCU
UUAGGGUGGCGGGAUUUGUAGUGCCAGCGUAGACAUGGACAUGCAUGUACGGUAGCGA
UGCUUGCUGAGUUUAAUAGUCGUAGUGACAUUCAGUAUAGGCUAUAAUAAAGUCAGUUA
GUACAUAAAUAUAGGCAAAGAUUUAGUUAGUAAACACCAUAGAUGCAGUAUCCACACCUU
GGUUAAUUUGAUUAAGAAGCGAGUAGUAGGGCUGCACGAUUCUGG-3’

Human RNA:

5’-
GGAUAUAUUCCAGCUGUAGUUGCCCAGUGUUUACUUAACACAUCUACAUUUUUUUCUUG
UCUAUUUUGGUCCCCUUGAUAGGAAAAGCUAUAAUUUUAGGCAGGACUAUACGUCGAUU
UGUAGCCAUGCUUCCUUCCUUUCCCUUGCUCAUCCAUGUUAGCUGGCAGUUUUUCUUUU
GAAAAGUUAAAACCGGGAUAUGUGCAAUAGAAAUAUAUAUAUAUAUGUAUAUUUUAAU
ACUUGUGGACAAAUGUUACAAGUUGUUUAAGAACAACAAAAUCACCAAUGUCUUCCAUUU
UGAGAUGUGUAUAGUUUUGUAAGCAUUAGUGCUUGGUAGCAUAUUGUAGUGCCAUGUUA
GGGGUUAGUGCAUGAGUCUAGUGAUUUUAAACUUCAGGAUGAAUUAUUGAUAAUAACAA
AUAGUGUAAAAAGAGUGGAAAAUCUAAACCUUUUCUUUUUCCAUAAUGUCUAAAUCUGUU
AUAUUCUUCCUGGGGAAAAGAGAUUAAGGCCCAAAAGACUCAUUUAUGAAUAGAAAUGUG
GGGUCAAAAUCUGAG -3’

Mouse RNA:

5’-
GUGGCACAUUUCCAUUUAUAGUCUUACUUGAGUUUUCUUCUUAAUAGCAGCAAAGACUAC
CAUUUUAGAGUUGUGCCUCCAUUUAUAGCUGUGCCUCCUCCCUUCUCCAUGUAAGCUGG
CAGUUUUUCUUUUGAAGUUUCUAAACAAAAACGGGAUAUGUGCAAUAUAUAUAAAUAUAU
AUAUAAUGUAUAUUUUAAUACUUGUGGACAAGUGUUACAAGUUGUUUAAGAACAACAAAA
UCACCAAUGUCUUCCAUUUUGAGAUGUGUAUAGUUUUGUAAGCAGUAGUGCUUGAUAGC
AUAUUGUAGUGCCAUGUUAGGGGUUAGUGCAUGAAUUUAGUAUAAUUUUAAACUUCAAG
AUGAAAUAUUGAUAAUAACCAGUAGUGUAAAAAGUGUGGAAAAUCCUAGACCUUUACUUU
UUUUGUAUAUGAAUAUUUGAUAGUCUCCCUGCGGAUAGGAGAAAAGGCCAAAACAACUAA
UGUCUGAACAGAAGUGUGGCUUCCCACUGAGAGCCACU-3’

Capillary electrophoresis primer sequences:

Human: 5’ GCT GGT TTC CAG ATA CCT TAT GG 3’

Mouse: 5’ GCT ATT CTA GCT CCC CCG TG 3’

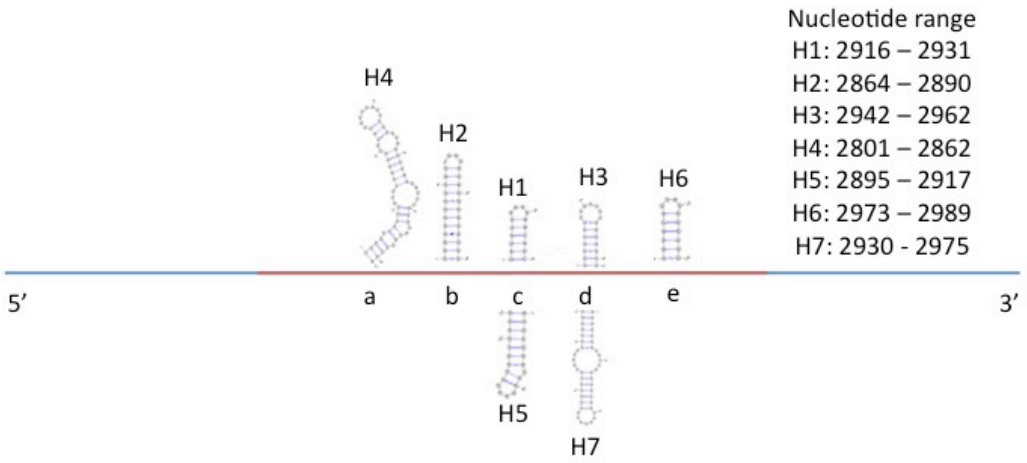


Figure B1: Relative location of each of the seven motifs predicted by Mfold for the 500 nucleotide human Cyrano lincRNA. Blue corresponds to the 500-nucleotide RNA, and red corresponds to the 300-nucleotide conserved region. In positions b and c, where two motifs are present, 50% of the structures predicted the top motif, while the other 50% of the structures predicted the lower motif for that region of the sequence.

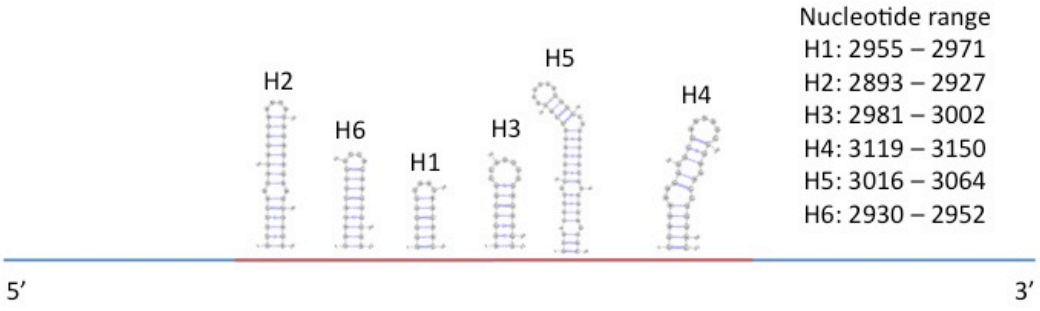


Figure B2: Relative location of each of the six motifs predicted by Mfold for the 500 nucleotide mouse Cyrano lincRNA. Blue corresponds to the 500-nucleotide RNA, and red corresponds to the 300-nucleotide conserved region.

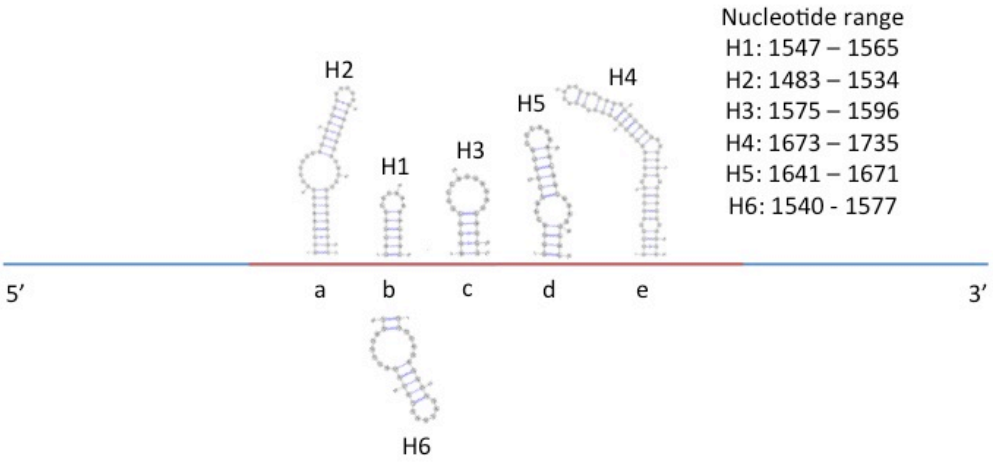


Figure B3: Relative location of each of the seven motifs predicted by Mfold for the 500 nucleotide zebrafish Cyrano lincRNA. Blue corresponds to the 500-nucleotide RNA, and red corresponds to the 300-nucleotide conserved region. In position b, where two motifs are present, 50% of the structures predicted the top motif, while the other 50% of the structures predicted the lower motif for that region of the sequence.

Table B1: Species containing the Cyrano lincRNA recovered from BLAST search used in alignments for comparative sequence analysis.

Cyrano Species
Physeter catodon *Pteropus vampyrus* *Gallus gallus* *Aquila chrysaetos* *Equus przewalskii* *Balaenoptera acutorostrata scammoni Sus scrofa* *Equus caballus* *Struthio camelus australis* *Bos taurus* *Bubalus bubalis* *Microcebus murinus* *Tarsius syrichta* *Mustela putorius furo* *Pathera tigris altaica* *Felis catus* *Oryctolagus cuniculus* *Myotis brandtii* *Haliaeetus albicilla* *Pygoscelis adeliae* *Corvus cornix* *Myotis davidii* *Aptenodytes forsteri* *Columba livia* *Cricetulus griseus* *Haliaeetus leucocephalus* *Microtus ochrogaster* *Mus musculus* *Rattus norvegicus* *Pelodiscus sinensis* *Chrysemys picta bellii* *Loxodonta africana* *Meleagris gallopavo* *Anas platyrhynchos* *Bison bison* *Zonotrichia albicollis* *Anolis carolinensis* *Pundamilia nyererei* *Poecilia reticulata* *Stegastes partitus* *Poecilia formosa* *Danio rerio* *Astyanax mexicanus* *Homo sapiens* *Pan troglodytes* *Pan paniscus* *Pongo abelii* *Cercocebus atys* *Papio anubis* *Callithrix jaccus* *Aotus nancymaae*

Python Scripts

Multiple python scripts were written to assist with basic tasks such as primer design and pulling out data from CMfinder and WAR server generated files. These are listed below, with what they do commented by a '#':

```
# to pull out the motif from CMfinder files and put them in a new file
from sys import argv
from os.path import exists

script, from_file, to_file = argv

myString = "#=GC SS_cons"
infile = open(from_file, 'r')
for line in (infile):
        if myString in line:
                outfile = open(to_file, 'w+')
                outfile.write(line)

# to replace thymine with uracil in a sequence
str = '''
'''



new_string = str.replace('T','U')
print new_string

# to print the reverse complement; useful for primer analysis
```

```
primer_sequence = ''
def reverse_complement(primer_sequence):
    sequence_dictionary = {'A':'T','T':'A','G':'C','C':'G'}
    return "".join([sequence_dictionary[base] for base in (primer_sequence)[::-1]])

print reverse_complement(primer_sequence)
print len(primer_sequence)

# to take an array of Stockholm format and obtain a percentage of occurrence for each position (useful for WAR server results)
arr = [
 '']

for index in range(0, 12):
 lpar_counter = 0.0
 rpar_counter = 0.0
 dot_counter = 0.0
 total = 0.0

 for string in arr:
   if string[index] == '.':
     dot_counter += 1
   elif string[index] == '(':
     lpar_counter += 1
   else:
     rpar_counter += 1
   total += 1
```

```
 print 'at index {} it is {}% dots, {}% left parentheses and {}% right parentheses'.format(
   index,
   dot_counter * 100 / total,
   lpar_counter * 100 / total,
   rpar_counter * 100 / total)

# to pull out a conserved region in a file and print it along with flanking bases
from sys import argv

script, filename = argv

with open(filename) as f:
        for line in f:
                if 'CAACAAAATCA' in line:
                        string = line
                        substring = string.find('CAACAAAATCA')
                        conserved_and_flanking = string[substring-50:substring + len('substring') +50]
print conserved_and_flanking

# to keep track of your productivity
def start():
        print "Hi, Jonesy."
        print "On a scale of * to *****, how productive are you being?"
```

```
        productivity = raw_input("> ")


        if productivity == "*":
                exit_programs()
        elif productivity == "**":
                exit_programs()
        elif productivity == "***":
                print "You're on the right track. Get back to work. Find your groove!"
        elif productivity in ("****", "*****"):
                print "Yay! Keep up the good work. I won't annoy you anymore."
        else:
                print "Jonesy. What?"


start()
```