The Role of Topological Constraints in RNA Tertiary Folding and Dynamics

by

Anthony M. Mustoe

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in The University of Michigan
2014

Doctoral Committee:

Professor Hashim M. Al-Hashimi, Co-Chair
Professor Charles L. Brooks III, Co-Chair
Professor Carol A. Fierke
Professor David S. Sept

2014

Acknowledgements

It goes without saying that this dissertation would not exist if it were not for my extraordinary advisors, Professors Charles Brooks and Hashim Al-Hashimi. I am incredibly lucky to have had the opportunity to work under them both, and cannot thank them enough for their dedicated mentorship and friendship throughout these past five years. They gave me the freedom to pursue my interests and make mistakes along the way, and pushed me to accomplish more than I thought possible.

I am also indebted to both of my committee members, Professors Carol Fierke and David Sept. Carol showed great faith to green-light our mitochondrial tRNA collaboration, and then after committing her students to make the tRNAs on my behalf, allow me, a theorist, to perform the melting experiments in her lab. These experiments are a cornerstone of this thesis, and the opportunity to perform them myself has been greatly rewarding. David's class on protein design that I took as an undergraduate at WashU was one of my primary inspirations to pursue a Ph.D. in Biophysics, and I am grateful for his continued mentorship.

I would additionally like to thank past and present members of the Al-Hashimi, Brooks, and Fierke groups, who have provided help and numerous insights throughout the years. Dr. Xin Liu, Yu Chen, Dr. Heidi Alvey, Dr. Shan Yang, Dr. Max Bailor, Dr. Scott Horowitz, Dr. Joe Yesselman, Dr. Alex Dickson, Dr. Loic Salmon, Dr. Sean Law, Dr. Jeet Chugh, Dr. Catherine Eichhorn, Dr. Liz Dethoff, Isaac Kimsey, Dr. Karunesh Arora, Dr. Aaron Frank, Kate Dyki, and David Braun in particular have contributed significantly to my success.

Finally, I would like to thank my family and friends for their unconditional support and encouragement. I am incredibly lucky to have such loving parents who can also give me advice on how to make it through a Ph.D. To my wife Alison, who bore the daily brunt, this thesis is just as much your accomplishment as it is mine. Thank you for everything.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Functional RNA molecules must fold into highly complex three-dimensional (3D) structures and undergo precise structural dynamics in order to carry out their biological functions. However, the principles that govern RNA 3D folding and dynamics remain poorly understood. Recent studies have proposed that topological constraints arising from the basic connectivity and steric properties of RNA secondary structure strongly confine the 3D conformation of RNA junctions and thus may contribute to the specificity of RNA 3D folding and dynamics. Herein, this hypothesis is explored in quantitative detail using a combination of computational heuristic models and the specially developed coarse-grained molecular dynamics model TOPRNA.

First, studies of two-way junctions provide new insight into the significance and mechanism of action of topological constraints. It is demonstrated that topological constraints explain the directionality and amplitude of bulge-induced bends, and that long-range tertiary interactions can modify topological constraints by disrupting non-canonical pairing in internal loops. Furthermore, topological constraints are shown to define free energy landscapes that coincide with the distribution of bulge conformations in structural databases and reproduce solution NMR measurements made on bulges.

Next, TOPRNA is used to investigate the contributions of topological constraints to tRNA folding and dynamics. Topological constraints strongly constrain tRNA 3D conformation and notably discriminate against formation of non-native tertiary contacts, providing a sequence-independent source of folding specificity. Furthermore, topological constraints are observed to

give rise to thermodynamic cooperativity between distinct tRNA tertiary interactions and encode functionally important 3D dynamics. Mutant tRNAs with unnatural secondary structures are shown to lack these favorable characteristics, suggesting that topological constraints underlie the evolutionary conservation of tRNA secondary structure. Additional studies of a non-canonical mitochondrial tRNA show that increased topological constraints can reduce the entropic cost of tertiary folding, and that disruptions of topological constraints explain the pathogenicity of a insertion mutation in this tRNA. UV melting experiments verify these findings.

Finally, TOPRNA is used to study the topological constraints of the 197 nucleotide *Azoarcus* Group I ribozyme. It is shown that topological constraints strongly confine this RNA and provide a mechanism for encoding tertiary structure specificity and cooperative hierarchical folding behavior.

Chapter 1: Introduction

Portions of this chapter have been adapted from the following publications:

Bailor, M. H., Mustoe, A. M., Brooks, C. L. III, and Al-Hashimi, H. M. (2011) Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Current Opinion in Structural Biology*, **21**, 296-305.

Bailor, M. H., Mustoe, A. M., Brooks, C. L. III, and Al-Hashimi, H. M. (2011) 3D maps of RNA interhelical junctions. *Nature Protocols*, **6**, 1536-1545.

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

Mustoe, A. M., Brooks, C. L. III, and Al-Hashimi, H. M. (2014) Hierarchy of RNA functional dynamics. *Annual Review of Biochemistry*, **83**, 441-466.

1.1 The Role of RNA 3D Structure and Dynamics in Cellular Function

1.1.1 RNA Biology

Composed of only four chemically similar nucleotides, RNA was long thought to lack the chemical and structural complexity needed to drive biochemical processes that power living cells, limited instead to a role as a rudimentary messenger. However, discoveries in molecular biology over the past three decades have shown that nothing could be further from the truth. Groundbreaking studies by Cech and Altman in the 1980s showed that RNA is capable of catalytic activity (1,2). Entire new classes of RNA such as microRNAs and long-non-coding RNAs (lncRNA), among many others, have been discovered and shown to play central roles in

gene regulation (3-5), and their dysfunction has been implicated in many human diseases (6,7). Sophisticated ribonucleoprotein complexes such as the spliceosome and ribosome, which were previously thought to derive their functionality from their protein components, have since been shown to be primarily RNA machines (8-11). Even classic RNAs such as transfer, messenger, and viral RNAs play surprisingly complex roles in protein synthesis (12-15). In a culmination of these studies, it is now believed that seventy-five percent of the human genome codes for RNA, the function of most of which we are only beginning to uncover, whereas less than 2% code for proteins (16,17).

1.1.2 Function RNA 3D Structure

For most RNAs, function is intimately related to and dependent on an ability to fold into a highly precise three-dimensional (3D) structure that is encoded by the RNA sequence (18). For example, early biophysical studies of transfer RNA (tRNA) molecules revealed that they universally fold to an 'L'-like structure that rivals the complexity of folds adopted by proteins (**Figure 1.1**) (19-21). This 3D fold is exactly complementary to the different tRNA binding sites in the ribosome, with the L-shape allowing tRNA to recognize a cognate mRNA codon at one end while simultaneously allowing the chemically-linked amino acid at the other end to interact with the ribosomal peptidyl transferase center >70 Å away (12,22). Other aspects of tRNA 3D structure, such as the specific arrangement of 'identity' nucleotides throughout the tRNA body, allow aminoacyl synthetase proteins to recognize tRNAs in a species-specific manner and hence ensure the fidelity of amino-acid charging (22). The highly conserved 'elbow' of the L also serves as an important recognition site for other tRNA processing enzymes and regulatory

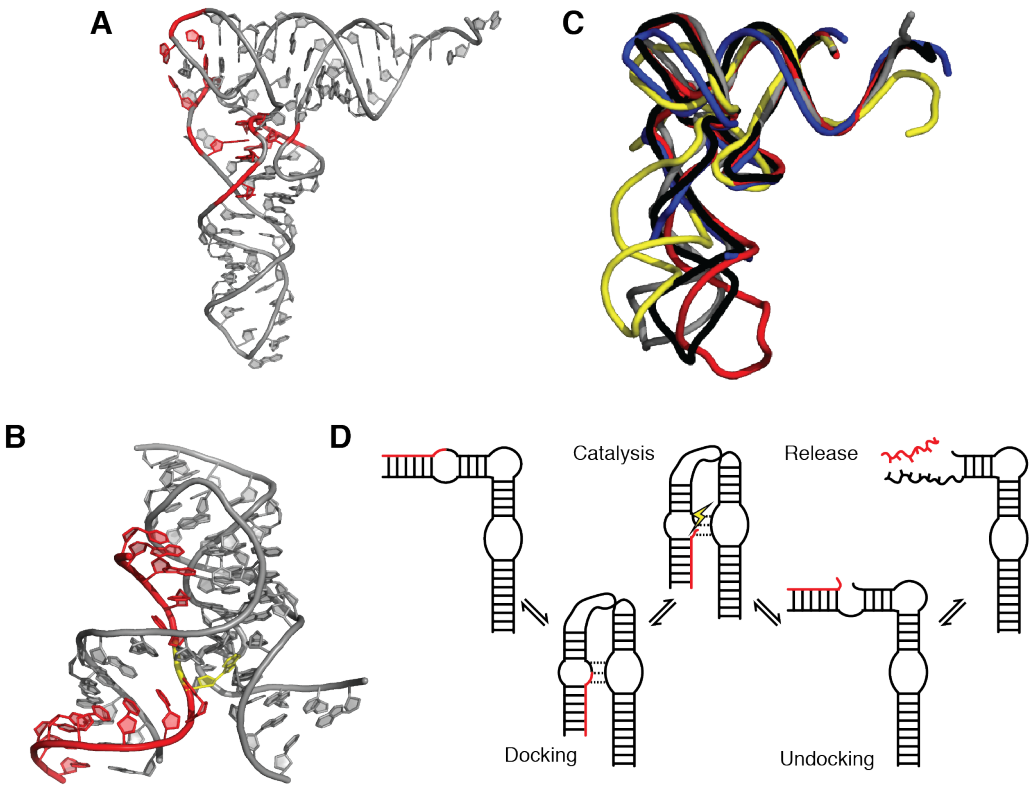

Figure 1.1: Functional RNA 3D structure and dynamics
(A) Crystal structure of tRNA (PDB 6TNA). Conserved long-range tertiary base pairs are shown in red. **(B)** Crystal structure of the hairpin ribozyme (PDB 1M5K). Cleaved strand is shown in red, with cleavage site shown in yellow. **(C)** Different X-ray structures of $tRNA^{Phe}$ in the unbound state (black, PDB 1EHZ), in complex with RNaseP (blue, engineered anticodon stem removed, PDB 3Q1Q), the ribosome in the P/E state (gray, PDB 3R8N), isopentenyl-tRNA transferase (red, PDB 3FOZ), and phenyalanyl-tRNA synthetase (yellow, PDB 1EIY). The structures are superimposed by the acceptor stem. **(D)** Enzymatic cycle of the hairpin ribozyme.

molecules (23,24). Mutations to the tRNA sequence that disrupt these structural features severely disrupt biological function (25,26).

Crystal structures and mechanistic studies have similarly revealed an intimate relationship between 3D structure and function in other RNAs (18,27). Catalytic RNA molecules, otherwise known as ribozymes, adopt complex folds that possess highly specific binding pockets that recognize RNA substrates with high affinity (**Figure 1.1**) (24,28-31). The bound substrate is in turn precisely aligned for catalysis by networks of hydrogen bonds, base-stacking interactions, and coordinated metal ions. So-called riboswitches, which are regulatory

RNAs that specifically bind small-molecule ligands, adopt folds with binding sites that completely encapsulate the target molecule (32,33). Other RNAs serve as scaffolds, with function arising from an ability to organize bound proteins in specific 3D arrangements (3,34). While the structure-function relationship is still poorly understood for many other RNAs such as lncRNAs (35), molecular biology dogma holds that 3D structure will eventually be shown to be central to the function of these molecules as well.

1.1.3 Function of RNA Dynamics

More recently, studies have also begun to emphasize the importance of structural dynamics to RNA biological function (36-39). RNAs possess an incredible ability to undergo complex structural changes in response to a wide range of cellular cues such as binding by proteins or other RNAs (40,41), small-molecule ligands (42,43), and metals (44), and changes in temperature (45) and pH (38,46). Again using tRNA as an illustrative example, microscopic changes in backbone orientation and local hydrogen bonding, as well as macroscopic reorientations of entire helical domains, allow tRNA to optimize intermolecular interactions with its diverse cellular partners (**Figure 1.1**) (47). The low energy barriers to such conformational changes are particularly important for allowing tRNA to efficiently traverse through the confines of the ribosome during translocation between the A-, P-, and E- ribosomal binding sites (48-50). In some cases, cofactor-induced structural changes can serve as transducers, triggering additional functional events. Notably, successive changes in inter-helical orientations induced by protein binding are thought to help order the assembly of complex ribonucleoprotein (RNP) machines, including the 30S ribosome (51-53), the signal recognition particle (54), and telomerase (55). Another important functional role of RNA 3D dynamics includes facilitating rapid substrate

turnover in ribozymes (56-58); a common mechanism utilizes an 'undocked' conformation to efficiently bind substrate, which is followed by a transition to a docked conformation where the substrate is aligned for catalysis in the active site, and finally a transition back to undocked state to mediate product release (**Figure 1.1**). Lastly, while not a topic of this thesis, it should be noted that large-scale changes in secondary structure provide a particularly rich source of RNA functional complexity. For example, riboswitches use secondary structure changes effected by small-molecule binding or other environmental factors to sequester or reveal regulatory sequence elements (59).

Much as RNA function depends a specific 3D fold, specificity is a critical feature of most functional RNA dynamics. This ensures, for example, that RNAs are able to adaptively bind to their correct cellular targets, yet not indiscriminately bind off-target cofactors (36). For RNAs such as ribozymes that cycle through multiple conformations during their functional lifecycle, dynamic specificity reduces the conformational space that must be searched between each transition, accelerating kinetics (60,61). Perhaps most importantly, dynamic specificity provides the foundation of RNA switching activity (59), ensuring that environmental effectors robustly generate the desired structural change.

The discussion above provides only a small sampling of the remarkable structural and dynamic diversity exhibited by RNA molecules. Nevertheless, the universal theme is that RNA biological function is intimately related to specific 3D structure and dynamics.

1.2 The RNA Free Energy Landscape

1.2.1 Overview

Understanding the physical principles that govern RNA folding and dynamics has been a long-standing challenge in biophysics. Free energy landscape theory, which was first developed to describe the folding and dynamics of proteins by Frauenfelder, Sligar, and Wolynes (62,63) is equally applicable to RNA and provides a powerful common language for describing how sequence encodes RNA behavior. In solution, a given RNA does not fold into a single structure, but rather forms a statistical distribution, or ensemble, of many inter-converting conformations. This statistical ensemble can be described in terms of a continuous free energy landscape, which specifies the free energy of every atomic configuration. The population of each configuration depends on its free energy, whereas the rates of inter-conversion between individual configurations depend on the height of the separating barriers.

Although the free energy landscape can in principle be arbitrarily complex, in RNA it is hierarchically organized into local energetic minima containing conformational sub-states (CS) that are separated by large barriers, each of which is in turn sub-divided into a larger number of local energetic minima that are separated by lower barriers, and so forth (**Figure 1.2**) (46,64,65). These hierarchically organized energetic layers form different "Tiers" (Tier 0, Tier 1, etc…), and RNA folding and dynamics can in turn be hierarchically organized in terms of transitions between CSs within different tiers.

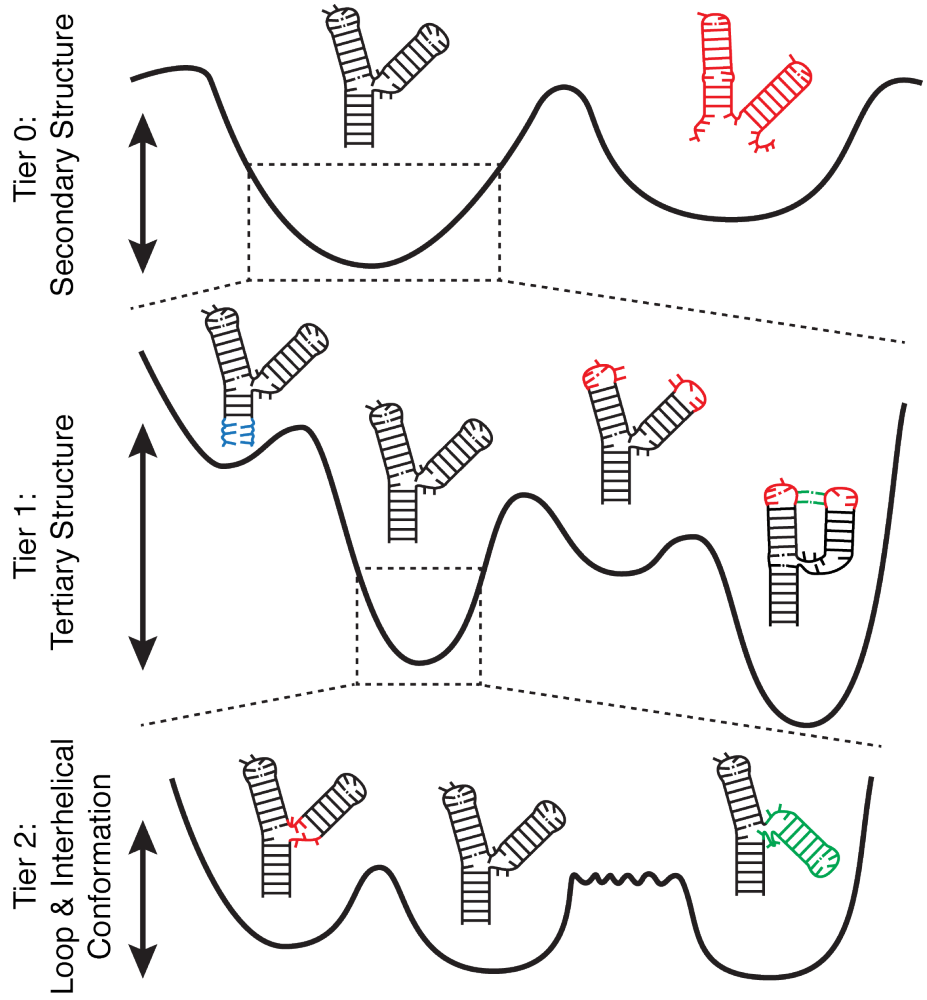

Figure 1.2: The different tiers of the RNA free energy landscape
Tier 0 defines broad secondary structure free-energy basins with high separating barriers. Tier 1 defines narrower free energy basins associated with alternative tertiary base-pairing arrangements of a given secondary structure. These include differences from base pair melting (blue, left), reshuffling (right, red), and tertiary pairing (far right, green). Tier 2 defines shallow free-energy basins associated with specific loop and interhelical conformations. These dynamics include loop dynamics (left, red) and interhelical dynamics (right, green). Although barrier heights between different loop and interhelical conformations are similar, interhelical dynamics typically occur more slowly due to the larger number of involved coordinates (long rough separating barrier).

1.2.2 Tier 0: Secondary structure folding and dynamics

The lowest tier of the RNA free energy landscape dictates secondary structure folding and dynamics (**Figure 1.2**). Driven by strong favorable base stacking interactions and cross-strand hydrogen bonding, RNA molecules have an overwhelming propensity to fold back on

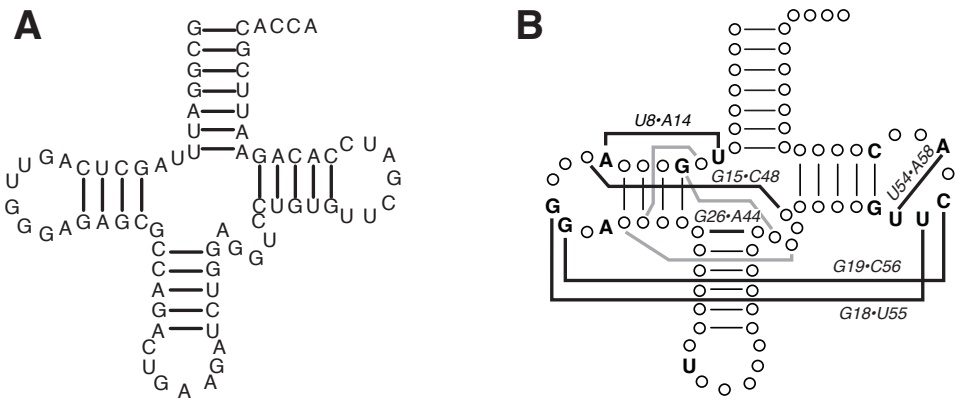

Figure 1.3: RNA secondary structure
(A) Secondary structure diagram of yeast $tRNA^{Phe}$. **(B)** Conservation of tRNA sequence across all organisms and amino acid species projected onto the secondary structure. Residues conserved in <90% of tRNA species are indicated by circles (72). Conserved long-range tertiary interactions are drawn by bold lines and labeled. Semi-conserved base-triples are drawn as gray lines.

---

themselves to form highly stable A-form helical elements comprised of Watson-Crick and G•U base pairs (66-68) (**Figure 1.3**). Each pair can be worth as much as -3.5 kcal/mol of free energy, and extended helices can have melting temperatures well above 80° C (67). Together, the collection of helical elements and remaining single stranded regions define the secondary structure of an RNA molecule.

Decades of studies have demonstrated that RNA secondary structure is independently stable and folds before higher-order 3D structure (64,65). To a first approximation, both the equilibrium conformation and stability of secondary structure can be predicted from primary sequence using simple additive thermodynamic models that take into account the differential stability of base pairs (G•C > A•U > G•U) and differences in the stacking energy between neighboring base pairs in a helix (69). As is discussed throughout this thesis, secondary structure plays a profound role in shaping the higher levels of the RNA free energy landscape and for this reason is typically the most highly evolutionary conserved aspect of RNA molecules (70,71). Notably, the high conservation of secondary structure contrasts with the low conservation of primary sequence; mutations of base-paired residues are readily tolerable as long they are offset

by compensatory mutations to the pairing partner. tRNA provides a classic example of an RNA that possesses little primary sequence conservation yet universal secondary structure conservation (**Figure 1.3**).

While most functional RNAs must fold to a single well-defined secondary structure in order to function, the free energy landscape is both complex and rough. Due to the inherent degeneracies of the energetics of base-pairing, many RNA molecules also misfold to alternative secondary structures (73-75). The stability of such misfolded helices in turn presents a large, sometimes-insurmountable barrier to refolding to the native secondary structure, and cells have evolved specialized chaperone proteins to help catalyze such secondary structure refolding (75,76). Thus, secondary structure folding plays a critical role in understanding the overall mechanism of RNA folding and remains an extremely active area of RNA research (77-79). As noted above, dynamics between alternative secondary structures also provide the foundation for many important RNA functions. Nevertheless, the remainder of this thesis will focus on the higher levels of RNA free energy landscape, assuming the existence of single prefolded native secondary structure.

1.2.3 Tier 1: Tertiary structure folding and dynamics

Within the free energy basin defined by global secondary structure, a diverse array of attractive tertiary interactions serves to stabilize specific 3D conformations of RNA helices, which together define the second tier of the RNA free energy landscape and govern 3D folding (**Figure 1.2**). The strongest of these interactions are the non-canonical pairing interactions (e.g. G•A pairs (80,81)) that occur between internal loop residues held in close proximity by neighboring canonical base pairs, as well as long-range canonical and non-canonical pairings between distal

loops (82). Other important interactions include long-range base-stacking interactions, favorable interactions between the RNA backbone and tightly bound cations, and weaker base-to-backbone hydrogen bonding interactions such as those found in A-minor motifs (27,82-84).

Apart from local non-canonical pairing in internal loops, individual tertiary interactions are typically less stable (<-2 kcal/mol) than secondary structure base pairs, and hence complex networks of multiple interactions are needed to stabilize RNA folds (82). The classic example of such a network can be found in tRNA, which for over 20 years remained the only example of a known RNA tertiary fold (21). Highly conserved canonical and non-canonical pairs within and between the D- and T- loops, base triples between the A/D- and V-loops and the D-stem, and a multitude of long-range stacking interactions are critical for maintaining the characteristic tRNA 'L'-shape (**Figure 1.1, 1.3**).

The past decade has seen a veritable explosion in the number of known RNA tertiary structures (27,85), revealing many insights into the interactions that stabilize RNA 3D structure. However, despite the growing catalogue of known tertiary interactions, the rules that govern tertiary folding remain poorly understood. While well conserved in tRNA, it has been much more difficult to identify conserved tertiary interactions in other RNAs (70,86). Most tertiary interactions are promiscuous, particularly those such as A-minor motifs that occur between the RNA backbone (87). The few tertiary interaction motifs that are specific, such as tetraloop/tetraloop-receptors, are reused multiple times in the same molecule (88). Furthermore, poorly understood cooperativity between tertiary interactions as much as 40 Å apart is responsible for a large proportion of the free energy of tertiary folding in many RNAs (89,90). Molecular crowding agents that indiscriminately promote RNA compaction have also been shown to stabilize native 3D structure, a somewhat surprising result given that such compaction

could naively be expected to also stabilize misfolds (91,92). Together, these facts pose the currently unanswered question: what is the source of tertiary structure folding specificity?

An essential feature of RNA tertiary folding is its strong dependence on solution cation concentration. This dependence is due to the negative charge of the RNA phosphate backbone, which generates strong electrostatic repulsions that disfavor close packing of helices. Folding requires the uptake of Na^+, K^+, and particularly Mg^{2+} ions to neutralize these repulsions, and this uptake poses a strong entropic penalty on folding when these ions are low abundance (93-95). While it was originally thought that ions were site-specifically bound by tertiary motifs, more recent studies have instead argued that in most cases ions diffusely associate with the RNA, forming an 'ion atmosphere' (95,96). The general success of mean-field Poisson-Boltzmann calculations in describing the salt dependence of RNA folding provides strong support for such a diffuse ion model (94,97), although *bona fide* examples of specifically coordinated ions exist (98).

While theoretical studies suggest that electrostatic forces between helices are significantly screened at the high salt concentrations needed for folding (97,99), the salt-dependence of RNA tertiary structure stability has provided the principle means for investigating Tier 1 of the energy landscape. Generally, experiments are initiated from a low (non-physiological) salt condition where the majority of secondary structure is prefolded and equilibrium tertiary folding is monitored as the salt concentration is increased (64,100,101). Alternatively, folding kinetics are monitored after a rapid jump in salt concentration. These and related studies have shown that charge neutralization induces RNA to rapidly and cooperatively collapse from a disordered ensemble of extended conformations to a native-like ensemble of compact intermediates. While tertiary structure is incompletely formed, the compact

intermediates are stabilized by native interactions (89). Interestingly, despite having substantial native character, a large fraction of the compact ensemble can consist of misfolded species, which is thought to be due to topological frustration or local mispairing of secondary structure (102). For the set of native compact intermediates, further increases in salt concentration leads to slow consolidation to the final 3D conformation where all native tertiary interactions are fully formed (100). While the atomic details of this consolidation process are unclear, particularly in large RNAs, it is thought to involve reorganization of non-specific tertiary interactions such as non-canonical pairs and interactions between tightly associated ions. Non-native intermediates also eventually fold at higher salt concentrations, although on much slower timescales.

A key theme arising from the RNA folding literature described above is that RNA molecules possessing fully native secondary structure generally possess native-like tertiary structure. In other words, it is rare to observe a 'tertiary-misfolded' RNA, even among compact intermediates that lack fully formed tertiary interactions. Thus, returning to the original question of what is the source of tertiary structure folding specificity, this observation suggests that specificity may be encoded at the secondary structure level. Exploring this hypothesis is a central aim of this dissertation.

In addition to encoding 3D structure, Tier 1 also plays an important role in defining tertiary structure dynamics that are important to many biological mechanisms. Examples of such dynamics include base-pair melting fluctuations, where base pairs transiently break apart (103), and base-pair isomerizations, where non-canonical and G•U base-pairs interconvert between alternative hydrogen bonding arrangements (81,104,105). Base pairs in internal loop motifs can also 'reshuffle', locally rearranging their pairing partners (106). Such local changes in base-pairing geometry and arrangement are typically destabilized by <3 kcal/mol relative to the most

favorable state, and thus they are readily populated. As explored in detail in Chapter 2, these conformational changes play important roles in shaping Tier 2 dynamics. On longer timescales, cooperative melting and reforming of tertiary interactions allow RNAs to transition between different functional conformations. A notable example is the 'docking' and 'undocking' transitions undergone by ribozymes during their catalytic lifecyles, where the 'tertiary unfolded' undocked state is only 1-2 kcal/mol higher in energy (61,107).

1.2.4 Tier 2: Interhelical and loop conformation and dynamics

Even after RNA has formed local non-canonical pairing and tertiary interactions, it still retains considerable structural flexibility (**Figure 1.2**). Unpaired residues in bulge motifs and other loops are able to change conformation with minimal energetic penalties (108), and helical domains linked by such single-stranded motifs can correspondingly undergo large-scale reorientations with respect to one another (109). On a microscopic scale, the sugar-phosphate backbone can repucker and reorient (110,111). Together, these various degrees of freedom define the relatively flat Tier 2 of the RNA free energy landscape.

The primary energetic determinants of the Tier 2 landscape are the favorable base-stacking interactions that occur between neighboring secondary structure helices and single-stranded residues. The strength of these interactions is highly context-dependent, with purine-purine stacks much stronger than pyrimidine-pyrimidine stacks (67). However, these interactions provide at most -3 kcal/mol and more typically <-1 kcal/mol in stabilizing energy (112,113), and thus unstacking and other rearrangements in loop conformation occur facilely. In turn, the dynamism of single-stranded loops allows helical domains to reorient with little to no energetic penalty. These so-called interhelical dynamics have been extensively characterized by the Al-

Hashimi lab in the HIV-1 trans-activation response element (TAR) model system, which consists of two helices linked by a 3-nt single-stranded bulge motif. A variety of NMR and combined NMR and molecular dynamics (MD) studies have revealed that a superposition of slower stacking and unstacking transitions on μs timescales and faster ns motions within a given 'stacked' energy basin give rise to 0°-90° changes in the interhelical bend (108,109,114). Similar large changes in interhelical conformation have been observed in many other RNAs by a variety of experimental methods (115-123). As illustrated by the different conformations adopted by tRNA in **Figure 1.1**, such interhelical reorientations can also occur in the context of a well-folded tertiary structure.

Despite the flatness of Tier 2 of the energy landscape, the structural dynamics that occur on this level of the hierarchy are highly specific (39,124). This has extremely important consequences for folding and for functions that harness Tier 2 dynamics. The source of this specificity is a central topic of this thesis and is discussed in greater detail in **Section 1.3**.

1.2.5 Coupling between Tiers

It is important to note that the above hierarchical description of the RNA free energy landscape is a simplification. For example, studies have shown that favorable tertiary interactions on Tier 1 can stabilize an otherwise unstable secondary structure (125-127). The tertiary structure stability of certain well-folded molecules like tRNA can also approach or exceed that of individual helical elements at high salt concentrations (128). Furthermore, when folding is initiated from a conformation lacking prefolded secondary structure, tertiary interactions often form concomitantly with native secondary structure and in turn bias RNA folding along specific pathways (77,129-131). Such behavior is likely more representative of RNA folding *in vivo*,

where RNAs fold co-transcriptionally as the nascent chain emerges from the RNA polymerase (127). Nevertheless, the hierarchical view of RNA folding has been well validated as a general model and provides an essential foundation for understanding the complexity of RNA folding and dynamics.

1.3 RNA junctions and topological constraints

1.3.1 RNA junctions as important determinants of RNA 3D conformation

Due to the clear importance of long-range tertiary interactions and electrostatics for RNA stability, most studies of RNA have focused on understanding how these forces shape Tier 1 of the free energy landscape. However, it is also known that secondary structure plays a powerful role in determining RNA 3D conformation. At the most basic level, secondary structure specifies the local helical structure of paired nucleotides. Global RNA structure is in turn largely dictated by the relative 3D orientation of these helical elements. Thus, the key link between secondary and 3D structure is the conformation of so-called interhelical junctions, which are the motifs of single-stranded residues or non-canonical base-pairs that link two or more helices (**Figure 1.4**) (46,132). A growing body of research has suggested that junctions inherit strong conformational preferences based on their number of helices and single-stranded residues, and that these preferences influence 3D folding and dynamics.

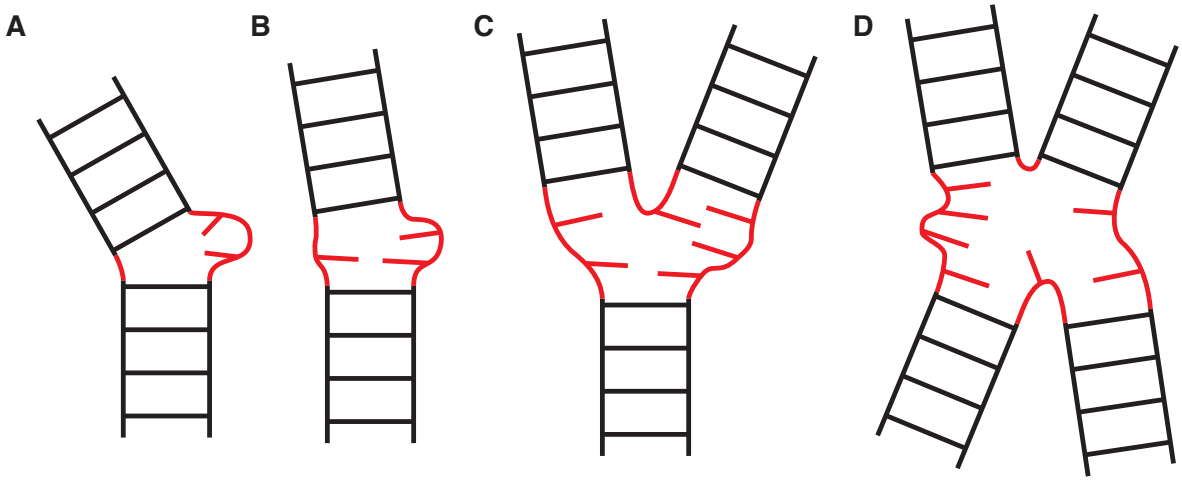

Figure 1.4: Cartoons of different RNA junctions
Junction linkages are shown in red for **(A)** two-way bulge, **(B)** two-way internal loop, **(C)** three-way, and **(D)** four-way junctions. Red straight lines indicate single-stranded or non-canonical paired residues, and black lines indicate flanking canonical base-paired helices.

The ability of junctions to influence RNA 3D conformation has been appreciated from the early 1990s. Using electropheretic gel mobility measurements, fluorescence resonance energy transfer, transient electric birefringence, and electron microscopy, Lilley (133,134), Griffith (135), Draper (136) and Hagerman (137,138) showed that bulges and internal loops induce directional bends in DNA and RNA duplexes by amounts dependent on the length, asymmetry and sequence of the junction. Longer bulges induced greater bends but only up to a certain length of 5–7 nucleotides. Similarly, studies of isolated three and four-way junctions indicated that these motifs adopt only a limited set of conformations, often involving coaxial stacking of one or two pairs of helices (139). Evidence of the importance of junctions to 3D folding was also hinted at by their strong evolutionary conservation. For example, whereas apical loops that cap helices generally vary in length across different species of the same class of RNA, the length of single-stranded loops that link helices are universally conserved (140,141).

These early observations have since been corroborated by the growing database of RNA X-ray structures. Surveys by Lescoute and Westhof (142) demonstrated that three-way junctions

generally adopt only one of three different classes of folded conformations, with class significantly determined by the length of single-stranded nucleotides within the junction. Similar surveys of four-way and higher-order junctions observed correlations between folded conformation and secondary structure (143,144).

The most compelling illustration of the role of junctions in RNA folding comes from a series of studies performed on the hairpin ribozyme, a paradigm system for studying RNA tertiary folding (145). In its simplest form, the hairpin ribozyme comprises two helices with embedded internal loops that are linked by a central two-way junction (**Figure 1.1**). When folded, the two internal loops form tertiary interactions that stabilize the helices in a parallel conformation and catalyze self-cleavage of one internal loop strand. However, other forms of the ribozyme exist, and in particular, the evolutionarily favored form of the ribozyme replaces the central two-way junction with a four-way junction. While the additional helices form no tertiary interactions, it was shown that the four-way junction stabilizes the folded conformation by -2 kcal/mol (107). Other modifications to the central junction, such as changing it to a three-way junction, or changing the length of the single-stranded linker in the two-way junction variant, also modulated folding stability by several kcal/mol (107,146). These changes in folding stability in turn correlated with changes in the catalytic rate of the ribozyme. Follow up studies later showed that the changes in stability were a result of changes in the entropic cost of folding (147), which in the four-way junction specifically increase the folding rate compared to the two-way junction (60).

From a different perspective, predicting junction conformation has proven to be a promising strategy for predicting global tertiary structure, and interest in this area has significantly increased with the development of methods that have made it relatively

straightforward to determine RNA secondary structure (14,148). An influential early study by Michael Levitt (19) used inferred interhelical stacking arrangements to arrive at a reasonable prediction of tRNA 3D structure. Later studies by Michel and Westhof (141) used similar principles to guide their remarkably accurate prediction of group I intron structure. More recently, a number of groups have had success using homology and knowledge-based approaches to predict the conformation of individual junctions and then using these predictions to infer global tertiary structure (149-154). Many others have demonstrated that combining knowledge of junction conformational biases with a limited number of inferred tertiary or experimental constraints is sufficient to accurately determine 3D structure (155-162).

Collectively, these studies strongly indicate that junctions play a potentially powerful role in determining RNA 3D conformation. However, a comprehensive understanding of the forces governing junction conformation and dynamics has remained as elusive as it has for tertiary structure. Experimental observations represent a convolution of multiple forces that has proven difficult to disentangle. These include electrostatics, attractive interactions between bases, potential non-canonical base-pairing interactions, as well as entropic contributions arising from the steric and connectivity properties of a given junction. Homology modeling methods, although successful, are ill-suited to provide a deeper understanding into the RNA free energy landscape.

1.3.2 Topological constraints

As alluded to above, the basic steric and connectivity properties of RNA junctions play important roles in governing 3D conformation even in the absence of attractive interactions and electrostatics. The single-stranded residues that tether helices together are short (<7 Å per nt) in comparison to the local diameter of the A-form helix (~17 Å across a base pair) (163). This gives

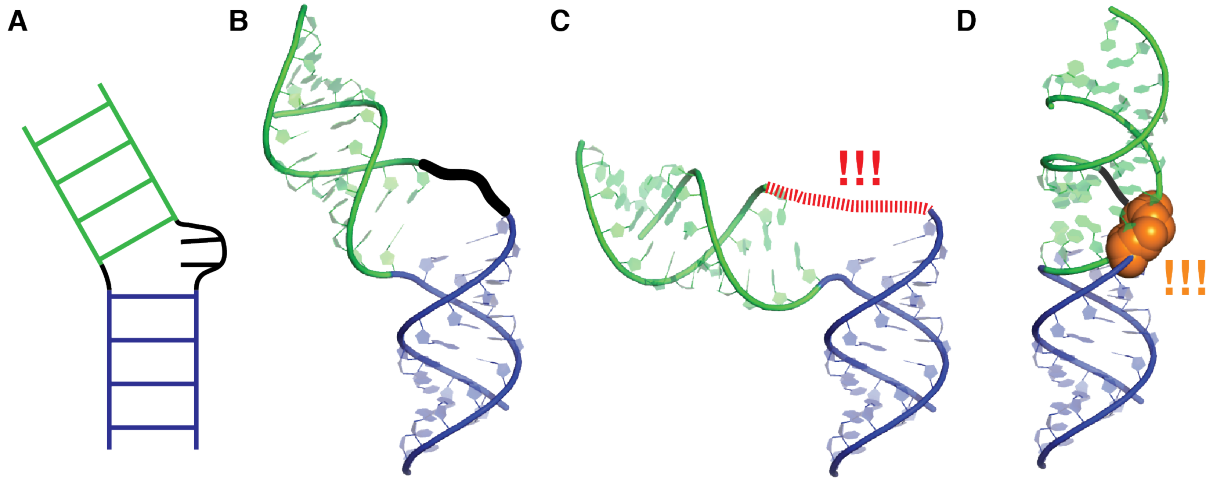

Figure 1.5: Illustration of topological constraints on a bulge
(A) Secondary structure cartoon of a 2-nt bulge motif. **(B)** A topologically allowed interhelical conformation that satisfies both the connectivity constraints of the bulge linker (black) and steric constraints of the helices. **(C)** An interhelical conformation that is disallowed due to connectivity constraints of the bulge (red). **(D)** An interhelical conformation that is disallowed due to steric constraints (orange).

---

rise to connectivity constraints – i.e. reorientation and translation of one helix relative to another is severely restricted by the short length of the tethers (**Figure 1.5**). The specific nature of these connectivity constraints will vary depending on the junction topology, including the number of helices in the junction (two-way, three-way, etc.), relative lengths of linking strands, and whether bases in junctions adopt a looped in or looped out conformation. In addition, steric constraints between RNA helices further restrict the allowed translation and reorientation of helices with respect to one another (**Figure 1.5**). These steric collisions can be long-range in nature, occurring between helices that are far apart in sequence. Together, these fundamental constraints comprise so-called topological constraints that strongly limit junction conformation.

The concept that topological constraints have a powerful influence on biomolecular structure has a long history, originating with the seminal description of the Ramachandran plot of allowed (φ,ϕ) backbone dihedral angles in proteins (164). Similar maps of allowed backbone dihedral conformations have been constructed for RNA (165,166). On a more fundamental level,

it has long been known from basic polymer physics that simple steric and connectivity properties of polymer chains give rise to forces that affect molecular conformation (167). However, application of these ideas on a macroscopic level to understand RNA junction conformation is only a recent development.

The first rigorous investigation of topological constraints in RNA junctions was prompted by studies of the two-way junction of HIV-1 TAR by the Al-Hashimi lab. Using both NMR (109) and a combined NMR-MD approach (168), they observed that at equilibrium TAR populates a wide range of interhelical conformations, characterized by 90° changes in interhelical bend and 60° changes in twist. Surprisingly, however, the distribution of populated conformations was highly anisotropic and directed along a very specific motional pathway. Studies of the related HIV-2 TAR, which replaces the 3-nt bulge found in HIV-1 with a shorter 2-nt bulge, revealed that it populates a similar although more limited distribution (168). To explore the generality of these observations, Bailor *et al.* (124) measured the interhelical conformations for all two-way junctions in the protein databank (PDB) using the approach outlined in **Section 1.4**. Strikingly, despite being derived from highly diverse structural and sequence contexts, the interhelical conformations observed among this collection of >3,500 junctions were limited to similar anisotropic distributions. Increasing the asymmetry of the internal loop resulted in an increase in the range of orientations sampled, whereas increasing the length of both loop strands resulted in a systematic shift in the interhelical twist by an amount comparable to the helical periodicity, assumed to be due to be non-canonical loop pairing.

To rationalize these observations, Bailor *et al.* (124) used a simple heuristic model to compute the set of interhelical orientations that are allowable given the topological constraints on junctions. These calculations revealed that topological constraints restrict two-way junctions to

highly anisotropic distributions of allowed states comprising only 4% of the total possible conformational space for 1-nt bulges, and growing to only 20% when the connectivity constraints are relaxed in 4-nt bulges. Impressively, despite containing such a limited number of conformations, the allowed distribution captured ~85% of the orientations observed in the PDB.

Concurrently, Chu *et al.* (169) used simulations of a model system consisting of two coarse-grained, idealized DNA helices adjoined by one or two polyethylene glycol (PEG) tethers to explore the effects that junction connectivity has on the 3D conformational landscape of RNA two-way junctions. The use of PEG ensured that sequence-specific interactions did not bias their results. Similar to the results of Bailor *et al.*, these simulations revealed that the PEG junction limits the set of interhelical conformations to a highly anisotropic distribution of allowed states. The approach taken by Chu *et al.* also allowed them to calculate free energy penalties imposed by the PEG linkers on different interhelical conformations, which ranged in magnitude by as much as 5 kcal/mol. Intriguingly, Chu *et al.* further demonstrated that these topological energy penalties could be used to discriminate against formation of non-native tertiary interactions between the helices, providing a potential strategy for encoding tertiary folding specificity in RNA.

Together, the results of Bailor *et al.* and Chu *et al.* pose the intriguing hypothesis that basic secondary-structure-encoded topological constraints play important roles in directing RNA folding and dynamics.

1.4 Quantifying RNA Interhelical Conformation

1.4.1 Motivation

In order to study the effects that topological constraints have on RNA structure, it is necessary to develop a unique geometric descriptor of RNA junction conformation. For example, the description of polypeptide backbones in terms of ϕ and ψ torsion angles was critical for the development of the Ramachandran map, and the (ϕ, ψ) description remains a central tool for analyzing protein structures (170-173). Local descriptors of nucleic acid structure have been developed for describing fine-grained helical and base pair parameters as well as sugar and phosphodiester torsions (163,174,175). However, there is a gap between atomic-level descriptors of local structure and descriptors that concisely relate the relative orientation and translation of secondary structure helices that comprise the fundamental building blocks of RNA global structure.

Traditionally, studies of RNA junctions have focused solely on the interhelical bend angle, which proves easier to measure without having to determine high-resolution structure (108,137,138,176-183). However, as helices are chiral objects, three angles are required to specify the orientation of one helix relative to another. The advent of weak alignment NMR and the measurement of residual dipolar couplings (RDCs) in partially aligned (184,185) RNA molecules provided a route for determining two, and in favorable cases all three, inter-helical angles without the need for determining a high resolution structure (109,124,186-189). This spurred the development and standardization of the Euler angle convention described below.

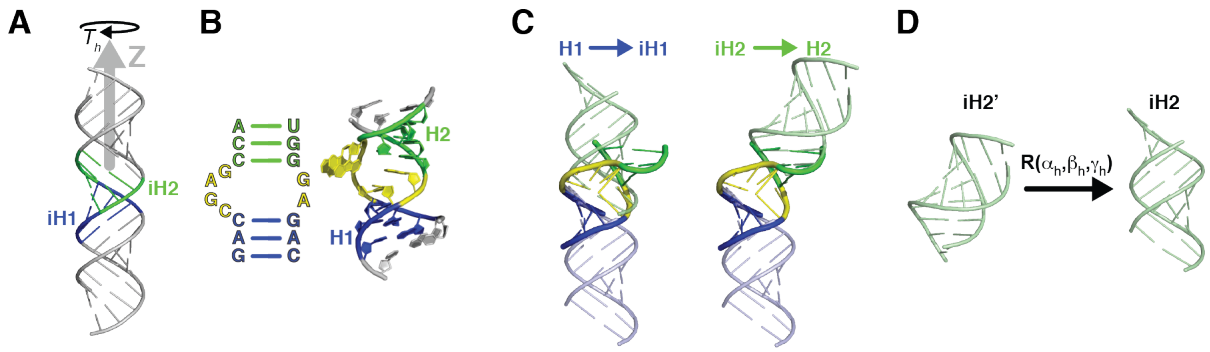

Figure 1.6: The (α_h, β_h, γ_h) convention
(A) The idealized reference. **(B)** An example internal loop junction from the PDB. **(C)** Alignment of the H1 and H2 helices to the reference. **(D)** Calculation of (α_h, β_h, γ_h) by aligning iH2' to iH2.

1.4.2 The (α_h, β_h, γ_h) convention

In order to describe the relative orientation of two helices across junction we rely on the fact that RNA helices overwhelmingly adopt standardized A-form structure (190); variations in structure observed across high-resolution X-ray structures are comparable to those expected due to thermal motions (163,191,192). This allows definition of a universal reference frame comprised of two coaxially stacked idealized A-form helices with the helix axis orientated along the positive Z-direction and twist angle $T_h = 51°$ between the junction-closing base pair of the lower helix (193) and Y-axis of the molecular frame (**Figure 1.6A**). We refer to the lower and upper helices as iH1 and iH2, respectively.

Using this reference frame, we are able to measure three Euler angles, which we term (α_h, β_h, γ_h), that describe the rotation needed to transform any two helices of a junction into an idealized stacked conformation (**Figure 1.6**) (124,189,194). In practice, this is performed by first superimposing one helix (H1) of the junction onto iH1. Next, a copy of iH2 is superimposed on the resulting H2 to generate iH2' (**Figure 1.6**). The rotation matrix **R** that brings iH2' into correspondence with iH2 is then determined and deconvoluted into Euler angles through:

$$R(\alpha_h\ \beta_h\ \gamma_h) = \begin{bmatrix} -\sin(\alpha_h)\sin(\gamma_h) + \cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) & -\sin(\alpha_h)\cos(\gamma_h) - \cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h) & \cos(\alpha_h)\sin(\beta_h) \\ \cos(\alpha_h)\sin(\gamma_h) + \sin(\alpha_h)\cos(\beta_h)\cos(\gamma_h) & \cos(\alpha_h)\cos(\gamma_h) - \sin(\alpha_h)\cos(\beta_h)\sin(\gamma_h) & \sin(\alpha_h)\sin(\beta_h) \\ -\cos(\gamma_h)\sin(\beta_h) & \sin(\gamma_h)\sin(\beta_h) & \cos(\beta_h) \end{bmatrix} \quad \textbf{[1.1]}$$

Equivalently, **R** can be computed as the rotation that brings H2 into correspondence with iH2, after H1 and iH1 have first been superimposed.

These angles can be intuitively related to 3D conformation by considering the inverse rotation, $\mathbf{R^{-1}}$, that transforms the reference frame (iH2) to a given orientation (iH2') (**Figure 1.6**). The inverse rotation is given by

$$R^{-1}(\alpha_h\ \beta_h\ \gamma_h) = R(-\gamma_h\ -\beta_h\ -\alpha_h). \quad \textbf{[1.2]}$$

Here, $-\gamma_h$, $-\beta_h$, and $-\alpha_h$ correspond to consecutive anti-clockwise rotations of H2 about the Z, Y, and Z axes, respectively (195). Equivalently, the angles γ_h, β_h, and α_h denote a clockwise rotation of H2 about its helical axis, followed by a rotation perpendicular to H1, followed by a rotation about the helical axis of H1. Thus, angles α_h and γ_h describe the relative twists of the two helices and β_h the interhelical bend.

1.4.3 Practical considerations when computing (α_h, β_h, γ_h) angles

To facilitate comparisons of (α_h, β_h, γ_h) angles between different junctions, it is necessary to consistently define the H1 and H2 helices. In two-way junctions, the two helices are typically tethered by two strands containing *X* and *Y* residues, where $X>Y$ (196). If the helices are tethered at only one strand then this is assigned as the 'Y' strand. Similarly, in higher-order junctions, one can assign one or two effective 'X' and 'Y' strands between helices, where the 'Y' strand has the fewest intervening helices and single-stranded residues. Given this definition of the strands, H1 is assigned as the helix that is linked to the 3' end of the 'Y' strand. H1 is arbitrarily assigned if $X=Y$.

Additional complications are associated with interpreting measured (α_h, β_h, γ_h) angles due to the non-linearity of the Euler angle space. First, Euler angles (α_h, β_h, γ_h) are degenerate with respect to ($\alpha_h \pm 180^o$, $-\beta_h$, $\gamma_h \pm 180^o$) and ($\alpha_h \pm 180^o$, $-\beta_h$, $\gamma_h \mp 180^o$). For the special case of having perfectly parallel or anti-parallel helices, there is a continuous coaxial degeneracy defined by ($\alpha_h \pm D$, $\beta_h = 0^o$ or 180^o, $\gamma_h \mp D$), where D is a constant. In our protocol, we lift these degeneracies by choosing the set of angles that have the minimal sum of squares, which serves to bias solutions to the pole and give the most compact distribution of Euler angles. Second, the L^2 distance between (α_h, β_h, γ_h) measured for two different junctions n and m does not provide an accurate measure of the similarity of their conformations. Instead, the difference in the conformations of two junctions is quantified as the amplitude, θ_{nm}, of the single axis rotation that transforms $H2_n$ into $H2_m$ following superposition of H1 (197,198). The relevant rotation, R_{nm}, which rotates $H2_n$ into $H2_m$ can be calculated through

$$R_{nm} = R_m^{-1} \cdot R_n = R\left(-\gamma_{h,m}, -\beta_{h,m}, -\alpha_{h,m}\right) \cdot R\left(\alpha_{h,n}, \beta_{h,n}, \gamma_{h,n}\right), \quad \textbf{[1.3]}$$

where ($\alpha_{h,n}$, $\beta_{h,n}$, $\gamma_{h,n}$) is interhelical Euler angle of junction n. θ_{nm} can then obtained from R_{nm} as

$$\cos\left(\theta_{nm}\right) = \tfrac{1}{2}\left[Trace\left(R_{nm}\right) - 1\right]. \quad \textbf{[1.4]}$$

1.5 Methods for studying topological constraints

1.5.1 Heuristic methods

The first studies of topological constraints by Bailor *et al* (124) relied on heuristic models to quantify the set of (α_h, β_h, γ_h) angles that were accessible to two-way junctions. These models were based on the assumption that helices possess only rotational degrees of freedom and were derived directly from the scheme used to measure interhelical Euler angles described in Section 1.4. Briefly, a phosphate on one strand was chosen as a stationary pivot and the inverse rotation

($-\gamma_h$, $-\beta_h$, $-\alpha_h$) was applied to the idealized reference helix iH2 for all possible combinations of (α_h, β_h, γ_h) on a 5° grid. Each generated interhelical conformation was then evaluated for steric collisions between the helices, and whether the distance between the helix termini opposite the pivot was less than the maximum length a bulge of *x* nucleotides could span. To model internal loops, the set of (α_h, β_h, γ_h) computed for bulges was shifted by -17°**y* in both α_h and γ_h, where *y* is length of the shorter strand in the internal loop. This shift was inferred based on the observation that (α_h, β_h, γ_h) angles in internal loops were systematically over-twisted by -33 Å per nucleotide in the shorter loop strand, and was suspected to be a result of non-canonical internal loop pairing.

As noted in **Section 1.3**, this original heuristic model has remarkable predictive power, capturing over ~85% of the junction conformations observed in the PDB. The simplicity of the model unambiguously demonstrates the extent to which topological constraints limit RNA conformation. It also offers the advantage of being very computationally efficient, which in turn allows exhaustive enumeration over all theoretically possible (α_h, β_h, γ_h) conformations to quantify the allowed fraction. However, it is difficult to extend this model to study higher-order junctions, where, for example, it cannot be assumed that helices are linked by a stationary pivot. The use of only idealized helices may additionally obscure possible sequence dependencies of topological constraints.

Alternatively, Sim and Levitt (199) recently developed a more scalable heuristic approach to explore the extent to which secondary structure limits molecular conformation in tRNA, the adenine riboswitch, and the TPP riboswitch. These molecules contain a four-way junction, a three-way junction, and both a three-way and two-way junction, respectively. In their approach, 3D conformations were randomly generated using a fragment assembly library while

enforcing native base pairing and ignoring attractive interactions. This circumvents the need to assume that helices are idealized and linked by pivots, but comes at the expense of convoluting biases of the fragment library with effects of topological constraints. Thus, while the authors observed that the ensembles of 3D conformations generated by their method were constrained, and in the case of tRNA and the adenine riboswitch, were intriguingly biased towards the native tertiary fold, it is difficult to unambiguously attribute these observations to topological constraints. Fragment assembly methods are also computationally expensive, presenting potential challenges for scaling to larger RNAs.

1.5.2 Free energy methods

Despite their advantages, heuristic models are limited by their inability to provide insight into the thermodynamic contributions of topological constraints. Free energy methods, namely molecular dynamics (MD) and Monte Carlo (MC) methods, provide a well-established and rigorous framework for exploring both the extent and thermodynamic contributions of topological constraints in complex RNAs (200,201).

In MD and MC methods, a Hamiltonian (also known as a force-field) is used to represent the potential energy associated with a microscopic molecular configuration. According to the laws of classical statistical mechanics, thermodynamic quantities in the ensemble of interest can then be exactly computed by integrating over all the degrees of freedom of the system, collectively termed phase-space. Given the high-dimensionality of phase-space, solving this integral is intractable using standard methods. MC accomplishes this integration by averaging over an ensemble of molecular configurations generated by randomly sampling phase-space with a probability proportional to the Boltzmann weight (202). In MD, the molecule's atoms are

assigned random velocities and the time-trajectory simulated using Newton's equations of motion (203,204). Thermodynamic quantities can be then obtained by time-averaging over the MD simulation due to the ergodic hypothesis that time-averaging is equivalent to ensemble-averaging. The primary thermodynamic quantity of interest in this dissertation is the free energy G_j associated with a macroscopic RNA conformation j (often a given (α_h, β_h, γ_h) conformation). In both MC and MD methods, G_j can be obtained through

$$\exp\left(\frac{-G_j}{k_B T}\right) = \frac{1}{N}\sum_i \chi_j(i), \qquad \textbf{[1.5]}$$

where χ_j is the indicator function of j, and the sum is performed over all N members of the generated ensemble. In reality, calculating absolute free energies G_j accurately is both impracticable and unnecessary, and thus the difference in free energy $\Delta G_{j,k}$ between j and some reference state k is reported. The other quantity of interest in this dissertation, the degree to which a molecule is constrained, can be obtained by calculating the total number of unique macroscopic conformations represented in the ensemble.

While both MC and MD yield formally correct results, it is difficult for both methods to generate ensembles that sufficiently cover phase-space due to the expense of computing the Hamiltonian. Propagating the system through time presents particular challenges for MD simulations because low-energy molecular configurations are often separated by high-energy configurations, resulting in an energy barrier in phase-space that is infrequently crossed. A variety of methods, most notably temperature replica exchange (205), have been developed to help accelerate barrier crossing in MD simulations (206). By contrast, MC methods are not bound by time-continuity and thus in theory can sample on either side of an energy barrier without having to traverse the barrier. However, as the density of a molecular system increases, most regions of phase-space become disallowed due to steric collisions. Thus, the probability of

randomly generating a configuration with a non-zero Boltzmann weight becomes small, necessitating the use of complicated schemes to explore phase space efficiently (201). For this and other reasons, MD simulation is typically the method of choice for studying large molecules.

Gold standard free energy methods use full atomic detail and a Hamilitonian parameterized from quantum mechanical calculations and other experimental information to model the system with as close to physical accuracy to possible. Multiple such all-atom models have been developed for RNA (207,208), and in principle these models could be adapted to study topological constraints. In particular, the all-atom Hamiltonian could be modified to eliminate attractive terms and electrostatics while introducing a scheme to enforce native secondary structure. Simulations could then be performed of an RNA subject only to the forces of its secondary structure topological constraints. However, the benefits offered by atomic-accuracy are outweighed by the extreme computational expense that would be required to fully sample the conformational space of an RNA junction. Topological constraints are anticipated to primarily act on macroscopic scale, and thus microscopic detail is likely unnecessary.

An increasingly important alternative is coarse-grained (CG) models, which retain the free-energy framework of all-atom models but instead represent the system using a reduced number of ‘pseudoatoms’ (209). For example, a common approach is to represent each amino acid of a protein, or nucleotide of a nucleic acid, with a single pseudoatom. The Hamiltonian can either be derived from first principles or using a knowledge-based approach, and provides an intermediate approximation of the biomolecular free energy landscape. The advantage offered by CG models is greatly improved computational efficiency because simulation time generally grows as the square of the number of simulated atoms. Additionally, by smoothing the energy landscape, CG models accelerate the rate at which an MD simulation explores the free energy

landscape. Many coarse-grained models have been developed and successfully used to study various aspects of RNA folding (154,159,210-217). Notably, Chu *et al.* (169) used a multi-scale coarse-grained model for their studies of topological constraints in two-way junctions linked by PEG tethers; the PEG tethers were treated atomistically while the helices treated using several pseuodoatoms. Development of a coarse-grained model specifically designed to explore the contributions of topological constraints to RNA folding and dynamics is discussed further in Chapter 3.

1.6 Dissertation objectives

While recent studies by Bailor *et al.* (124) and Chu *et al.* (169) suggest that topological constraints encoded by junction secondary structure are important determinants of RNA 3D conformation and dynamics, many questions remain. In particular, the role of topological constraints in shaping the free energy landscape of both two-way and higher-order RNA junctions, and how this may influence tertiary folding and dynamics of complex RNA molecules, is unknown. In this dissertation, I first explore in greater detail the mechanism of action of topological constraints in two-way junctions, including their dependence on helix structure, connectivity, and non-canonical internal loop pairing. Then, with the aim of answering the question of what role topological constraints play in the folding and dynamics complex RNAs, I develop the TOPRNA coarse-grained molecular dynamics model to provide a scalable method for studying the contributions of topological constraints to the RNA free energy landscape. These contributions are explored in two-way junctions, followed by tRNA and variants thereof, and finally the large multiple-junction-containing *Azoarcus* Group I ribozyme.

1.7 References

1. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849-857.
2. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E. and Cech, T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147-157.
3. Rinn, J.L. and Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, **81**, 145-166.
4. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215-233.
5. Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA. *Hum Mol Genet*, **15 Spec No 1**, R17-29.
6. Cooper, T.A., Wan, L. and Dreyfuss, G. (2009) RNA and disease. *Cell*, **136**, 777-793.
7. Esteller, M. (2011) Non-coding RNAs in human disease. *Nat Rev Genet*, **12**, 861-874.
8. Fica, S.M., Tuttle, N., Novak, T., Li, N.S., Lu, J., Koodathingal, P., Dai, Q., Staley, J.P. and Piccirilli, J.A. (2013) RNA catalyses nuclear pre-mRNA splicing. *Nature*, **503**, 229-234.
9. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327-339.
10. Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905-920.
11. Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F. *et al.* (2000) Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, **102**, 615-623.
12. Voorhees, R.M. and Ramakrishnan, V. (2013) Structural basis of the translational elongation cycle. *Annu Rev Biochem*, **82**, 203-236.
13. Zaher, H.S. and Green, R. (2009) Fidelity at the molecular level: lessons from protein synthesis. *Cell*, **136**, 746-762.
14. Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Jr., Swanstrom, R., Burch, C.L. and Weeks, K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711-716.
15. Houck-Loomis, B., Durney, M.A., Salguero, C., Shankar, N., Nagle, J.M., Goff, S.P. and D'Souza, V.M. (2011) An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature*, **480**, 561-564.
16. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101-108.
17. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C. and Snyder, M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
18. Gesteland, R.F., Cech, T.R. and Atkins, J.F. (2006) *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor Laboratory Press.
19. Levitt, M. (1969) Detailed molecular model for transfer ribonucleic acid. *Nature*, **224**, 759-763.

20. Cole, P.E., Crothers, D.M. and Yang, S.K. (1972) Conformational-Changes of Transfer Ribonucleic-Acid - Equilibrium Phase-Diagrams. *Biochemistry*, **11**, 4358-&.
21. Kim, S.H., Suddath, F.L., Quigley, G.J., McPherson, A., Sussman, J.L., Wang, A.H., Seeman, N.C. and Rich, A. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, **185**, 435-440.
22. Giege, R., Juhling, F., Putz, J., Stadler, P., Sauter, C. and Florentz, C. (2012) Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip Rev RNA*, **3**, 37-61.
23. Zhang, J. and Ferre-D'Amare, A.R. (2013) Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. *Nature*, **500**, 363-366.
24. Reiter, N.J., Osterman, A., Torres-Larios, A., Swinger, K.K., Pan, T. and Mondragon, A. (2010) Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature*, **468**, 784-789.
25. Behlen, L.S., Sampson, J.R., DiRenzo, A.B. and Uhlenbeck, O.C. (1990) Lead-catalyzed cleavage of yeast tRNAPhe mutants. *Biochemistry*, **29**, 2515-2523.
26. Sampson, J.R., DiRenzo, A.B., Behlen, L.S. and Uhlenbeck, O.C. (1990) Role of the tertiary nucleotides in the interaction of yeast phenylalanine tRNA with its cognate synthetase. *Biochemistry*, **29**, 2523-2532.
27. Holbrook, S.R. (2008) Structural principles from large RNAs. *Annu Rev Biophys*, **37**, 445-464.
28. Marcia, M. and Pyle, A.M. (2012) Visualizing group II intron catalysis through the stages of splicing. *Cell*, **151**, 497-507.
29. Scott, W.G., Finch, J.T. and Klug, A. (1995) The crystal structure of an aii-rnahammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell*, **81**, 991-1002.
30. Ferre-D'Amare, A.R., Zhou, K. and Doudna, J.A. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567-574.
31. Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J. and Strobel, S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45-50.
32. Serganov, A. and Patel, D.J. (2012) Metabolite recognition principles and molecular mechanisms underlying riboswitch function. *Annu Rev Biophys*, **41**, 343-370.
33. Batey, R.T., Gilbert, S.D. and Montange, R.K. (2004) Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, **432**, 411-415.
34. Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E. and Chang, H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689-693.
35. Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26-46.
36. Boehr, D.D., Nussinov, R. and Wright, P.E. (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol*, **5**, 789-796.
37. Al-Hashimi, H.M. and Walter, N.G. (2008) RNA dynamics: it is about time. *Curr Opin Struct Biol*, **18**, 321-329.
38. Dethoff, E.A., Chugh, J., Mustoe, A.M. and Al-Hashimi, H.M. (2012) Functional complexity and regulation through RNA dynamics. *Nature*, **482**, 322-330.
39. Mustoe, A.M., Brooks, C.L. and Al-Hashimi, H.M. (2014) Hierarchy of RNA functional dynamics. *Annu Rev Biochem*, **83**, 441-466.

40. Green, N.J., Grundy, F.J. and Henkin, T.M. (2010) The T box mechanism: tRNA as a regulatory molecule. *FEBS letters*, **584**, 318-324.
41. Ray, P.S., Jia, J., Yao, P., Majumder, M., Hatzoglou, M. and Fox, P.L. (2009) A stress-responsive RNA switch regulates VEGFA expression. *Nature*, **457**, 915-919.
42. Haller, A., Souliere, M.F. and Micura, R. (2011) The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Acc Chem Res*, **44**, 1339-1348.
43. Winkler, W., Nahvi, A. and Breaker, R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952-956.
44. Cromie, M.J., Shi, Y., Latifi, T. and Groisman, E.A. (2006) An RNA sensor for intracellular Mg(2+). *Cell*, **125**, 71-84.
45. Giuliodori, A.M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C.O. and Pon, C.L. (2010) The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Mol Cell*, **37**, 21-33.
46. Cruz, J.A. and Westhof, E. (2009) The dynamic landscapes of RNA architecture. *Cell*, **136**, 604-609.
47. Alexander, R.W., Eargle, J. and Luthey-Schulten, Z. (2010) Experimental and computational determination of tRNA dynamics. *FEBS Lett*, **584**, 376-386.
48. Fischer, N., Konevega, A.L., Wintermeyer, W., Rodnina, M.V. and Stark, H. (2010) Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature*, **466**, 329-333.
49. Zhang, W., Dunkle, J.A. and Cate, J.H. (2009) Structures of the ribosome in intermediate states of ratcheting. *Science*, **325**, 1014-1017.
50. Agirrezabala, X., Liao, H.Y., Schreiner, E., Fu, J., Ortiz-Meoz, R.F., Schulten, K., Green, R. and Frank, J. (2012) Structural characterization of mRNA-tRNA translocation intermediates. *Proc Natl Acad Sci USA*, **109**, 6094-6099.
51. Mulder, A.M., Yoshioka, C., Beck, A.H., Bunner, A.E., Milligan, R.A., Potter, C.S., Carragher, B. and Williamson, J.R. (2010) Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. *Science*, **330**, 673-677.
52. Adilakshmi, T., Bellur, D.L. and Woodson, S.A. (2008) Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly. *Nature*, **455**, 1268-1272.
53. Kim, H., Abeysirigunawarden, S.C., Chen, K., Mayerle, M., Ragunathan, K., Luthey-Schulten, Z., Ha, T. and Woodson, S.A. (2014) Protein-guided RNA dynamics during early ribosome assembly. *Nature*, **506**, 334-338.
54. Menichelli, E., Isel, C., Oubridge, C. and Nagai, K. (2007) Protein-induced conformational changes of RNA during the assembly of human signal recognition particle. *J Mol Biol*, **367**, 187-203.
55. Stone, M.D., Mihalusova, M., O'Connor C, M., Prathapam, R., Collins, K. and Zhuang, X. (2007) Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein. *Nature*, **446**, 458-461.
56. Herschlag, D. (1992) Evidence for processivity and two-step binding of the RNA substrate from studies of J1/2 mutants of the Tetrahymena ribozyme. *Biochemistry*, **31**, 1386-1399.
57. Bevilacqua, P.C., Kierzek, R., Johnson, K.A. and Turner, D.H. (1992) Dynamics of ribozyme binding of substrate revealed by fluorescence-detected stopped-flow methods. *Science*, **258**, 1355-1358.

58. Zhuang, X., Kim, H., Pereira, M.J., Babcock, H.P., Walter, N.G. and Chu, S. (2002) Correlating structural dynamics and function in single ribozyme molecules. *Science*, **296**, 1473-1476.
59. Breaker, R.R. (2012) Riboswitches and the RNA World. *Cold Spring Harb Perspect Biol*, **4**.
60. Tan, E., Wilson, T.J., Nahas, M.K., Clegg, R.M., Lilley, D.M. and Ha, T. (2003) A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc Natl Acad Sci USA*, **100**, 9308-9313.
61. Shi, X., Solomatin, S.V. and Herschlag, D. (2012) A role for a single-stranded junction in RNA binding and specificity by the Tetrahymena group I ribozyme. *J Am Chem Soc*, **134**, 1910-1913.
62. Frauenfelder, H., Sligar, S.G. and Wolynes, P.G. (1991) The energy landscapes and motions of proteins. *Science*, **254**, 1598-1603.
63. Onuchic, J.N., Luthey-Schulten, Z. and Wolynes, P.G. (1997) Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem*, **48**, 545-600.
64. Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct*, **26**, 113-137.
65. Tinoco, I., Jr. and Bustamante, C. (1999) How RNA folds. *J Mol Biol*, **293**, 271-281.
66. Schultes, E.A., Spasic, A., Mohanty, U. and Bartel, D.P. (2005) Compact and ordered collapse of randomly generated RNA sequences. *Nat Struct Mol Biol*, **12**, 1130-1136.
67. Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) RNA structure prediction. *Annu Rev Biophys Biophys Chem*, **17**, 167-192.
68. Tinoco, I., Jr., Borer, P.N., Dengler, B., Levin, M.D., Uhlenbeck, O.C., Crothers, D.M. and Bralla, J. (1973) Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol*, **246**, 40-41.
69. Mathews, D.H., Moss, W.N. and Turner, D.H. (2010) Folding and finding RNA secondary structure. *Cold Spring Harb Perspect Biol*, **2**, a003665.
70. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
71. Fox, G.W. and Woese, C.R. (1975) 5S RNA secondary structure. *Nature*, **256**, 505-507.
72. Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res*, **37**, D159-162.
73. Uhlenbeck, O.C. (1995) Keeping RNA happy. *RNA*, **1**, 4-6.
74. Treiber, D.K. and Williamson, J.R. (2001) Beyond kinetic traps in RNA folding. *Curr Opin Struct Biol*, **11**, 309-314.
75. Herschlag, D. (1995) RNA chaperones and the RNA folding problem. *J Biol Chem*, **270**, 20871-20874.
76. Woodson, S.A. (2010) Taming free energy landscapes with RNA chaperones. *RNA Biology*, **7**, 677-686.
77. Lin, J.C., Hyeon, C. and Thirumalai, D. (2012) RNA under tension: Folding Landscapes, Kinetic partitioning Mechanism, and Molecular Tensegrity. *J Phys Chem Lett*, **3**, 3616-3625.

78. Zhu, J.Y., Steif, A., Proctor, J.R. and Meyer, I.M. (2013) Transient RNA structure features are evolutionarily conserved and can be computationally predicted. *Nucleic Acids Res*, **41**, 6273-6285.
79. Grohman, J.K., Gorelick, R.J., Lickwar, C.R., Lieb, J.D., Bower, B.D., Znosko, B.M. and Weeks, K.M. (2013) A Guanosine-Centric Mechanism for RNA Chaperone Function. *Science*, **340**, 190-195.
80. SantaLucia, J., Jr. and Turner, D.H. (1993) Structure of (rGGCGAGCC)2 in solution from NMR and restrained molecular dynamics. *Biochemistry*, **32**, 12612-12623.
81. Leontis, N.B., Stombaugh, J. and Westhof, E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res*, **30**, 3497-3531.
82. Butcher, S.E. and Pyle, A.M. (2011) The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc Chem Res*, **44**, 1302-1311.
83. Doherty, E.A., Batey, R.T., Masquida, B. and Doudna, J.A. (2001) A universal mode of helix packing in RNA. *Nat Struct Biol*, **8**, 339-343.
84. Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B. and Steitz, T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci USA*, **98**, 4899-4903.
85. Reiter, N.J., Chan, C.W. and Mondragon, A. (2011) Emerging structural themes in large RNA molecules. *Curr Opin Struct Biol*, **21**, 319-326.
86. Gautheret, D., Damberger, S.H. and Gutell, R.R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J Mol Biol*, **248**, 27-43.
87. Battle, D.J. and Doudna, J.A. (2002) Specificity of RNA-RNA helix recognition. *Proc Natl Acad Sci USA*, **99**, 11676-11681.
88. Costa, M. and Michel, F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*, **14**, 1276-1285.
89. Behrouzi, R., Roh, J.H., Kilburn, D., Briber, R.M. and Woodson, S.A. (2012) Cooperative tertiary interaction network guides RNA folding. *Cell*, **149**, 348-357.
90. Sattin, B.D., Zhao, W., Travers, K., Chu, S. and Herschlag, D. (2008) Direct measurement of tertiary contact cooperativity in RNA folding. *J Am Chem Soc*, **130**, 6085-6087.
91. Kilburn, D., Roh, J.H., Behrouzi, R., Briber, R.M. and Woodson, S.A. (2013) Crowders perturb the entropy of RNA energy landscapes to favor folding. *J Am Chem Soc*, **135**, 10055-10063.
92. Kilburn, D., Roh, J.H., Guo, L., Briber, R.M. and Woodson, S.A. (2010) Molecular crowding stabilizes folded RNA structure by the excluded volume effect. *J Am Chem Soc*, **132**, 8690-8696.
93. Fiore, J.L., Holmstrom, E.D. and Nesbitt, D.J. (2012) Entropic origin of Mg2+-facilitated RNA folding. *Proc Natl Acad Sci USA*, **109**, 2902-2907.
94. Misra, V.K. and Draper, D.E. (2000) Mg(2+) binding to tRNA revisited: the nonlinear Poisson-Boltzmann model. *J Mol Biol*, **299**, 813-825.
95. Lipfert, J., Doniach, S., Das, R. and Herschlag, D. (2014) Understanding nucleic Acid-ion interactions. *Annu Rev Biochem*, **83**, 813-841.
96. Draper, D.E. (2008) RNA folding: thermodynamic and molecular descriptions of the roles of ions. *Biophys J*, **95**, 5489-5495.

97. Bai, Y., Chu, V.B., Lipfert, J., Pande, V.S., Herschlag, D. and Doniach, S. (2008) Critical assessment of nucleic acid electrostatics via experimental and computational investigation of an unfolded state ensemble. *J Am Chem Soc*, **130**, 12334-12341.
98. Leipply, D. and Draper, D.E. (2011) Evidence for a thermodynamically distinct Mg2+ ion associated with formation of an RNA tertiary structure. *J Am Chem Soc*, **133**, 13397-13405.
99. Tan, Z.J. and Chen, S.J. (2012) Ion-mediated RNA structural collapse: effect of spatial confinement. *Biophys J*, **103**, 827-836.
100. Woodson, S.A. (2010) Compact intermediates in RNA folding. *Annu Rev Biophys*, **39**, 61-77.
101. Thirumalai, D. and Hyeon, C. (2009) In Walter, N. G., Woodson, S. A. and Batey, R. (eds.), *Non-Protein Coding RNAs*. Springer-Verlag Berlin, Heidelberger Platz 3, D-14197 Berlin, Germany, Vol. 13, pp. 27-47.
102. Russell, R., Das, R., Suh, H., Travers, K.J., Laederach, A., Engelhardt, M.A. and Herschlag, D. (2006) The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J Mol Biol*, **363**, 531-544.
103. Chen, C., Jiang, L., Michalczyk, R. and Russu, I.M. (2006) Structural energetics and base-pair opening dynamics in sarcin-ricin domain RNA. *Biochemistry*, **45**, 13606-13613.
104. Burkard, M.E. and Turner, D.H. (2000) NMR structures of r(GCAGGCGUGC)2 and determinants of stability for single guanosine-guanosine base pairs. *Biochemistry*, **39**, 11748-11762.
105. Yildirim, I., Park, H., Disney, M.D. and Schatz, G.C. (2013) A dynamic structural model of expanded RNA CAG repeats: a refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *J Am Chem Soc*, **135**, 3528-3538.
106. Dethoff, E.A., Petzold, K., Chugh, J., Casiano-Negroni, A. and Al-Hashimi, H.M. (2012) Visualizing transient low-populated structures of RNA. *Nature*, **491**, 724-728.
107. Walter, N.G., Burke, J.M. and Millar, D.P. (1999) Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. *Nat Struct Biol*, **6**, 544-549.
108. Zhang, Q., Sun, X., Watt, E.D. and Al-Hashimi, H.M. (2006) Resolving the motional modes that code for RNA adaptation. *Science*, **311**, 653-656.
109. Zhang, Q., Stelzer, A.C., Fisher, C.K. and Al-Hashimi, H.M. (2007) Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, **450**, 1263-1267.
110. Jucker, F.M., Heus, H.A., Yip, P.F., Moors, E.H. and Pardi, A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J Mol Biol*, **264**, 968-980.
111. Johnson, J.E., Jr. and Hoogstraten, C.G. (2008) Extensive backbone dynamics in the GCAA RNA tetraloop analyzed using 13C NMR spin relaxation and specific isotope labeling. *J Am Chem Soc*, **130**, 16757-16769.
112. Walter, A.E., Turner, D.H., Kim, J., Lyttle, M.H., Muller, P., Mathews, D.H. and Zuker, M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci USA*, **91**, 9218-9222.
113. Tyagi, R. and Mathews, D.H. (2007) Predicting helical coaxial stacking in RNA multibranch loops. *RNA*, **13**, 939-951.

114. Salmon, L., Bascom, G., Andricioaei, I. and Al-Hashimi, H.M. (2013) A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed. *J Am Chem Soc*, **135**, 5457-5466.
115. Getz, M.M., Andrews, A.J., Fierke, C.A. and Al-Hashimi, H.M. (2007) Structural plasticity and Mg2+ binding properties of RNase P P4 from combined analysis of NMR residual dipolar couplings and motionally decoupled spin relaxation. *RNA*, **13**, 251-266.
116. Olsen, G.L., Bardaro, M.F., Jr., Echodu, D.C., Drobny, G.P. and Varani, G. (2010) Intermediate rate atomic trajectories of RNA by solid-state NMR spectroscopy. *J Am Chem Soc*, **132**, 303-308.
117. Hohng, S., Wilson, T.J., Tan, E., Clegg, R.M., Lilley, D.M. and Ha, T. (2004) Conformational flexibility of four-way junctions in RNA. *J Mol Biol*, **336**, 69-79.
118. Melcher, S.E., Wilson, T.J. and Lilley, D.M. (2003) The dynamic nature of the four-way junction of the hepatitis C virus IRES. *RNA*, **9**, 809-820.
119. Reblova, K., Sponer, J. and Lankas, F. (2012) Structure and mechanical properties of the ribosomal L1 stalk three-way junction. *Nucleic Acids Res*, **40**, 6290-6303.
120. Besseova, I., Reblova, K., Leontis, N.B. and Sponer, J. (2010) Molecular dynamics simulations suggest that RNA three-way junctions can act as flexible RNA structural elements in the ribosome. *Nucleic Acids Res*, **38**, 6247-6264.
121. Grant, G.P., Boyd, N., Herschlag, D. and Qin, P.Z. (2009) Motions of the substrate recognition duplex in a group I intron assessed by site-directed spin labeling. *J Am Chem Soc*, **131**, 3136-3137.
122. Zhang, Q., Kim, N.K., Peterson, R.D., Wang, Z. and Feigon, J. (2010) Structurally conserved five nucleotide bulge determines the overall topology of the core domain of human telomerase RNA. *Proc Natl Acad Sci USA*, **107**, 18761-18768.
123. Haller, A., Altman, R.B., Souliere, M.F., Blanchard, S.C. and Micura, R. (2013) Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. *Proc Natl Acad Sci USA*, **110**, 4188-4193.
124. Bailor, M.H., Sun, X.Y. and Al-Hashimi, H.M. (2010) Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation. *Science*, **327**, 202-206.
125. Wu, M. and Tinoco, I., Jr. (1998) RNA folding causes secondary structure rearrangement. *Proc Natl Acad Sci USA*, **95**, 11555-11560.
126. Koculi, E., Cho, S.S., Desai, R., Thirumalai, D. and Woodson, S.A. (2012) Folding path of P5abc RNA involves direct coupling of secondary and tertiary structures. *Nucleic Acids Res*, **40**, 8011-8020.
127. Frieda, K.L. and Block, S.M. (2012) Direct observation of cotranscriptional folding in an adenine riboswitch. *Science*, **338**, 397-400.
128. Privalov, P.L. and Filimonov, V.V. (1978) Thermodynamic analysis of transfer RNA unfolding. *J Mol Biol*, **122**, 447-464.
129. Cho, S.S., Pincus, D.L. and Thirumalai, D. (2009) Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *Proc Natl Acad Sci USA*, **106**, 17349-17354.
130. Feng, J., Walter, N.G. and Brooks, C.L., III. (2011) Cooperative and directional folding of the preQ1 riboswitch aptamer domain. *J Am Chem Soc*, **133**, 4196-4199.
131. Lin, J.C. and Thirumalai, D. (2008) Relative stability of helices determines the folding landscape of adenine riboswitch aptamers. *J Am Chem Soc*, **130**, 14080-14081.

132. Bindewald, E., Hayes, R., Yingling, Y.G., Kasprzak, W. and Shapiro, B.A. (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res*, **36**, D392-397.
133. Bhattacharyya, A., Murchie, A.I. and Lilley, D.M. (1990) RNA bulges and the helical periodicity of double-stranded RNA. *Nature*, **343**, 484-487.
134. Gohlke, C., Murchie, A.I.H., Lilley, D.M.J. and Clegg, R.M. (1994) Kinking of DNA and RNA Helices by Bulged Nucleotides Observed by Fluorescence Resonance Energy-Transfer. *Proc Natl Acad Sci USA*, **91**, 11660-11664.
135. Hsieh, C.H. and Griffith, J.D. (1989) Deletions of bases in one strand of duplex DNA, in contrast to single-base mismatches, produce highly kinked molecules: possible relevance to the folding of single-stranded nucleic acids. *Proc Natl Acad Sci USA*, **86**, 4833-4837.
136. Tang, R.S. and Draper, D.E. (1990) Bulge loops used to measure the helical twist of RNA in solution. *Biochemistry*, **29**, 5232-5237.
137. Zacharias, M. and Hagerman, P.J. (1995) Bulge-Induced Bends in RNA - Quantification by Transient Electric Birefringence. *J Mol Biol*, **247**, 486-500.
138. Zacharias, M. and Hagerman, P.J. (1996) The influence of symmetric internal loops on the flexibility of RNA. *J Mol Biol*, **257**, 276-289.
139. Lilley, D.M. (2000) Structures of helical junctions in nucleic acids. *Q Rev Biophys*, **33**, 109-159.
140. Marck, C. and Grosjean, H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189-1232.
141. Michel, F. and Westhof, E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol*, **216**, 585-610.
142. Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83-93.
143. Laing, C. and Schlick, T. (2009) Analysis of four-way junctions in RNA structures. *J Mol Biol*, **390**, 547-559.
144. Laing, C., Jung, S., Iqbal, A. and Schlick, T. (2009) Tertiary motifs revealed in analyses of higher-order RNA junctions. *J Mol Biol*, **393**, 67-82.
145. Muller, S., Appel, B., Krellenberg, T. and Petkovic, S. (2012) The many faces of the hairpin ribozyme: structural and functional variants of a small catalytic RNA. *IUBMB Life*, **64**, 36-47.
146. Zhao, Z.Y., Wilson, T.J., Maxwell, K. and Lilley, D.M. (2000) The folding of the hairpin ribozyme: dependence on the loops and the junction. *RNA*, **6**, 1833-1846.
147. Klostermeier, D. and Millar, D.P. (2000) Helical junctions as determinants for RNA folding: origin of tertiary structure stability of the hairpin ribozyme. *Biochemistry*, **39**, 12970-12978.
148. Hajdin, C.E., Bellaousov, S., Huggins, W., Leonard, C.W., Mathews, D.H. and Weeks, K.M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci USA*, **110**, 5498-5503.
149. Bida, J.P. and Maher, L.J., III. (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. *RNA*, **18**, 385-393.
150. Das, R., Karanicolas, J. and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods*, **7**, 291-294.

151. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51-55.
152. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K.J., Lukasiak, P., Bartol, N., Blazewicz, J. and Adamiak, R.W. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res*, **40**, e112.
153. Kim, N., Laing, C., Elmetwaly, S., Jung, S., Curuksu, J. and Schlick, T. (2014) Graph-based sampling for approximating global helical topologies of RNA. *Proc Natl Acad Sci USA*, **111**, 4079-4084.
154. Cao, S. and Chen, S.J. (2011) Physics-based de novo prediction of RNA 3D structures. *J Phys Chem B*, **115**, 4216-4226.
155. Lavender, C.A., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2010) Robust and Generic RNA Modeling Using Inferred Constraints: A Structure for the Hepatitis C Virus IRES Pseudoknot Domain. *Biochemistry*, **49**, 4931-4933.
156. Ding, F., Lavender, C.A., Weeks, K.M. and Dokholyan, N.V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods*, **9**, 603-608.
157. Seetin, M.G. and Mathews, D.H. (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J Comput Chem*, **32**, 2232-2244.
158. Flores, S.C. and Altman, R.B. (2010) Turning limited experimental information into 3D models of RNA. *RNA*, **16**, 1769-1778.
159. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189-199.
160. Das, R., Kudaravalli, M., Jonikas, M., Laederach, A., Fong, R., Schwans, J.P., Baker, D., Piccirilli, J.A., Altman, R.B. and Herschlag, D. (2008) Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci USA*, **105**, 4144-4149.
161. Gherghe, C.M., Leonard, C.W., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2009) Native-like RNA Tertiary Structures Using a Sequence-Encoded Cleavage Agent and Refinement by Discrete Molecular Dynamics. *J Am Chem Soc*, **131**, 2541-2546.
162. Yang, S.C., Parisien, M., Major, F. and Roux, B. (2010) RNA Structure Determination Using SAXS Data. *J Phys Chem B*, **114**, 10039-10048.
163. Neidle, S. (1999) *Oxford handbook of nucleic acid structure*. Oxford University Press.
164. Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol*, **7**, 95-99.
165. Murray, L.J., Arendall, W.B., 3rd, Richardson, D.C. and Richardson, J.S. (2003) RNA backbone is rotameric. *Proc Natl Acad Sci USA*, **100**, 13904-13909.
166. Humphris-Narayanan, E. and Pyle, A.M. (2012) Discrete RNA libraries from pseudo-torsional space. *J Mol Biol*, **421**, 6-26.
167. Richards, E.G. (1980) *An introduction to the physical properties of large molecules in solution*. Cambridge University Press, New York.
168. Frank, A.T., Stelzer, A.C., Al-Hashimi, H.M. and Andricioaei, I. (2009) Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res*, **37**, 3670-3679.

169. Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S. and Herschlag, D. (2009) Do conformational biases of simple helical junctions influence RNA folding stability and specificity? *RNA*, **15**, 2195-2205.
170. Hooft, R.W., Sander, C. and Vriend, G. (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. *Comput Appl Biosci*, **13**, 425-430.
171. Bertini, I., Cavallaro, G., Luchinat, C. and Poli, I. (2003) A use of Ramachandran potentials in protein solution structure determinations. *J Biomol NMR*, **26**, 355-366.
172. Kumar, M.V. and Swaminathan, R. (2010) A novel approach to segregate and identify functional loop regions in protein structures using their Ramachandran maps. *Proteins*, **78**, 900-916.
173. Wu, D., Jernigan, R. and Wu, Z. (2007) Refinement of NMR-determined protein structures with database derived mean-force potentials. *Proteins*, **68**, 232-242.
174. Duarte, C.M. and Pyle, A.M. (1998) Stepping through an RNA structure: A novel approach to conformational analysis. *J Mol Biol*, **284**, 1465-1478.
175. Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.J., Neidle, S., Shakked, Z. *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol*, **313**, 229-237.
176. Gast, F.U., Amiri, K.M. and Hagerman, P.J. (1994) Interhelix geometry of stems I and II of a self-cleaving hammerhead RNA. *Biochemistry*, **33**, 1788-1796.
177. Leehey, M.A., Squassoni, C.A., Friederich, M.W., Mills, J.B. and Hagerman, P.J. (1995) A noncanonical tertiary conformation of a human mitochondrial transfer RNA. *Biochemistry*, **34**, 16235-16239.
178. Friederich, M.W., Gast, F.U., Vacano, E. and Hagerman, P.J. (1995) Determination of the Angle between the Anticodon and Aminoacyl Acceptor Stems of Yeast Phenylalanyl Transfer-Rna in Solution. *Proc Natl Acad Sci USA*, **92**, 4803-4807.
179. Zacharias, M. and Hagerman, P.J. (1997) Influence of static and dynamic bends on the birefringence decay profile of RNA helices: Brownian dynamics simulations. *Biophys J*, **73**, 318-332.
180. Nakamura, T.M., Wang, Y.H., Zaug, A.J., Griffith, J.D. and Cech, T.R. (1995) Relative orientation of RNA helices in a group 1 ribozyme determined by helix extension electron microscopy. *EMBO J*, **14**, 4849-4859.
181. Tang, R.S. and Draper, D.E. (1994) On the use of phasing experiments to measure helical repeat and bulge loop-associated twist in RNA. *Nucleic Acids Res*, **22**, 835-841.
182. Al-Hashimi, H.M., Gosser, Y., Gorin, A., Hu, W., Majumdar, A. and Patel, D.J. (2002) Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. *J Mol Biol*, **315**, 95-102.
183. Zhang, Q. and Al-Hashimi, H.M. (2009) Domain-elongation NMR spectroscopy yields new insights into RNA dynamics and adaptive recognition. *RNA*, **15**, 1941-1948.
184. Bax, A. and Grishaev, A. (2005) Weak alignment NMR: a hawk-eyed view of biomolecular structure. *Curr Opin Struct Biol*, **15**, 563-570.
185. MacDonald, D. and Lu, P. (2002) Residual dipolar couplings in nucleic acid structure determination. *Curr Opin Struct Biol*, **12**, 337-343.
186. Zhang, Q. and Al-Hashimi, H.M. (2008) Extending the NMR spatial resolution limit for RNA by motional couplings. *Nat Methods*, **5**, 243-245.
187. Casiano-Negroni, A., Sun, X. and Al-Hashimi, H.M. (2007) Probing Na(+)-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar

couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry*, **46**, 6525-6535.
188. Dethoff, E.A., Hansen, A.L., Zhang, Q. and Al-Hashimi, H.M. (2010) Variable helix elongation as a tool to modulate RNA alignment and motional couplings. *J Magn Reson*, **202**, 117-121.
189. Bailor, M.H., Musselman, C., Hansen, A.L., Gulati, K., Patel, D.J. and Al-Hashimi, H.M. (2007) Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat Protoc*, **2**, 1536-1546.
190. Musselman, C., Pitt, S.W., Gulati, K., Foster, L.L., Andricioaei, I. and Al-Hashimi, H.M. (2006) Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J Biomol NMR*, **36**, 235-249.
191. Holbrook, S.R. and Kim, S.H. (1984) Local mobility of nucleic acids as determined from crystallographic data. I. RNA and B form DNA. *J Mol Biol*, **173**, 361-388.
192. Foloppe, N. and MacKerell, A.D. (1998) Conformational properties of the deoxyribose and ribose moieties of nucleic acids: A quantum mechanical study. *J Phys Chem B*, **102**, 6669-6678.
193. Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. and Westhof, E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*, **31**, 3450-3460.
194. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) 3D maps of RNA interhelical junctions. *Nat Protoc*, **6**, 1536-1545.
195. Varshalovich, D.A., Moskalev, A. and Khersonskii, V. (1988) *Quantum theory of angular momentum*. World Scientific.
196. Lilley, D.M.J., Clegg, R.M., Diekmann, S., Seeman, N.C., Vonkitzing, E. and Hagerman, P.J. (1995) A nomenclature of junctions and branchpoints in nucleic acids. *Nucleic Acids Res*, **23**, 3363-3364.
197. Fisher, C.K. and Al-Hashimi, H.M. (2009) Approximate reconstruction of continuous spatially complex domain motions by multialignment NMR residual dipolar couplings. *J Phys Chem B*, **113**, 6173-6176.
198. Tolman, J.R., Al-Hashimi, H.M., Kay, L.E. and Prestegard, J.H. (2001) Structural and dynamic analysis of residual dipolar coupling data for proteins. *J Am Chem Soc*, **123**, 1416-1424.
199. Sim, A.Y.L. and Levitt, M. (2011) Clustering to identify RNA conformations constrained by secondary structure. *Proc Natl Acad Sci USA*, **108**, 3590-3595.
200. Brooks, C.L., Karplus, M. and Pettitt, B.M. (1987) *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*. J. Wiley.
201. Zuckerman, D.M. (2011) Equilibrium sampling in biomolecular simulations. *Annu Rev Biophys*, **40**, 41-62.
202. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J Chem Phys*, **21**, 1087-1092.
203. Rahman, A. (1964) Correlations in the motion of atoms in liquid argon. *Phys Rev*, **136**, A405.
204. McCammon, J.A., Gelin, B.R. and Karplus, M. (1977) Dynamics of folded proteins. *Nature*, **267**, 585-590.

205. Sugita, Y. and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*, **314**, 141-151.
206. Abrams, C. and Bussi, G. (2013) Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, **16**, 163-199.
207. Denning, E.J., Priyakumar, U.D., Nilsson, L. and Mackerell, A.D., Jr. (2011) Impact of 2'-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA. *J Comput Chem*, **32**, 1929-1943.
208. Zgarbova, M., Otyepka, M., Sponer, J., Mladek, A., Banas, P., Cheatham, T.E., 3rd and Jurecka, P. (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput*, **7**, 2886-2902.
209. Hyeon, C. and Thirumalai, D. (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat Commun*, **2**, 487.
210. Tan, R.K.Z., Petrov, A.S. and Harvey, S.C. (2006) YUP: A molecular simulation program for coarse-grained and multiscaled models. *J Chem Theory Comput*, **2**, 529-540.
211. Hyeon, C. and Thirumalai, D. (2005) Mechanical unfolding of RNA hairpins. *Proc Natl Acad Sci USA*, **102**, 6789-6794.
212. Denesyuk, N.A. and Thirumalai, D. (2013) Coarse-grained model for predicting RNA folding thermodynamics. *J Phys Chem B*, **117**, 4901-4911.
213. Ding, F., Sharma, S., Chalasani, P., Demidov, V.V., Broude, N.E. and Dokholyan, N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164-1173.
214. Sharma, S., Ding, F. and Dokholyan, N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951-1952.
215. Bernauer, J., Huang, X.H., Sim, A.Y.L. and Levitt, M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066-1075.
216. Xia, Z., Gardner, D.P., Gutell, R.R. and Ren, P. (2010) Coarse-grained model for simulation of RNA three-dimensional structures. *J Phys Chem B*, **114**, 13497-13506.
217. Pasquali, S. and Derreumaux, P. (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. *J Phys Chem B*, **114**, 11957-11966.

Chapter 2: The Dependence of Topological Constraints on Helix Structure, Connectivity, and Non-Canonical Pairing

This chapter has been adapted from the following publication:

Mustoe, A. M., Bailor, M. H., Teixeira, R. M., Brooks, C. L. III, and Al-Hashimi, H. M. (2012) New insights into the fundamental role of topological constraints as a determinant of two-way junction conformation. *Nucleic Acids Research*, **40**, 892-904.

2.1 Introduction

Prior studies by Bailor, Sun, and Al-Hashimi (1) and the Herschlag group (2) have shown that two-way junctions impose steric and connectivity constraints that restrict the relative orientation of A-form helices. These constraints include interhelix steric clashes and connectivity constraints due to the finite length of the junction nucleotides linking the helices. Using a simple heuristic model that assumed canonical A-form helix geometry and a maximum inter-nucleotide distance of 4.9 Å, Bailor *et al.* showed that these constraints alone restrict the range of accessible inter-helical conformations to between 4%-20%, and on average 7%, of the total interhelical conformational space (1). Despite this narrow range, the heuristic model accommodates ~85% of all interhelical conformations observed in all two-way junctions in the PDB, including bulges and internal loops (1).

In this Chapter, we explore how various parameters of interest influence the topologically allowed space with the goal of developing a deeper and more predictive understanding of how two-way junctions encode 3D RNA conformation and dynamics. In particular, we examine how

the allowed conformational space varies with local variations in A-form helix geometry, the presence/absence of non-canonical base pairs at the junction, relative lengths of the 5' and 3' helices, and overall junction topology (bulge, symmetric and asymmetric internal loops). We also provide a firm quantitative basis for the well known yet poorly understood directionality of junction-induced bends as well as explain the origins of recently reported correlated variations in the inter-helical twist and bend angles observed across two-way junctions (1,3). Additionally, we survey the structures of symmetric and asymmetric internal loops and show that the resulting allowed inter-helical conformational space is accurately modeled by assuming that internal loops will maximize non-canonical base-pairing, marking a significant simplification in the link between RNA secondary and tertiary structure. Our results also uncover a new simple mechanism for coupling junction-induced topological constraints with tertiary interactions.

2.2 Methods

2.2.1 Computing topologically allowed inter-helical orientations

We utilize the $H_iS_XH_jS_Y$ notation (1,4) to describe the topology of a two-helix junction where i and j specify the length of the 5' and 3' helices, respectively, and X and Y ($X \geq Y$) denote the number of single stranded nucleotides in the 5' and 3' strands with the respective strands assigned in accordance to our prior conventions (1). For simplicity we refer to a junction topology in the Chapter as S_XS_Y, with the values of i and j implied by context; junctions from the PDB possess $i,j \geq 3$, and in reference to our model, unless otherwise stated, $i=j=4$. For a structure of a given two-way junction topology the inter-helical conformation can be defined through the use of the three Euler angles needed to describe the orientation of the chiral helical objects with respect to one another. As described in **Chapter 1.4** and (1,5), these angles, denoted (α_h, β_h, γ_h),

are obtained from the rotation matrix needed to transform the 3' helix from its observed orientation to one where it is coaxially stacked upon its 5' partner, and where α_h and γ_h are interpreted as twists around the 3' and 5' helices and β_h as an inter-helical bend angle.

As described by Bailor *et al.* (1), this representation of inter-helical conformation can readily inverted to compute the sterically and connectivity allowed conformations. Briefly, an idealized A-form helix with helical axis oriented along the positive z-axis from 5' to 3' direction was divided into two sub-helices, denoted the 5' and 3' helices, and with the angle between the 5' helix closing base pair y-axis and the molecular y-axis, T_h, set to 51.1° (5). The entire system was translated such that the first phosphorous atom belonging to the 5' helix on the strand oriented along the negative z-direction from the 5' to 3' end was located at the origin, serving as the rotation pivot point. A given (α_h, β_h, γ_h) inter-helical orientation was generated by applying the Euler rotation $R_Z(-\gamma_h)$ *x* $R_Y(-\beta_h)$ *x* $R_Z(-\alpha_h)$ to the 5' helix while keeping the 3' helix fixed, where R_Z and R_Y refer to Euler rotations of an object about the Z and Y axis respectively (5). Rotations were performed for all non-degenerate (5) permutations of (α_h, β_h, γ_h) on a 5° grid. Conformations that resulted in inter-helix collisions, taken as an inter-helical atomic distance of <1.4 Å excluding hydrogen atoms and O1P and O2P atoms connected to the pivot P atom, were classified as sterically disallowed. If the distance between the last O3' atom of the 5' helix and first P atom of the 3' helix at the division between the two helices on the strand opposite the pivot is greater than the linker cutoff of 7.26 Å per bulged nucleotide than the conformation was classified as connectivity disallowed. This corresponds to a cutoff of 7.26 Å, 14.52 Å, 21.78 Å, and 29.04 Å when calculating connectivity constraints for S_1S_0, S_2S_0, S_3S_0, S_4S_0 bulges respectively. The fraction of space allowed for a given junction is computed by dividing the number of allowed conformations by the number of unique (α_h, β_h, γ_h) permutations, 191,736.

2.2.2 Computing topologically allowed space with translational relaxation

We computed the topologically allowed conformations when allowing for some degree of interhelix translation that may relieve interhelix steric collisions. Translational degrees of freedom were defined as deviations in the x, y, and z coordinates of a 3' helix relative to the position of an idealized reference 3' helix that has the same (α_h, β_h, γ_h) orientation. Analysis of crystal structures of S_1S_0, S_2S_0, S_3S_0, and S_4S_0 junctions in our database (see **Section 2.2.5**) revealed that these translational deviations are roughly isotropic in x, y, and z, centered near 0, and on average $\sqrt{x^2 + y^2 + z^2} \leq 2$ Å. We therefore allowed 1-2 Å of translational freedom in each of the x, y, and z directions to be representative of that found in the PDB. To simulate the effects that these extra degrees of freedom may have to relieve inter-helix steric collisions we modified our rotation protocol. As in the translation-free procedure, a rotation $R_Z(-\gamma_h)$ *x* $R_Y(-\beta_h)$ *x* $R_Z(-\alpha_h)$ is applied to the 3' idealized helix, iHII, while keeping the 5' helix, iHI, fixed. Following each rotation, iHII was translated by all combinations of -1, 0, and 1 Å in the x, y, and z directions for a 1 Å cut-off, or by all combinations of -2, -1, 0, 1, and 2 Å for a 2 Å cut-off. This procedure captures >80% of the inter-helix translations observed in the PDB. A given inter-helical angle was accepted provided one of the translations alleviated any existing steric collisions. Connectivity allowed conformations were taken to be the same those for the translation free case.

2.2.3 Computing the topologically allowed space with non-idealized helices

Eighteen 8-bp continuous helices of different sequences were chosen from junctions returned from a search of the RNA FRABASE (6) and were divided into 2×4-bp helices, named the 5' and 3' helices respectively (**Table A.1**). The helices were chosen such that all combinations of WC pairs and two GU wobble pairs were represented at the 'junction' between the two 4-bp

helices. To remove any inter-helical bends or twists the 5' and 3' helices of each 2×4-bp system were individually superimposed on top of their corresponding helices in the idealized A-form system used in the idealized topologically allowed computations. In cases where the pivot P atom was not located at the origin after the superpositions a translation was applied to both 5' and 3' helices such that the pivot was at the origin as desired. After these 'non-idealized' systems were prepared the same rotation procedure used for the idealized helices was used.

2.2.4 Cylinder rotations

An in-house Python script was used to generate two cylinders of diameter 17.5 Å, height 24 Å, and surfaces represented by a ~1.5 Å square mesh (distance between points was closer on cylinder ends). The chosen diameter of 17.5 Å corresponds to the distance between P(i) and O3'(j) across the idealized A-form helix in Watson-Crick base-paired residues i and j. The bottom cylinder axes were vectors originating at (8.75,0,-0.75) and (0,0,0) and running along the negative z direction for the bulge and center connected systems, respectively. The top cylinder axes began at (8.75,0,0.75) for the bulge and (0,0,14.52) for the center connected systems and ran along the positive z direction. Orientations that resulted in mesh points from opposing cylinders being <1.5 Å apart were determined to be sterically disallowed. Connectivity constraints for the bulge-like system were implemented by requiring that the distance between the linker termini, (17.5,0,-0.75) and R($-\gamma_h$,$-\beta_h$,$-\alpha_h$) *x* (17.5,0.0.75), were less than 14.52 Å. The linker-length of the center-connected cylinder system is invariant across all rotations and is thus free of connectivity constraints.

2.2.5 PDB survey of two-way junctions

A search of RNA two-way junctions was performed using the RNA FRABASE (6) on June 25, 2010 for $H_3S_1H_3S_0$, $H_3S_2H_3S_0$, $H_3S_3H_3S_0$, and $H_3S_4H_3S_0$ bulge motifs; on July 27, 2010 for $H_3S_1H_3S_1$, $H_3S_2H_3S_1$, $H_3S_3H_3S_1$, $H_3S_4H_3S_1$, $H_3S_2H_3S_2$, $H_3S_3H_3S_2$, $H_3S_4H_3S_2$, $H_3S_3H_3S_3$, $H_3S_4H_3S_3$, and $H_3S_4H_3S_4$ internal loop motifs; and September 22, 2010 for all H_6H_6 helices (or 6-bp continuous helices) as described previously (1). Inter-helical orientations for each identified junction were measured using in-house Perl scripts and the (α_h, β_h, γ_h) convention of Bailor *et al.* (1,5,7). Unless otherwise noted, angles for NMR structures were measured on the first model reported in the ensemble. The less than 2% of junctions identified by the RNA FRABASE whose flanking helices were >2 Å in backbone RMSD from idealized A-form, had a strand possessing all deoxyribose sugars, or had non-continuous chain connectivity at the junction were excluded from our analysis. Measured orientations are designated as falling within the topologically allowed space if the shortest distance between the (α_h^i, β_h^i, γ_h^i) measured for conformation *i* and (α_h^T, β_h^T, γ_h^T) for topologically allowed point *T*, $d = \sqrt{\left(\alpha_h^i - \alpha_h^T\right)^2 + \left(\beta_h^i - \beta_h^T\right)^2 + \left(\gamma_h^i - \gamma_h^T\right)^2}$, is <5°. Bulge-linker distances were calculated along the bulge strand between the O3' atom of the 5' helix closing base pair and the P atom of the 3' helix's first base pair with an in-house Perl script (1).

In-house Perl scripts were also used for the determination of the inter-helical translations of each junction by adapting the procedure used to determine inter-helical angles. In obtaining the (α_h, β_h, γ_h) for a junction a reference helix, denoted iH2', is superimposed on top of the 3' helix and then this iH2' is transformed to an orientation that is perfectly coaxially stacked upon the 5' helix (5). This transformation is done by first translating the center of mass of iH2' to the origin, performing a rotation, and then translating the rotated iH2' from the origin to its coaxially

stacked position. The difference between these first and second translations define a net translation of the center of mass of iH2' from its coaxially stacked position. This measured net translation is composed of two terms, the movement of the center of mass due the rotation of the helix, and the deviation of the helix away from this 'pure rotation' position. We calculated the 'pure rotation' position of a given (α_h, β_h, γ_h) by applying a rotation of ($-\gamma_h$, $-\beta_h$, $-\alpha_h$) to the 3' helix of an idealized, coaxially stacked helical system, and then determined the net translation incurred when transforming this 'pure rotation' conformation back to its original coaxially stacked state. The difference between the net translation measured for the PDB conformation and net translation of the 'pure rotation' conformation was taken as the translational deviation of the junction.

2.2.6 FR3D analysis of internal loop motifs

FR3D (Find RNA 3D) (8) pair wise interaction files that were available for PDBs within our database were downloaded from the FR3D website at http://rna.bgsu.edu/FR3D/AnalyzedStructures/**PDB**/**PDB**_interactions.html on August 23, 2010, where **PDB** is replaced with the name of the PDB structure of interest. The interaction files were parsed with in-house Python scripts to search for base pairing and stacking interactions. Bases were assigned as 'looped-in' if FR3D indicated that they possessed a stacking interaction with residues of either strand of the junction or with the bases comprising the closing base pairs of the flanking helices. The <2% of junctions without available FR3D files were excluded from the analysis.

2.2.7 Identification of tertiary interactions

Intra-RNA tertiary interactions were indentified by searching the FR3D interaction files for pairing or stacking interactions between junction or closing base pair nucleotides and non-junction nucleotides with an in-house Python script. Junctions without available FR3D files were by default classified as having tertiary contacts. A separate in-house python script was used to search for protein, ligand, and RNA tertiary interactions that were not identified by FR3D. PDB files corresponding to each junction in our database were searched exhaustively for any C, N, S, O, or P atoms within 4 Å of junction or closing base pair atoms while excluding atoms from the chain(s) containing the junction and all 'HOH' O atoms.

2.2.8 Manual annotation of $H_3S_{\geq 2}H_3S_2$ junctions

Crystal structures of $H_3S_{\geq 2}H_3S_2$ junctions within our database were sorted according to sequence and inter-helical angles. Junctions with identical sequences and within an 11.5° distance in (α_h, β_h, γ_h) space were clustered together and assumed to have common features. One junction from each group was then manually inspected and each loop base was qualitatively classified as to whether it was non-canonically base paired, 'near non-canonically paired', or neither, and whether it was stacked into the junction. Base paired residues were defined as two bases from opposing strands having at least one hydrogen bond donor to acceptor distance of ≤3.6 Å and angle ≥125°. Near non-canonical base pairs were defined as two bases from opposing strands that were roughly coplanar and looped into the junction but did not possess clear hydrogen bonds. Tertiary contacts, defined as protein, ligand, crystal or other non-junction RNA hydrogen bonding contacts with any junction or closing base pair base (excluding backbone contacts) were also recorded.

2.3 Results and Discussion

2.3.1 Sterics

We performed an exhaustive search of the inter-helical conformational space accessible to a bulge by carrying out a series of rigid body rotations to two, four base-pair long, idealized A-form helices. In **Figure 2.1B**, we show the inter-helical orientations allowed after excluding all conformations that possess inter-helix steric clashes. The angles α_h and γ_h, denoting the twists of the 3' and 5' helices about their respective helical axes, and β_h, the inter-helical bend angle, are plotted as 3×2D projections of the 3D (α_h, β_h, γ_h) space of inter-helical conformations (5). Of the theoretically possible (α_h, β_h, γ_h) orientations (5), 53% are excluded by steric constraints alone. Visual inspection of the projections also reveals that the allowed orientations are asymmetric in α_h and γ_h. Calculations using perfectly symmetric cylinders (**Figure A.1**) reveal that the asymmetry of the sterically allowed space arises from the inherent asymmetry and chirality of the helices.

A notable consequence of the anisotropy of the sterically allowed space is that it encodes a specific directionality to any bending that occurs between the two helices (**Figure 2.1C**). Steric collisions between the closing base pairs of the two helices make it such that the density of accessible states is much greater for bends that correspond to movement of the 3' helix into the space above the plane of the 5' closing base pair and 'away' from the 5' helix, rather than into the closing base pair plane and 'towards' the 5' helix. This result provides a quantitative and simple explanation for the directionality of RNA and DNA bulge-induced bends (9-13).

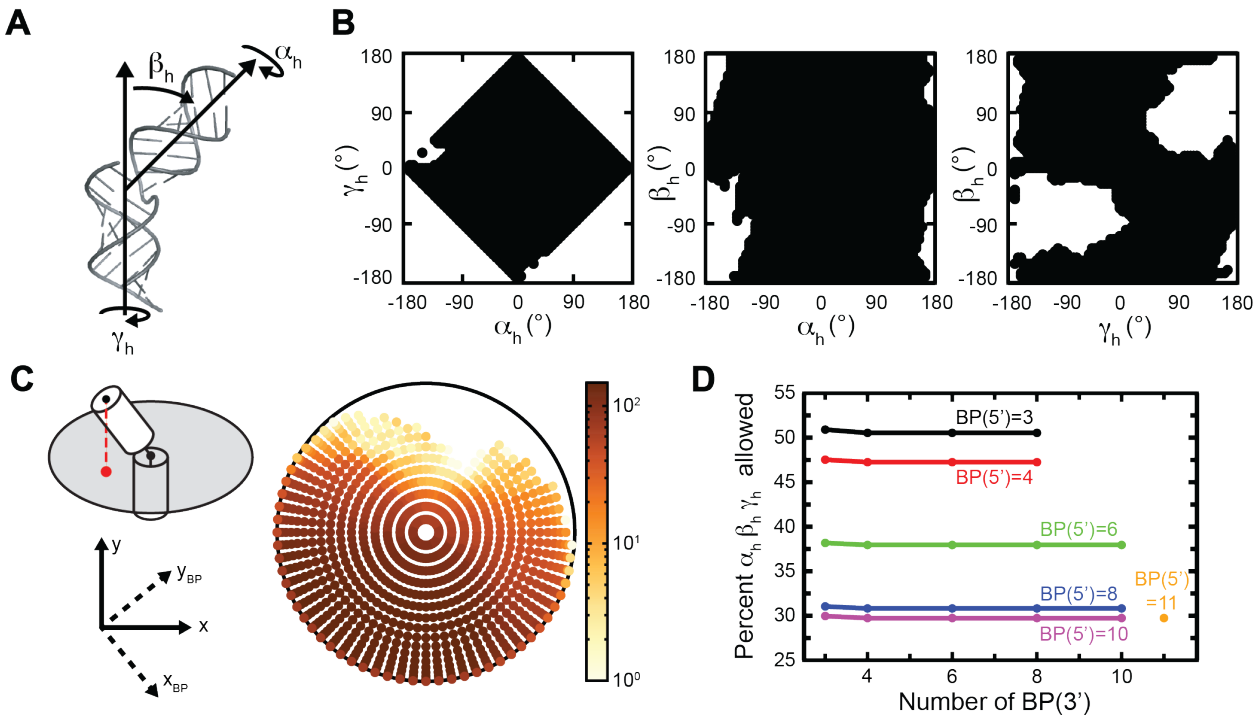

Figure 2.1: Sterically allowed inter-helical conformations
(A) (α_h, β_h, γ_h) convention used to describe inter-helical orientations and **(B)** α_h-γ_h, α_h-β_h, and γ_h-β_h projections of the conformations allowed given the steric constraints. **(C)** Density of allowed conformations of the upper (3') helical axis in relation to the fixed lower (5') helical axis projected in the plane perpendicular to the lower helical axis. A cartoon of the relationship between a helix conformation and a projected point as well as the coordinate frame of the closing bp of the 5' helix (14) in relation to the coordinate frame of the projection is shown for reference. **(D)** Fraction of (α_h, β_h, γ_h) space sterically accessible as a function of length of the 5' and 3' helices.

Interestingly, the sterically allowed orientations are also dependent on helix length. While 53% of conformations are excluded with 4-bp helices, the fraction excluded increases to 62% and 69% for 6-bp and 8-bp helices respectively and decreases to 49% with 3-bp helices. The steric constraints only depend on the length of the 5' helix when the 3' helix has >3 base pairs (**Figure 2.1D**). This can be attributed to the fact that the helices are asymmetric about the rotation origin (pivot P atom), which is located closer to atoms in the 5' helix as compared to the 3' helix. As a result, elongation of the 5' helix results in steric occlusion whereas the new volume occupied by elongation of the 3' helix falls outside of the sphere swept by the 3' helix. Thus, varying the length of the 5' helix can in principle modulate the allowed inter-helical space.

The above steric model assumes an idealized A-form helix geometry. However, RNA helices can deviate from the canonical geometry and this in turn can affect the sterically allowed conformations. To examine these contributions, we computed the sterically allowed conformations when using 18 non-idealized 2×4 base-pair helices taken from the PDB (**Table A.1**). Each sequence had different closing base pairs at the site of the junction, one representative for each of the 16 unique canonical base pairs, and two sequences possessing a GU wobble pair at the junction. The fraction of sterically disallowed conformations varied between 35%-50% and on average 44% of the total space, in good agreement with 53% when using idealized A-form helices. On average, 97% of the orientations that are allowed when using idealized A-form helices are also allowed when using non-idealized helices, whereas an average of 82% of the orientations sampled by the non-idealized helices are accessible to idealized A-from helices. Thus, local distortions in the helix geometry can slightly increase the space that can be sampled and allow access to otherwise sterically disallowed inter-helical conformations.

Our model also assumes that the two helices are fixed at the pivot point. However, some of the steric constraints can in principle be relaxed through translations of the helices relative to one another. Analysis of the bulge motifs within our structural database revealed that the average net translation from the origin is ≤2 Å. If the sterically allowed space is recalculated with an allowance for ~1 Å translational relaxation the fraction of sterically allowed conformations increases from 47% to 68% and with ~2 Å to 80%. Studies that apply a more sophisticated treatment of the coupling between rotational and translational degrees of freedom are needed to clarify the significance of inter-helix translations in determining junction conformation.

As a whole, these results suggest that steric constraints exclude 20-53% of the possible inter-helical orientations. However, we emphasize that these values represent a lower bound to

the degree of steric confinement in RNA junctions. The heavy atom van der Waals diameter of 1.4 Å used in our models is well below commonly accepted values of 3-4 Å. Inclusion of hydrogens and the solvation shell will also increase the effective van der Waals radii. Finally, it is likely that electrostatic repulsion may serve to further narrow the space (15).

2.3.2 Bulge connectivity constraints

The set of conformations that are accessible to a bulge junction is also subject to connectivity constraints arising from the finite length of the bulged nucleotides linking the helices. We approximate these constraints by requiring that the end-to-end distance between the last O3' atom of the lower (5') helix and the first P atom of the upper (3') helix to be less than a given cutoff length. Bailor *et al.* previously used a bulge length cut-off equal to 4.9 Å per nucleotide, which was assumed to be the maximum length of a given residue (1). Here, we compute the average per bulge residue distance length for all $H_3S_{4\geq X\geq 1}H_3S_0$ junctions within our database. The computed distribution is roughly Gaussian with a mean of 5.78 Å and a standard deviation of 1.48 Å (**Figures A.2, A.3**). In the new calculations, we use a cutoff of 7.26 Å per bulged nucleotide, or the mean plus one standard deviation, which accommodates 84% of the per-nucleotide bulge lengths found within the PDB. With the distance cut-off the connectivity constraints alone exclude 95%, 83%, 62%, and 31% of the possible orientations for S_1S_0, S_2S_0, S_3S_0, S_4S_0 bulges, respectively (**Figure 2.2A**).

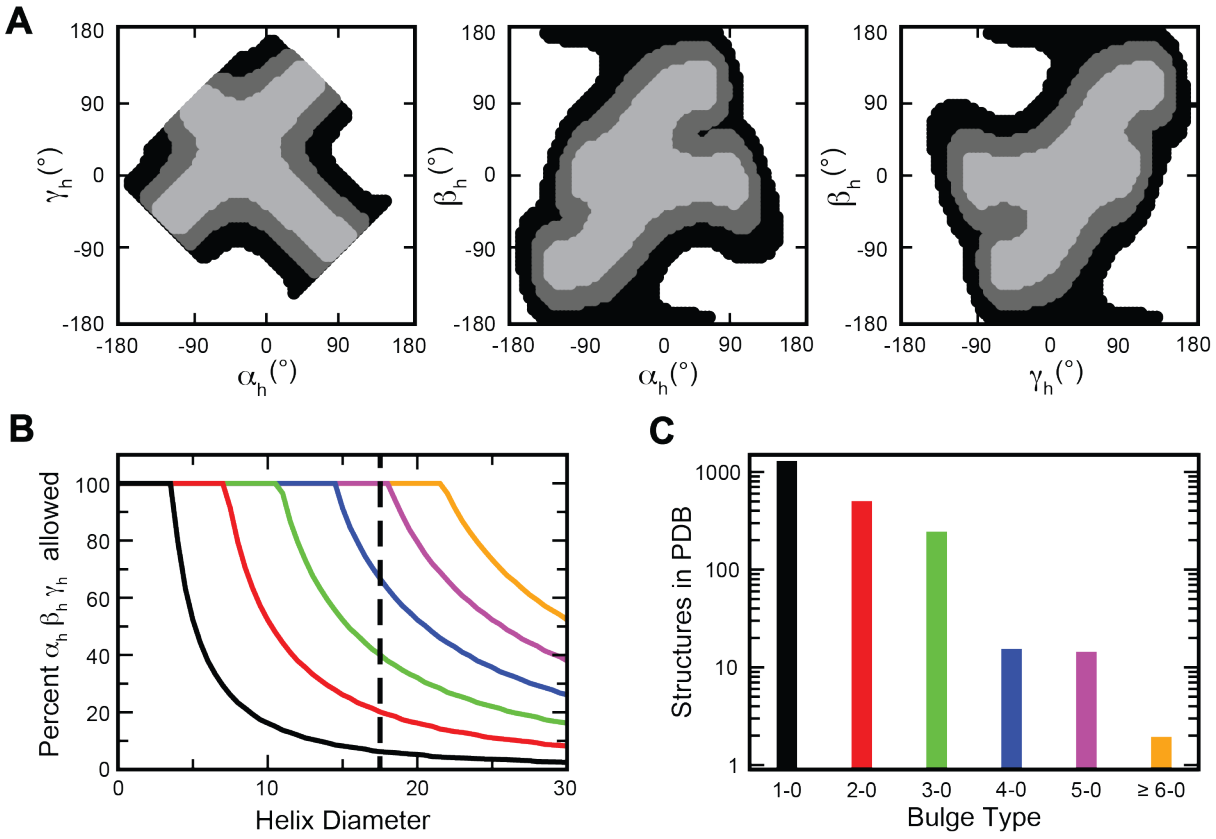

Figure 2.2: Connectivity allowed inter-helical conformations.
(A) The set of conformations allowed given the connectivity constraints of one (light grey), two (dark grey), and three (black) bulged nucleotides, where the allowed conformations for *i* number of bulged nucleotides include those for *i-1, i-2, ...1* bulged nucleotides. **(B)** The percentage of (α_h, β_h, γ_h) orientations that are connectivity allowed as a function of effective helix diameter for a linker cutoff of one (black), two (red), three (green), four (blue), five (orange), or six (magenta) bulged nucleotides. The effective linker diameter of an A-form helix, 17.5 Å, is marked with a vertical dashed line. **(C)** The number of S_1S_0 (black), S_2S_0 (red), S_3S_0 (green), S_4S_0 (blue), S_5S_0 (orange), and S_6S_0 (magenta) structures in PDB.

Our analysis does not account for intra-linker and linker-helix steric clashes. As a first estimate of the impact of such steric effects, we identified orientations where the line segment drawn between the O3' and P atoms (the linker termini) was less than 2 Å from either helix and within 2 Å of the cutoff distance, meaning a bulge connecting the two termini would have no alternative path. This analysis suggests that 0%, 1%, 2%, and 4% of the connectivity allowed conformations may be inaccessible due linker-helix steric overlaps for S_1S_0, S_2S_0, S_3S_0, and S_4S_0 motifs respectively. When considering both steric and connectivity constraints, as discussed in

greater detail in following sections, 0%, 2%, 8%, and 11% of the S_1S_0, S_2S_0, S_3S_0, and S_4S_0 allowed conformations are estimated to contain linker-helix steric overlaps.

The connectivity constraints, and their variation with bulge length, depend critically on the ratio of the per-nucleotide length of the bulge, which is determined by the backbone structure, to the helix diameter, as determined by base-pair structure. In **Figure 2.2B**, we plot the results of a derived mathematical relationship (**Appendix A.1**, **Figure A.4**) that relates the set of accessible (α_h, β_h, γ_h) angles to the number of bulged nucleotides and the 'helix' diameter. At the effective diameter of the A-form helix, increasing the number of bulged nucleotides results in incremental increases in the size of the allowed space up to five nucleotides, above which there are no connectivity constraints. However, for a hypothetically smaller helix diameter, there are fewer, more coarsely spaced putative levels. Conversely, there are greater, more finely space levels for larger helix diameters, underlining the unique tunability of RNA junction conformation that is endowed by RNA's chemical structure.

A strong prediction of **Figure 2.2B** as well as the computed topologically allowed space is that while the accessible space increases with bulge length, it plateaus at five nucleotides. This prediction is in general agreement with the FRET experiments of Gohkle et al (16) where the FRET efficiency between probes located at the ends of the two helices increased as the number bulged nucleotides increases from one to five but then changes very little for bulges that are seven and nine nucleotides long. Additionally, gel electrophoresis experiments also indicate a convergence to a common electrophoretic mobility for bulges that are longer than six nucleotides (10,17,18). Interestingly, a search of bulge motifs in PDB with the RNA FRABASE (6) reveals a marked scarcity of bulges with greater than five nucleotides (**Figure 2.2C**). This suggests that

bulges primarily play the role of tuning the inter-helical orientation and there is rarely reason to have bulges >5 nucleotides long.

2.3.3 Interhelical correlations

A notable feature of **Figure 2.2A** is that the inter-helical angles do not vary independently of one another; rather we observe two linear correlations, one positive and one anti, in the α_h-γ_h projection. Thus, each helix can only twist about its axis a certain amount, generally determined by the length of the linker, before the other helix must compensate with a correlated twist or bend to maintain linker connectivity (**Figure A.5**). This codependence can be attributed to the unique topology of the nucleic acid two-way junction where the helices are linked at two positions around the helix circumference. By comparison, an artificial system in which helices are connected through a single, centrally located linker exhibits no correlations between α_h and γ_h (**Figure A.1**).

The source of this correlation can be found in the mathematical relationship linking the linker-length, L, to the fixed helix diameter, D, and the applied (α_h, β_h, γ_h) rotation transformation (**Appendix A.1**, **Figure A.4**):

$$\frac{L^2}{D^2} = 2 - 2\cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) + 2\sin(\alpha_h)\sin(\gamma_h) \qquad \textbf{[2.1]}$$

A derivative of **[2.1]** with respect to an arbitrary time, t, while maintaining a constant inter-helical bend, β_h, and constant linker-length, shows that changes in α_h must be offset by corresponding changes in γ_h through the equality

$$\frac{d\alpha_h}{dt} = \frac{\sin(\alpha_h)\cos(\gamma_h) + \cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h)}{-\cos(\alpha_h)\sin(\gamma_h) - \sin(\alpha_h)\cos(\beta_h)\cos(\gamma_h)}\frac{d\gamma_h}{dt} \qquad \textbf{[2.2]}$$

highlighting the interdependence of these quantities. In the proper limits, **[2.2]** reduces to

$$\frac{d\alpha_h}{dt} = \pm \frac{d\gamma_h}{dt} \qquad \textbf{[2.3]}$$

and changes in α_h, which denotes a twist of the 3' helix, must be exactly offset corresponding changes in γ_h, or twist of the 5' helix, and explaining the positive and negative α_h-γ_h correlations in **Figure 2.2A.**

Generally, we do not expect β_h and linker length to remain fixed; rather all variables will be sampling values throughout their allowed range. Under this scenario, α_h and γ_h are no longer directly correlated as a change in α_h or γ_h could be offset by a change in the linker length or β_h. However, the limited range of the linker length and a limited β_h range due to steric constraints ensure that the system will experience the α_h-γ_h constraints, giving rise to on-average correlated movements of the two helices as has been previously observed from experiment (1,3). These correlations will be particularly strong in junctions with smaller bulges where the range of accessible linker lengths and β_h's is more significantly restricted.

2.3.4 Union of bulge connectivity and sterics

The full range of accessible inter-helical conformations is restrained by both connectivity and steric constraints. In **Figure 2.3A** we show the intersection of both of the spaces. Visual comparison between **Figure 2.2A** and **Figure 2.3A** reveals that sterics and connectivity exclude distinct parts of conformational space. For the case of S_1S_0 and S_2S_0 bulges, 95% and 83% of the space is disallowed when considering connectivity alone, yet when combined with the steric constraints, 98% and 95% respectively of the space is excluded. While the connectivity constraints are symmetric, this symmetry is broken through the imposition of the steric constraints. Conspicuously, the negative α_h-γ_h correlation for the S_1S_0 and S_2S_0 bulges

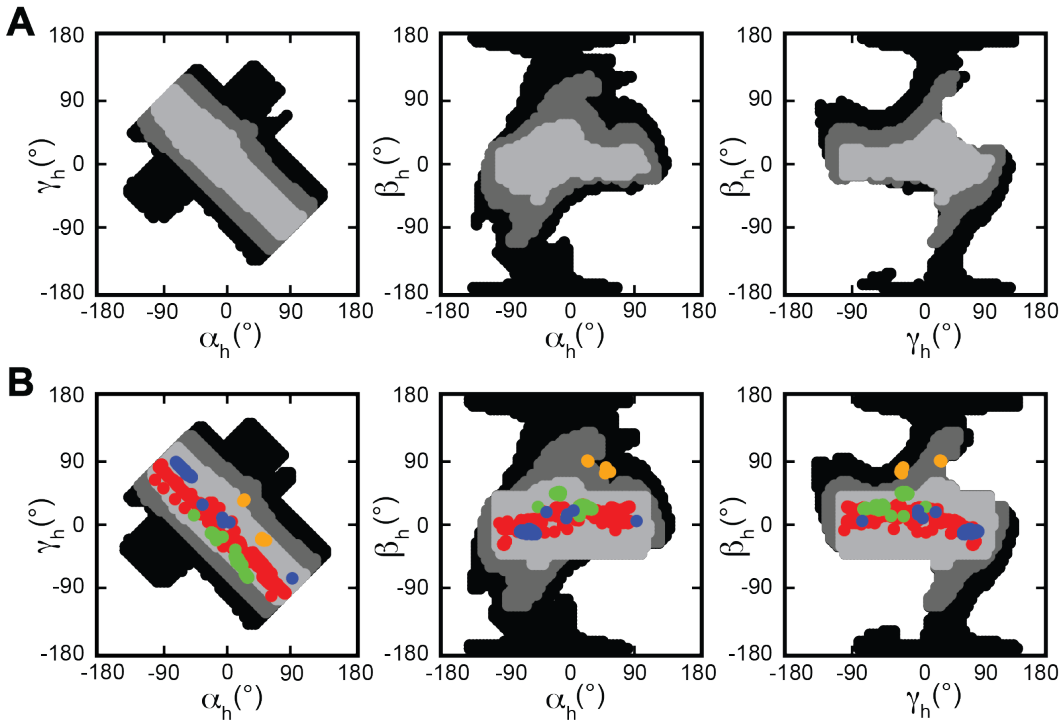

Figure 2.3: Union of steric and connectivity constraints
(A) Plane projections of the set of interhelical conformations satisfying both the steric and connectivity constraints for one (light grey), two (dark grey), or three (black) bulged nucleotides, where allowed conformations for *i* number of bulged nucleotides are inclusive of those for *i-1, i-2, ...1* bulged nucleotides. **(B)** Projections of the total topologically allowed space, obtained by taking the union of that in **(A)** with those conformations accessible to continuous helices and an additional 5° error. Interhelical conformations measured for 728 $H_3S_1H_3S_0$, 160 $H_3S_2H_3S_0$, 65 $H_3S_3H_3S_0$, and 10 $H_3S_4H_3S_0$ crystal structure junctions in the PDB are plotted in red, green, blue, and orange respectively.

disappears and the range of accessible β_h angles decreases; this implies that bends that maintain or shorten the distance between the two closing base pairs, and therefore preserve a short linker length, are precluded due to steric collisions, whereas bends that result in an opening of the space between the pairs, and are accessible to bulges with longer linker lengths, are not.

Nevertheless, this picture of the topological constraints is incomplete – A-form helices themselves possess intrinsic orientational degrees freedom that are independent of the constraints imposed by the bulge. Analysis of all coaxially stacked 2×3 bp helices, or all continuous 6-bp helices, in the PDB revealed that the (α_h, β_h, γ_h) values for these systems varied with -

$20 \leq \zeta = \alpha_h + \gamma_h \leq 20$ and $-25 \leq \beta_h \leq 35$ (**Figure A.6**). To obtain a complete topologically allowed space we take the union of those conformations that satisfy either these constraints or those calculated for the bulge topology of interest with an additional 5° in each angle added to the bounds of this space. This 5° padding accounts for the errors inherent in the measurement of an arbitrary (α_h, β_h, γ_h) angle from the PDB as well as for capturing conformational states that may be accessible due to deviations from idealized A-form helical structure.

To test our framework we measured how well we could capture the conformations found in the PDB. In **Figure 2.3B** we plot the predicted allowed conformations overlaid with conformations measured for the 963 non-GU closing base pair crystal structure junctions in our database. **Table 2.1** summarizes the results for the both the fraction of PDB structures that fall within the predicted idealized allowed space, and the fraction of possible (α_h, β_h, γ_h) states that the space is confined to. It is remarkable that our model samples as little as 5% of the allowed space, and on average 18%, and yet is able to accommodate 100% of all known bulge structures. These results compare favorably to our prior study (1) where the topologically allowed space was confined to 4%, 8%, 14%, and 20% of possible states for S_1S_0, S_2S_0, S_3S_0, and S_4S_0 motifs

Bulge type	Fraction of PDBS sampled (total)	Idealized (α_h, β_h, γ_h) allowed	Range of non-ideal (α_h, β_h, γ_h) allowed (mean)	Translation (α_h, β_h, γ_h) allowed (w/o 5° err)
S_1S_0	1.0 (728)	0.048	0.047-0.055 (0.049)	0.062 (0.028)
S_2S_0	1.0 (160)	0.094	0.090-0.10 (0.098)	0.14 (0.093)
S_3S_0	1.0 (65)	0.19	0.20-0.24 (0.22)	0.33 (0.23)
S_4S_0	1.0 (10)	0.37	0.38-0.45 (0.42)	0.61 (0.49)

Table 2.1: Statistics of the bulge topologically allowed space
Fraction of crystal structures that fall within the predicted topologically allowed space for non-GU closing base pair crystallographic RNA junctions identified with the RNA FRABASE (6) and fraction of total unique (α_h, β_h, γ_h) permutations allowed for the idealized, non-ideal, and idealized with translation helical systems. Topologically allowed spaces represent the intersection of the sterically and connectivity allowed spaces with the addition of the (α_h, β_h, γ_h) accessible to a continuous helix and 5° error padding in each angle.

respectively, accommodating ~84% of 1195 cryo-EM, NMR, and crystal structure junctions. The differences in the degree of confinement for the larger bulge sizes are a function of the longer linker length used in the current model (7.26 Å vs. 4.9 Å per nucleotide).

While we have focused on X-ray structures in the above analysis, 100% and 97% of the 73 cryo-EM and 58 NMR junctions without GU closing base pairs, respectively, fall within the predicted region (**Figure A.7**). The two NMR outliers (19) are from the same set of experimental restraints and have α_h β_h γ_h angles 12° and 16° away from the closest predicted allowed point. A MolProbity (20) analysis of the structures revealed that there are significant steric clashes and deviations from accepted stereochemistry at the site of the S_3S_0 junction in question, suggesting that structural inaccuracies may be the cause of this disagreement. We also excluded junctions containing GU closing base pairs due to their propensity to induce significant local distortions to the base pairing geometry. Stand alone analysis of all 899 GU-containing junctions showed that 87% fall within our predicted ranges, where those falling outside are on average 9° from the nearest predicted point (**Figure A.7**). This suggests that sequence induced helix distortions can play a role in modifying the topologically allowed space. Indeed, 100% of the junctions containing GU pairs fall within at least one of two allowed distributions generated from non-idealized GU closing base pair sequences. However, it should be noted that comparison between these distributions and the idealized derived distribution share the majority of their points.

An interesting observation of **Figure 2.3B** is that certain regions of the allowed space are rarely, if ever, sampled in the PDB. This includes conformations where helices bend into each other along the major-groove face, which likely leads to unfavorable electrostatic effects. Further refinement of the model with incorporation of electrostatic interactions and larger van der Waals radii, as mentioned above, would likely reveal that such states are largely inaccessible.

2.3.5 Dynamic properties of the bulge topologically allowed space

Prior studies of two-way junction motifs suggest that these systems are highly dynamic on the nanosecond to microsecond timescales (3,21), and in the case of HIV-1 TAR (1,22), roughly sample the entire topologically allowed space. If bulge systems are in fact sampling the entire topologically allowed space one would expect an ensemble average over all conformations to be similar to that observed from experiments with sensitivities less than these timescales. Remarkably, if we assume each conformation is equally weighted and calculate the ensemble average bend, or $|\beta_h|$, expected from the topologically allowed space, we obtain a result that closely replicates the values Zacharias and Hagerman (17) obtained for polyA and polyU bulges using transient electronic birefringence (**Figure 2.4A**). These values are also in agreement with estimates from FRET measurements of RNA of Gohkle *et al.* (16) and in qualitative agreement with studies performed on DNA (16,23,24). A corollary to these results is that the bent state is entropically favored for longer loops. Disregarding the conformational preferences of the loop, there is a significant entropic penalty to forming bent states ($|\beta_h|>\sim60$) for S_1S_0 and S_2S_0 bulges, while the opposite holds true for $S_{X\geq3}S_0$ bulges, where forming a coaxial stack (restricting $|\beta_h|<\sim35$) is disfavored (**Figure 2.4B**). Thus, the varying conformational preferences of different junction topologies can be attributed to secondary structure encoded modulation of the topologically allowed space.

2.3.6 Structural survey of internal loop conformations

Bailor *et al.* (1) previously showed that the behavior of asymmetric internal loops can be explained by transforming a given internal loop into a reduced bulge representation. This is achieved by maximizing stacking and non-canonical base-pairing between residues in the

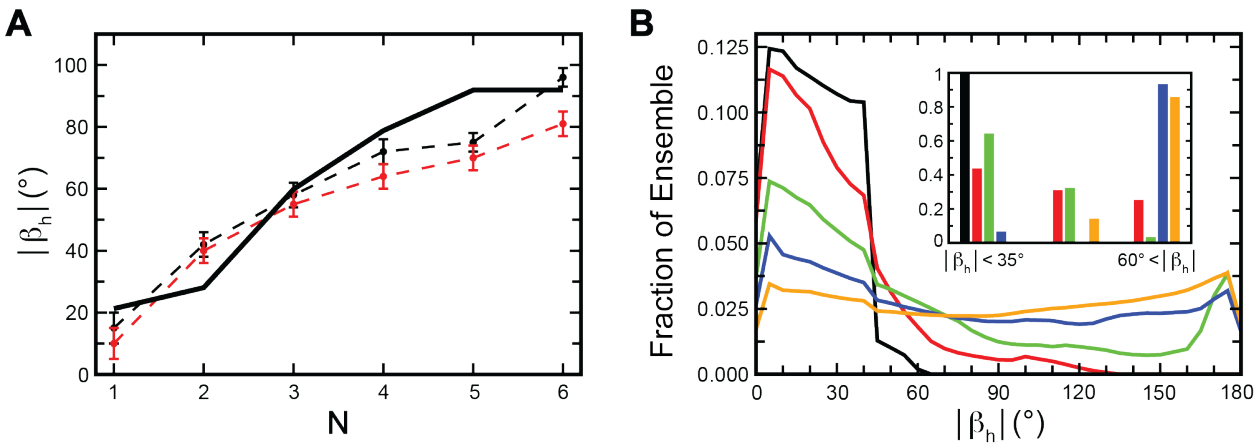

Figure 2.4: Dynamic properties of bulges encoded by topological constraints
(A) The inter-helical bend angle for *N* bulged nucleotides for poly-A (dashed black) and poly-U (dashed red) RNA bulges as measured by Zacharias and Hagerman (17) in the absence of Mg^{2+}, and $|\beta_h|$ averaged over the total topologically allowed space as a function of *N* bulged nucleotides. **(B)** Populations of topologically accessible conformations as a function of $|\beta_h|$ for S_1S_0 (black), S_2S_0 (red), S_3S_0 (green), S_4S_0 (blue), and S_5S_0 (orange) bulges. The inset shows the fraction of structures from the PDB with coaxial ($|\beta_h|<35°$), semi-bent ($35°\leq|\beta_h|\leq60°$), or bent ($|\beta_h|>60°$) conformations for 1253 S_1S_0 (black), 487 S_2S_0 (red), 238 S_3S_0 (green), 15 S_4S_0 (blue), and 14 S_5S_0 (orange) junctions.

internal loop. This 'reduced' bulge motif is subject to the characteristic topological constraints, but with the addition of *Y* base-pairs inserted into the junction. By assuming that each inserted base-pair results in a change in $\alpha_h+\gamma_h$ = -34°, the twist of the A-form helix, one could readily generate the allowed space for asymmetric internal loops by shifting the α_h and γ_h of the reduced bulge set of topologically allowed conformations by *Y**-17°. This simple model was able to explain the observation that the (α_h, β_h, γ_h) measured for asymmetric internal loops mirrored the distribution found in bulges with systematic shifts in α_h and γ_h.

To investigate this assumption of maximum stacking of internal loop residues, we used the FR3D (8) nucleic acid structure analysis software to probe the conformation of nucleotides inside internal loops. Analysis of our PDB library consisting of crystal structures for all S_1S_1, S_2S_1, S_3S_1, S_4S_1, S_2S_2, S_3S_2, S_4S_2, S_3S_3, S_4S_3, and S_4S_4 junctions obtained from the RNA FRABASE (6) reveals a clear preference for bases in the *Y* strand to loop into the junction (**Figure 2.5**). Moreover, in >90% of the cases where all *Y* bases are looped in, at least an

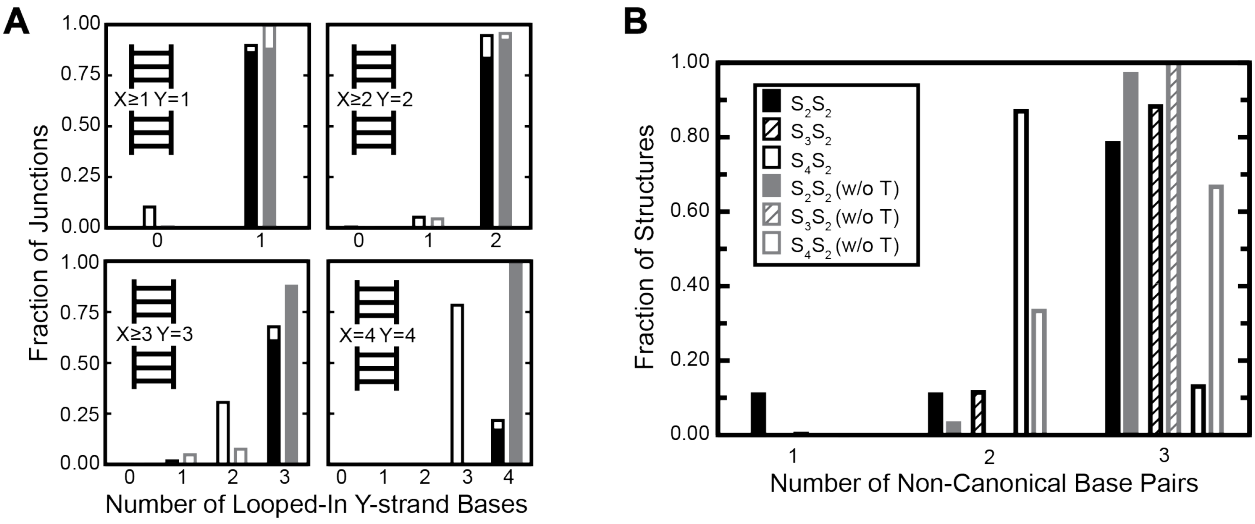

Figure 2.5: Analysis of non-canonical pairing in internal loops
(A) Fraction of looped-in *Y* strand bases determined from FR3D (8) analysis of internal loops for $H_3S_{X\geq1}H_sS_{Y=1}$, $H_3S_{X\geq2}H_sS_{Y=2}$, $H_3S_{X\geq1}H_sS_{Y=3}$, and $H_3S_{X=4}H_sS_{Y=4}$ junctions. Black bars represent all structures within our database for which FR3D analysis files were available and red represent the subset of these structures identified as not having tertiary interactions. The degree to which the rightmost bars (fully looped-in conformations) are filled represents the fraction of those structures that have at least *Y* number of looped-in bases on the *X* strand. Results represent 474 [135] S_1S_1, 468 [97] S_2S_1, 194 [1] S_3S_1, 168 [6] S_4S_1, 315 [103] S_2S_2, 341 [77] S_3S_2, 23 [4] S_4S_2, 631 [74] S_3S_3, 165 [32] S_4S_3, and 153 [5] S_4S_4 motifs where [ ] denote the number structures without tertiary interactions. **(B)** The number of non-canonical and near non-canonical base pairs observed in all $H_3S_{X\geq2}H_sS_2$ crystal structures obtained through a manual annotation of structures. The fraction of structures with 0, 1, or 2 pairs is plotted for S_2S_2 (solid), S_3S_2 (dashed), and S_4S_2 (open) motifs in black, with the subset of those structures free of tertiary contacts shown in grey. There are 323 [97], 349 [46] and 23 [3] S_2S_2, S_3S_2, S_4S_2 motifs respectively, where [ ] denote the number of motifs without tertiary contacts.

equivalent number of *X* strand bases are looped-in as well (filled bars of **Figure 2.5**). Interestingly, the majority of the loops that do not adopt fully looped-in conformations are directly involved in tertiary interactions (see **Section 2.2.7**). Manual examination of the structures that were not fully looped-in and did not have a tertiary contact revealed that all were either influenced by crystal contacts or appeared to be missing coordinates where a protein would be located.

As hypothesized from the observations of Bailor et al (1), we expected that not only would the junctions adopt fully looped-in conformations, but that the looped in bases should form non-canonical base pairs that propagate the flanking helices. To examine what fraction of looped in bases form non-canonical base pairs we hand annotated all S_2S_2, S_3S_2, and S_4S_2

junctions. The results revealed that >80% of the junctions exhibited a maximum number of non-canonical base pairs or near non-canonical base pairs, which we define as cases when bases from opposing strands were looped into the junction and roughly coplanar but did not appear to form clear hydrogen bonds (**Figure 2.5B**). Results for the fraction of bases looped-in obtained from this annotation closely resemble those from FR3D (data not shown). Comparable to the results of **Figure 2.5**, the fraction of structures with two non-canonical or near non-canonical pairs increases to 97% when we only consider structures without tertiary interactions. Of the four outliers where we do not observe maximal pairing and which do not have tertiary interactions, two show maximal pairing in other structures reported of the same junction. Of the remaining two, one is located at the surface of the ribosome and we believe, given the level of disorder of the junction nucleotides, may be experiencing the influence of tertiary interactions not reported in the coordinate file, and the other is bound in close proximity by a protein and under considerable tensile stress. We also note that when removing junctions with tertiary interactions only three $H_3S_4H_3S_2$ junctions remain, one of which is a candidate for having unreported tertiary interactions, making it impossible to draw conclusive findings for this particular motif.

2.3.7 Internal loop topologically allowed space

The results from our structural survey motivate the assumption that the junction residues will fully loop in and maximize non-canonical base pairs. Thus, the topologically allowed space for a S_XS_Y motif is identical to that for a $S_{X\text{-}Y}S_0$ motif shifted by Y*-17° in α_h and γ_h to account for the additional twist of Y non-canonical pairs. In the case of symmetric internal loops, S_XS_X motifs reduce to S_0S_0 motifs and are subject to the same constraints on (α_h, β_h, γ_h) that we observed for continuous helices in the PDB (**Figure A.6**). With the correction for the additional twist of -17°

in both α_h and γ_h per *X* non-canonical base pairs, the S_XS_X allowed space contains all (α_h, β_h, γ_h) satisfying $-25 \leq \beta_h \leq 35$ and $-20-34*X \leq (\alpha_h+\gamma_h) \leq 20-34*X$ on a 5°-grid, with an additional 5° in each dimension to account for alignment errors and translational degrees of freedom.

As shown in **Table 2.2**, this model is able to capture the significant majority of conformations observed within the PDB. When tertiary interactions are excluded, the fraction of PDB conformations sampled is largely 100%, where the exceptions are crystal structures that have crystal packing interactions at the junction, or are very disordered junctions that appear to be missing protein or other coordinates in close proximity to the junction site.

Examination of the tertiary contact free cryo-EM (cEM) internal loop junctions shows that 89% of 75 junctions fall within our predicted region. Of the eight outliers, seven are from a *de novo* cEM-docked model of the ribosome built before a complete crystal structure was available (25), and the other is a cEM-docked homology model of an rRNA segment whose full crystal structure possesses extensive protein interactions (26). Analysis of 83 tertiary contact free NMR internal loops revealed that 85% of the NMR structures agree with our predictions, where outliers were on average ~10° from the nearest allowed conformation. This decrease in performance is a result of outlier junctions adopting conformations that are not fully non-canonically or near-non-canonically paired, or, when paired, propagating with the helical twist that is significantly different than -34°. While it is certainly possible that the more flexible RNAs characterized through NMR adopt conformations not in agreement with our predictions, we note that ten of the twelve outliers, and all of those >10° outside, are structures solved without the use of residual dipolar coupling (RDC) restraints. As recently shown by Summers and colleagues (27) these structures are particularly reliant on the force field (and any deficiencies found therein) used in the structure refinement. That these outliers may be due to structural

inaccuracies is also supported by MolProbity (20) analysis showing that each outlier suffers from steric violations and/or deviations from standard backbone angle stereochemistry around the site of the junction (data not shown). Finally, we note that analysis of the entire structural ensemble of each outlier showed that, expect for one molecule (28), at least several, and on average ~40%, of the reported structures fell within the allowed space, demonstrating that experimental constraints are not precluding agreement with our model.

We note that Bailor *et al.* previously approximated the symmetric loop case using the set of allowed conformations for a S_1S_0 bulge system that was shifted by -17° in α_h and -17° in γ_h per looped base (1). This approximation achieves similar results as those above, but fails to provide a physical model for its success. The reason for this agreement is because of significant overlap between the two allowed distributions – approximately 50% of the points are the same.

Combined, the results of our structural surveys and the success of our internal loop predicted topologically allowed space strongly argues for a 'reduced' bulge model of internal loops in the absence of external tertiary contacts. This view represents a significant simplification in the link between secondary and tertiary structure and suggests an additional level of constraints that may be incorporated in modeling RNA 3D structure from secondary structure (29). In particular, these constraints could be implemented in the *de novo* prediction of smaller RNAs (30,31) that do not possess long-range tertiary contacts, significantly reducing the conformational space that must be searched and thereby limiting computation time. The topologically allowed space might also be useful as a global orientation restraint in the refinement of NMR structures. Finally, in larger RNAs where tertiary contacts are expected to be present, such constraints could be implemented as a function of the proximity of sequence-

Junction type	PDBs sampled (total)	PDBs w/o tert. contacts sampled (total)	(α_h, β_h, γ_h) allowed
S_1S_1	0.99 (482)	1.00 (135)	0.041
S_2S_2	0.79 (323)	0.99 (103)	0.041
S_3S_3	0.94 (641)	1.00 (74)	0.040
S_4S_4	0.97 (155)	0.60 (5)	0.039
S_2S_1	1.00 (482)	1.00 (97)	0.048
S_3S_2	0.99 (349)	0.96 (77)	0.047
S_4S_3	0.26 (165)	0.56 (32)	0.046
S_3S_1	0.95 (198)	1.00 (1)	0.093
S_4S_2	1.00 (23)	1.00 (4)	0.092
S_4S_1	1.00 (170)	1.00 (6)	0.19

Table 2.2: Statistics of the internal loop topologically allowed space
The fraction of internal loop motifs that fall within the predicted topologically allowed space, for all crystallographic junctions indentified with the RNA FRABASE (6), the subset of those junctions without tertiary interactions, and the fraction of the total (α_h, β_h, γ_h) conformations predicted as allowed.

distant RNA to the internal loop in question; on close approach the constraints would be relaxed, allowing greater conformational freedom.

2.3.8 Interplay between tertiary interactions and topologically allowed conformations

Recent studies (2) have put forth the idea that topological constraints can act to preferentially select for the formation certain tertiary contacts over others, thereby providing a mechanism for secondary structure control of RNA folding and 3D structure. Intriguingly, our results suggest that there may be another dimension to the relationship between the secondary structure encoded conformational space and tertiary interactions. An implication of **Table 2.2** is that tertiary interactions can modify the topologically allowed space by looping out junction residues from their default, maximally non-canonically paired conformation. Thus tertiary interactions can affect the global conformation of RNA by modifying the topological constraints encoded by junctions. Several examples of tertiary contacts significantly modifying the conformation of an internal loop can be found in the literature. Most dramatically, the S_2S_2 motif of the L11 protein binding site in the 23S rRNA shifts from a fully non-canonically paired conformation to one

where a base distal in sequence from the junction inserts to the middle of the junction and is stabilized by interactions with L11 (32,33), with similar behavior observed for closely related mutants (34). In the process, the junction is shifted from within the topologically allowed space to >45° outside. Also from the ribosome, the UAA/GAN internal loop is significantly perturbed from its fully non-canonically paired conformation in the free state when bound by proteins (35-37). Examples of RNA tertiary contacts modifying internal loop conformation can be found in the GAAA receptor of *Tetrahymena* ribozyme P4-P6 domain (28,38), and stem A and stem B of the hairpin ribozyme (39-41). In some cases, downstream structure may also act to stabilize transient deviations from what would otherwise be a fully looped-in internal loop conformation, consistent with behavior of the J5/5a hinge of the *Tetrahymena* ribozyme in its folded and unfolded states (38,42-44). It is well known that internal loops serve as important sites in protein binding and intra-RNA tertiary interactions (45-48), and that their sequence-specific ability to recognize their binding partners plays a crucial role in stabilizing specific RNA 3D folds (46,49,50). However, that these tertiary interactions, once formed, may in fact be also enabling the molecule to sample new, otherwise inaccessible inter-helical conformations could offer a new paradigm to understanding RNA folding.

2.4 Conclusions

In this Chapter, we refined the rigid-body heuristic models developed by Bailor *et al.* (1) to provide a more detailed understanding of the role that topological constraints play in governing two-way junction conformation. Our results demonstrate that sterics play a predominant role in encoding both the anisotropy and limited size of the topologically allowed space, constraining two-way junctions to <47% of their theoretically possible interhelical conformations states. Connectivity constraints further limit the number of allowed states in a tunable manner, and are

responsible for giving rise to correlations between the twists of the two junction helices. When steric and connectivity constraints are combined, our refined heuristic model predicts that two-way junctions are limited to <5-40% of their theoretically possible 3D conformations. Significantly, 100% of the interhelical bulge conformations observed in PDB crystal structures that do not have GU closing base pairs are accommodated within our predicted topologically allowed space. Furthermore, the interhelical bend averaged over the topologically allowed space closely matches the interhelical bend measured experimentally in different bulge constructs in the absence of Mg^{2+}. Collectively, these results affirm the importance of topological constraints as a determinant of bulge conformation.

In this Chapter we also explored the dependence of topological constraints on properties such as deviations from ideal A-form structure of the junction helices, potential translational degrees of freedom possessed by the helices, and steric constraints posed by the bulge linker. We demonstrate that each of these properties has the potential to modify the topologically allowed space. Interhelical translations in particular can allow helices to access many conformations that would otherwise be sterically forbidden. Notably, variations in sequence and particularly the presence of GU base pairs at the junction can also allow bulges to access otherwise forbidden conformations.

Finally, an extensive survey of internal loops in the PDB provides strong supporting data for the previous assumption by Bailor *et al.* (1) that internal loops adopt fully non-canonically or near non-canonically paired conformations. This finding justifies that internal loops can be reduced to bulges, and thus experience the same topological constraints as bulges with a simple adjustment for the twist added by the intervening non-canonical pairs. Interestingly, this survey also reveals that long-range tertiary interactions can modify the topologically allowed space by

modifying internal loop pairing, providing a potential mechanism for transducing the formation of a tertiary interaction into a conformational change elsewhere in an RNA molecule.

2.5 References

1. Bailor, M.H., Sun, X.Y. and Al-Hashimi, H.M. (2010) Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation. *Science*, **327**, 202-206.
2. Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S. and Herschlag, D. (2009) Do conformational biases of simple helical junctions influence RNA folding stability and specificity? *RNA*, **15**, 2195-2205.
3. Zhang, Q., Stelzer, A.C., Fisher, C.K. and Al-Hashimi, H.M. (2007) Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, **450**, 1263-1267.
4. Lilley, D.M.J., Clegg, R.M., Diekmann, S., Seeman, N.C., Vonkitzing, E. and Hagerman, P.J. (1995) A nomenclature of junctions and branchpoints in nucleic acids. *Nucleic Acids Res*, **23**, 3363-3364.
5. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) 3D Maps of RNA Inter-helical Junctions. *Nat Protoc*, **6**, 1536-1545.
6. Popenda, M., Blazewicz, M., Szachniuk, M. and Adamiak, R.W. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res*, **36**, D386-391.
7. Bailor, M.H., Musselman, C., Hansen, A.L., Gulati, K., Patel, D.J. and Al-Hashimi, H.M. (2007) Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat Protoc*, **2**, 1536-1546.
8. Sarver, M., Zirbel, C.L., Stombaugh, J., Mokdad, A. and Leontis, N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol*, **56**, 215-252.
9. Tang, R.S. and Draper, D.E. (1994) On the use of phasing experiments to measure helical repeat and bulge loop-associated twist in RNA. *Nucleic Acids Res*, **22**, 835-841.
10. Bhattacharyya, A., Murchie, A.I. and Lilley, D.M. (1990) RNA bulges and the helical periodicity of double-stranded RNA. *Nature*, **343**, 484-487.
11. Tang, R.S. and Draper, D.E. (1990) Bulge loops used to measure the helical twist of RNA in solution. *Biochemistry*, **29**, 5232-5237.
12. Lilley, D.M.J. (1995) Kinking of DNA and Rna by Base Bulges. *Proc Natl Acad Sci USA*, **92**, 7140-7142.
13. Riordan, F.A., Bhattacharyya, A., McAteer, S. and Lilley, D.M. (1992) Kinking of RNA helices by bulged bases, and the structure of the human immunodeficiency virus transactivator response element. *J Mol Biol*, **226**, 305-310.
14. Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. and Westhof, E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*, **31**, 3450-3460.
15. Bai, Y., Chu, V.B., Lipfert, J., Pande, V.S., Herschlag, D. and Doniach, S. (2008) Critical assessment of nucleic acid electrostatics via experimental and computational investigation of an unfolded state ensemble. *J Am Chem Soc*, **130**, 12334-12341.

16. Gohlke, C., Murchie, A.I.H., Lilley, D.M.J. and Clegg, R.M. (1994) Kinking of DNA and RNA Helices by Bulged Nucleotides Observed by Fluorescence Resonance Energy-Transfer. *Proc Natl Acad Sci USA*, **91**, 11660-11664.
17. Zacharias, M. and Hagerman, P.J. (1995) Bulge-Induced Bends in Rna - Quantification by Transient Electric Birefringence. *J Mol Biol*, **247**, 486-500.
18. Zacharias, M. and Hagerman, P.J. (1995) The bend in RNA created by the trans-activation response element bulge of human immunodeficiency virus is straightened by arginine and by Tat-derived peptide. *Proc Natl Acad Sci USA*, **92**, 6052-6056.
19. Diener, J.L. and Moore, P.B. (1998) Solution structure of a substrate for the archaeal pre-tRNA splicing endonucleases: the bulge-helix-bulge motif. *Mol Cell*, **1**, 883-894.
20. Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., III, Snoeyink, J., Richardson, J.S. *et al.* (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*, **35**, W375-383.
21. Olsen, G.L., Bardaro, M.F., Jr., Echodu, D.C., Drobny, G.P. and Varani, G. (2010) Intermediate rate atomic trajectories of RNA by solid-state NMR spectroscopy. *J Am Chem Soc*, **132**, 303-308.
22. Frank, A.T., Stelzer, A.C., Al-Hashimi, H.M. and Andricioaei, I. (2009) Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res*, **37**, 3670-3679.
23. Stuhmeier, F., Hillisch, A., Clegg, R.M. and Diekman, S. (2000) Fluorescence energy transfer analysis of DNA structures containing several bulges and their interaction with CAP. *J Mol Biol*, **302**, 1081-1100.
24. Wozniak, A.K., Schroder, G.F., Grubmuller, H., Seidel, C.A. and Oesterhelt, F. (2008) Single-molecule FRET measures bends and kinks in DNA. *Proc Natl Acad Sci USA*, **105**, 18337-18342.
25. Mueller, F., Sommer, I., Baranov, P., Matadeen, R., Stoldt, M., Wohnert, J., Gorlach, M., van Heel, M. and Brimacombe, R. (2000) The 3D arrangement of the 23 S and 5 S rRNA in the Escherichia coli 50 S ribosomal subunit based on a cryo-electron microscopic reconstruction at 7.5 A resolution. *J Mol Biol*, **298**, 35-59.
26. Halic, M., Gartmann, M., Schlenker, O., Mielke, T., Pool, M.R., Sinning, I. and Beckmann, R. (2006) Signal recognition particle receptor exposes the ribosomal translocon binding site. *Science*, **312**, 745-747.
27. Tolbert, B.S., Miyazaki, Y., Barton, S., Kinde, B., Starck, P., Singh, R., Bax, A., Case, D.A. and Summers, M.F. (2010) Major groove width variations in RNA structures determined by NMR and impact of 13C residual chemical shift anisotropy and 1H-13C residual dipolar coupling on refinement. *J Biomol NMR*, **47**, 205-219.
28. Butcher, S.E., Dieckmann, T. and Feigon, J. (1997) Solution structure of a GAAA tetraloop receptor RNA. *EMBO J*, **16**, 7490-7499.
29. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Curr Opin Struct Biol*.
30. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51-55.

31. Das, R., Karanicolas, J. and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods*, **7**, 291-294.
32. Conn, G.L., Draper, D.E., Lattman, E.E. and Gittis, A.G. (1999) Crystal structure of a conserved ribosomal protein-RNA complex. *Science*, **284**, 1171-1174.
33. Wang, Y.X., Huang, S. and Draper, D.E. (1996) Structure of a U.U pair within a conserved ribosomal RNA hairpin. *Nucleic Acids Res*, **24**, 2666-2672.
34. Shankar, N., Xia, T., Kennedy, S.D., Krugh, T.R., Mathews, D.H. and Turner, D.H. (2007) NMR reveals the absence of hydrogen bonding in adjacent UU and AG mismatches in an isolated internal loop from ribosomal RNA. *Biochemistry*, **46**, 12665-12678.
35. Shankar, N., Kennedy, S.D., Chen, G., Krugh, T.R. and Turner, D.H. (2006) The NMR structure of an internal loop from 23S ribosomal RNA differs from its structure in crystals of 50s ribosomal subunits. *Biochemistry*, **45**, 11776-11789.
36. Lee, J.C., Gutell, R.R. and Russell, R. (2006) The UAA/GAN internal loop motif: a new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions. *J Mol Biol*, **360**, 978-988.
37. Reblova, K., Strelcova, Z., Kulhanek, P., Besseova, I., Mathews, D.H., Nostrand, K.V., Yildirim, I., Turner, D.H. and Sponer, J. (2010) An RNA molecular switch: Intrinsic flexibility of 23S rRNA Helices 40 and 68 5'-UAA/5'-GAN internal loops studied by molecular dynamics methods. *J Chem Theory Comput*, **2010**, 910-929.
38. Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Kundrot, C.E., Cech, T.R. and Doudna, J.A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**, 1678-1685.
39. Cai, Z. and Tinoco, I., Jr. (1996) Solution structure of loop A from the hairpin ribozyme from tobacco ringspot virus satellite. *Biochemistry*, **35**, 6026-6036.
40. Butcher, S.E., Allain, F.H. and Feigon, J. (1999) Solution structure of the loop B domain from the hairpin ribozyme. *Nat Struct Biol*, **6**, 212-216.
41. Rupert, P.B. and Ferre-D'Amare, A.R. (2001) Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature*, **410**, 780-786.
42. Schlatterer, J.C., Kwok, L.W., Lamb, J.S., Park, H.Y., Andresen, K., Brenowitz, M. and Pollack, L. (2008) Hinge stiffness is a barrier to RNA folding. *J Mol Biol*, **379**, 859-870.
43. Takamoto, K., Das, R., He, Q., Doniach, S., Brenowitz, M., Herschlag, D. and Chance, M.R. (2004) Principles of RNA compaction: insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations. *J Mol Biol*, **343**, 1195-1206.
44. Szewczak, A.A. and Cech, T.R. (1997) An RNA internal loop acts as a hinge to facilitate ribozyme folding and catalysis. *RNA*, **3**, 838-849.
45. Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B. and Steitz, T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci USA*, **98**, 4899-4903.
46. Costa, M. and Michel, F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*, **14**, 1276-1285.
47. Draper, D.E. (1995) Protein-RNA recognition. *Annu Rev Biochem*, **64**, 593-620.
48. Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Szewczak, A.A., Kundrot, C.E., Cech, T.R. and Doudna, J.A. (1996) RNA tertiary structure mediation by adenosine platforms. *Science*, **273**, 1696-1699.

49. Tan, E., Wilson, T.J., Nahas, M.K., Clegg, R.M., Lilley, D.M. and Ha, T. (2003) A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc Natl Acad Sci USA*, **100**, 9308-9313.
50. Geary, C., Chworos, A. and Jaeger, L. (2011) Promoting RNA helical stacking via A-minor junctions. *Nucleic Acids Res*, **39**, 1066-1080.

Chapter 3: Development of TOPRNA and the Contributions of Topological Constraints to Bulge Conformational Free Energy

This chapter has been adapted from the following publication:

Mustoe, A. M., Al-Hashimi, H. M., and Brooks, C. L. III. (2014) Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges, *Journal of Physical Chemistry B*, **118**, 2615-2627.

3.1 Introduction

The studies of **Chapter 2** affirmed the importance of topological constraints in limiting two-way junction conformation (1). However, they also highlighted how features such as non-ideal helical structure, interhelical translations, and steric constraints posed by the single-stranded bulge linker may modify the topologically allowed space. Our heuristic models cannot easily capture these properties of RNA junctions. More significantly, our heuristic models are poorly suited for studying the potential thermodynamic contributions of topological constraints. Notably, the multi-scale stochastic dynamics studies by Chu *et al.* (2) on PEG-linked DNA helices indicated that topological constraints destabilize certain junction conformations by more than 5 kcal/mol. Topological constraints can be reasonably expected to make similar contributions to the conformational free energy landscape of RNA bulges. Finally, it is ultimately our goal to characterize the topological constraints of higher-order junctions, and those in RNAs containing multiple junctions. The simplifications inherent to our heuristic models, namely that interhelical

motions are limited to pivoted rigid-body rotations, make them impractical for studying more complex junctions. Thus, it is desirable to develop a new method that can more accurately model topological constraints, provide thermodynamic information, and which can be readily scaled to study arbitrarily complex RNA systems.

Many computational tools have been developed that can potentially satisfy these objectives. All-atom molecular dynamics methods (3,4), and hybrid methods thereof (5,6), offer the highest degree of physical accuracy but are difficult to scale to large systems due to computational expense. Coarse-grained (CG) molecular dynamics methods represent a good compromise (7). Existing CG approaches include: the NAST (5) and YUP (8) models, which use one pseudo-atom to represent each nucleotide; the three pseudo-atom models of Thirumalai and coworkers (9,10), Dokhoylan and coworkers (11,12), and Chen and coworkers (13); the five pseudo-atom models of Levitt and coworkers (14) and Ren and coworkers (15); and the six to seven pseudo-atom HiRE-RNA (16) model. However, none of these CG models are ideally suited to study topological constraints. Notably, one pseudo-atom models cannot fully capture the stereochemical constraints of the RNA backbone. Higher resolution CG models generally allow breaking of secondary structure pairs and include the full suite of RNA forces; while more realistic, these complexities can make it difficult to isolate effects of topological constraints which are dependent only on secondary structure. Moreover, many of these CG models are implemented as specialty codes, which presents development challenges.

In this Chapter, we introduce a new CG model, TOPRNA (TOPological modeling of RNA), which is implemented in CHARMM (17). TOPRNA uses a three pseudo-atom per nucleotide representation similar to that used in preexisting CG models (9,11,13,18), but otherwise differs in that its sole design purpose is to isolate the effects of topological constraints

on RNA structure. Nucleotides participating in canonical base pairs are permanently bonded together and parameterized to maintain helical structures, but all other nucleotides are treated as freely rotatable chains. In addition, attractive interactions involving single-stranded nucleotides and electrostatics are ignored. Thus, biases from non-topological-constraint energy terms are minimized and only negligible energy barriers separate alternative conformations. This approach is similar to that employed by the NAST (5) and YUP (8) models, though these models use one pseudo-atom per nucleotide representations and are primarily optimized for structure prediction applications.

As a first application, we use TOPRNA to reevaluate the role of topological constraints in two-way junction bulge motifs. Through extensive simulations, we corroborate our findings in **Chapter 2** while demonstrating that the greater physical accuracy of TOPRNA captures new aspects of the topological constraints on bulges. Significantly, we also demonstrate that topological constraints encode complex free energy landscapes that appear to play a central role in dictating the 3D conformation and dynamics of bulges.

3.2 Materials and Methods

3.2.1 Model development

TOPRNA uses three pseudo-atoms to represent the phosphate (P), sugar (S), and base (B) moieties of each nucleotide (**Figure 3.1**) (9,11,13,18). The B pseudoatom was taken as a positional average of a base's cyclic nitrogen and carbon atoms; the S pseudoatom an average over the C1', C2', C3', C4', C5' and O4' atoms; and the P pseudo-atom as the phosphorus atom. Given a user-input secondary structure, base-paired nucleotides are permanently bonded together and contiguously paired regions are parameterized to adopt A-form helical structure. By contrast,

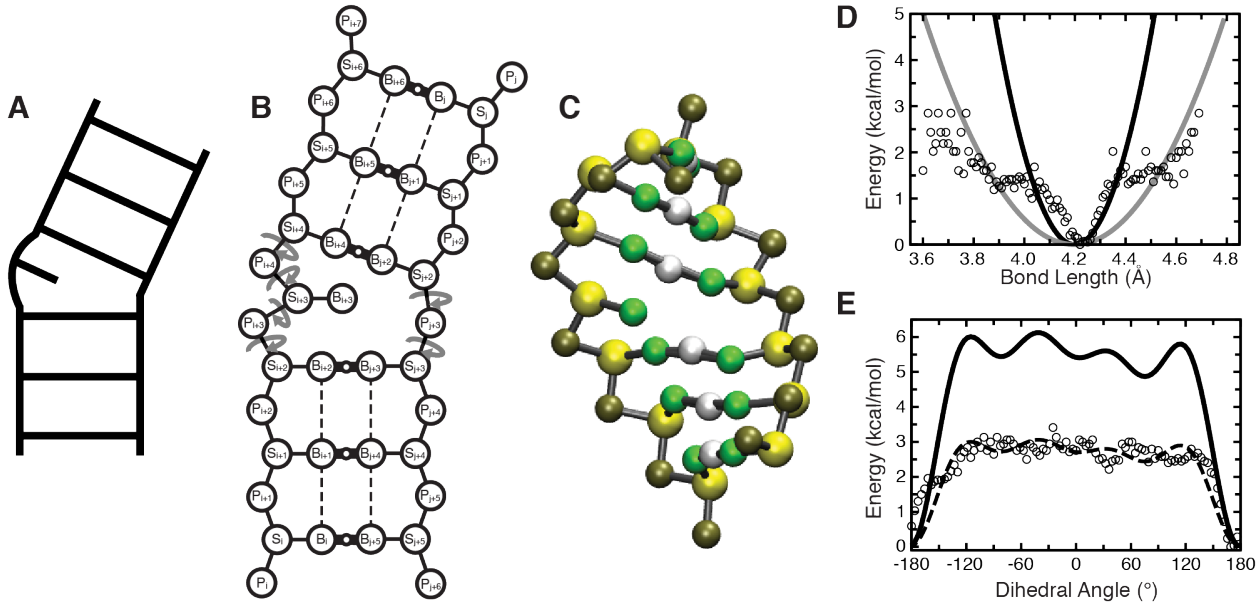

Figure 3.1: Outline of the TOPRNA model
(**A, B, C**) A secondary structure element is shown according to its ladder cartoon, and in 2D and 3D TOPRNA representations. In (**B**), circular arrows denote freely rotatable bonds, thick solid lines with small open circles denote permanent base pair bonds with an accompanying 'M' filler pseudo-atom, and dashed lines denote improper dihedral angles used to maintain helical twist. (**D**) The adenine P_i-S_i bond potential is shown as a representative harmonic potential. Black and grey lines indicate the different potentials used for base-paired and single-stranded nucleotides, respectively, following the observation that these bonds exhibited strong and weak harmonic potential behavior at short and large deviations. Two different potentials were also used for angles involving this bond, but other bonds and angles were assigned the same K regardless of base-pairing status. (**E**) The S_i-P_{i+1}-S_{i+1}-P_{i+2} dihedral potential placed between sequentially paired residues *i* and *i+1* is shown as a representative example. The dashed line indicates the original cosine series fit to the statistical potential, with the solid line indicating the final TOPRNA potential after the *K*'s obtained from the original fit were uniformly doubled. Statistical potentials were calculated by binning every 0.01 Å and 3.6° for bonds and dihedrals, respectively, with unpopulated bins excluded, and are shown using open circles in (**D**) and (**E**).

---

all nucleotides not in AU, GC, or GU pairs are left without backbone dihedral potentials and are freely rotatable (**Figure 3.1B**). Backbone dihedrals that link distinct helices are also freely rotatable (**Figure 3.1B**). All electrostatic interactions are ignored, and aside for a small attractive interaction only experienced between base-paired B pseudo-atoms (see below), all attractive interactions such as stacking or hydrogen bonding of single-stranded nucleotides are ignored.

We implemented TOPRNA in CHARMM (17) using the standard CHARMM potential

$$U = \sum_{bonds} K_b\left(b - b_0\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_0\right)^2 + \sum_{dihedrals}\sum_{n} K_\varphi\left(1 + \cos\left(n\varphi - \delta\right)\right) + \sum_{impropers} K_\omega\left(\omega - \omega_0\right)^2 + \sum_{\substack{non-bonded \\ pairs}} \varepsilon_{ij}^{\min}\left[\left(\frac{R_{ij}^{\min}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}^{\min}}{r_{ij}}\right)^{6}\right] \quad \textbf{[3.1]}$$

Nucleic-acid-like geometry between pseudo-atoms is maintained through the application of bond, angle, and S-pseudo-atom chiral center improper torsion potentials to each nucleotide. The helical conformation of base-paired nucleotides is maintained by dihedral potentials placed along the backbone and across the base-pair bond (**Figure 3.1**). Potential parameters were derived from fits to statistical potentials constructed through Boltzmann conversions at 300 K

$$E_v(x) = -k_B T * \ln\left(\frac{P_v(x)}{\max\left(P_v(x)\right)}\right), \quad \textbf{[3.2]}$$

where $P_v(x)$ is the probability of a structural parameter v having a value of x **(Figure 3.1**). Basic connectivity and geometry parameters, as well as backbone dihedrals for base-paired nucleotides, were derived from fits to the RNA05 structural database (19). Base-pairing specific parameters were derived from fits to a database comprising the 6,677 4-base-pair continuous helices returned by an RNA FRABASE (20) search performed on May 28, 2010.

Backbone bond and angle potentials, b_0 and θ_0, were set as the RNA05 database mean values and K_b and K_θ were manually chosen to allow a range of motions consistent with that exhibited by the statistical potentials (**Figure 3.1D**). The P_i-S_i bond and angles involving this bond exhibited stiffer and weaker harmonic potential behavior at small and large deviations (**Figure 3.1D**). Given that the minimum of these potentials corresponded to helical conformations, we assigned the stiff potential to base-paired nucleotides postulating that they should be confined to values near the helical minimum. Conversely, we assigned the weaker

harmonic to single-stranded nucleotides, allowing them to adopt the full range of conformations observed in the database. All other *K* of a given bond or angle type are invariant with respect to nucleotide-identity or base-pairing status. While b_0 and θ_0 were set on a nucleotide-specific basis, they vary by only <~0.6 Å and <~10° respectively. The improper potential applied to the S pseudo-atom chiral center was given parameters of ω_0=30° and K_ω=3.5 kcal/mol/radian^2 for all nucleotide types.

Base-pairs (AU, GC, and GU) were implemented by placing bond, angle, and dihedral potentials directly between paired B pseudo-atoms. Bond and angle parameters were determined as above using our helix database. Dihedrals across the base-pair bond, of type S_i-B_i-B_j-S_j, were treated as n=1 cosines, with K_φ and δ determined through fits to the statistical potentials using the Algorithm::CurveFit module of Perl; K_φ of all pairs were later increased to 5 kcal/mol to increase helical rigidity (see below). Dihedrals of type P_i-S_i-B_i-B_j were found to be harmonic, and parameterized as weak impropers (ω_0 set to the helical-database means and K_ω=0.5 kcal/mol/radian^2).

The helical conformation of contiguous base pairs was enforced through backbone dihedral potentials of type B_i-S_i-P_{i+1}-S_{i+1} and S_i-P_{i+1}-S_{i+1}-P_{i+2} that were placed between base-paired nucleotides *i* and *i+1*. An improper dihedral, indicated by the dashed lines in **Figure 3.1B**, is also placed between consecutive base pairs. As emphasized above, none of these potentials are applied to single-stranded nucleotides or to regions that serve as pivots between distinct helices (**Figure 3.1**). The ω_0 and K_ω of the improper were derived from our helical-database consistent with the procedure used for other harmonic potentials. The two backbone dihedrals were parameterized as four-term cosine series fit to RNA05-derived statistical potentials as above. S_i-P_{i+1}-S_{i+1}-P_{i+2} potentials were treated as residue-type independent and were derived from statistics

combined from all RNA05 trinucleotide sequences (**Figure 3.1E**). K_φ was later uniformly doubled for all backbone potentials to increase helical rigidity (**Figure 3.1E**). The need for these increases, along with those of the K_φ of the base-pair bond dihedrals mentioned above, is likely due to two factors: cosine series cannot capture the full steepness of the A-form helix energy-well; and for the backbone dihedrals, the statistical potentials were derived from a mix of both helical and non-helical conformations, causing them to overestimate helical flexibility.

Electrostatics were eliminated by setting all charges to 0 and non-bonded van der Waals (VDW) interactions were only considered between pseudo-atoms separated by four or more bonds. These interactions were truncated at 8 Å using a switching function turned on at 6 Å. VDW parameters were uniformly assigned with ε_{ij}^{min}=0.01 kcal/mol, which effectively eliminates attractive VDW energies but also reduces the repulsive component of these potentials. Thus, $R_i^{min}/2$ ($R_{ij}^{min}=R_i^{min}/2+ R_j^{min}/2$) were obtained as the values that produced a repulsive energy of $U_{vdw}\approx$0.6 kcal/mol at radial distances approximating the shortest radial dimension of the chemical moiety being represented. For example, $R_i^{min}/2$ was set to 4.0 Å for P pseudo-atoms, which produces a repulsive energy of $U_{vdw}\approx$0.6 kcal/mol at a radius of 2.7 Å. While this parameterization fairly approximates sugar and phosphate moieties, it does a poorer job of capturing the oblong steric profile of bases. For single-stranded nucleotides, this simply means that TOPRNA provides a lower bound approximation of the steric constraints on RNA. However, for paired nucleotides this decreased steric profile results in a ~2 Å steric gap between the paired B pseudo-atoms that in rare cases allowed helices to interpenetrate one another. To prevent such behavior we placed a fourth filler pseudo-atom (termed M) in parallel to the base-pair bond of paired bases. M pseudo-atoms were parameterized to be collinear with and at the midpoint of their respective B-B bonds, with K_b and K_θ set to values ~10% of those used for the

B-B bonds and angles, and with VDW radii approximating 1 Å. This parameterization ensures that M pseudo-atoms minimally affect the dynamics of the base-pair bond, serving only to fill the steric gap.

As an exception to default non-bonded parameters, attractive VDW interactions were placed between the B pseudo-atoms of base-paired nucleotides to simulate intra-helix stacking using the CHARMM NBFIX functionality. We reiterate that these interactions are not experienced by single-stranded nucleotides. For two helical B pseudo-atoms of nucleotide types *X* and *Y*, ε_{ij}^{min} was set to the most negative of the $\Delta G°_{37}$ measured for either an unpaired *X* stacking 3' to a paired *Y* or an unpaired *Y* stacking 3' to a paired *X* (21). These ε_{ij}^{min} ranged in value from -0.1 kcal/mol to -1.7 kcal/mol. R_{ij}^{min} was determined by computing the average 5'→3' and 3'→5' distances between the B pseudo-atoms of stacked *X* and *Y* nucleotides from our helical database, choosing the minimum of these two values, and subtracting 0.15 Å (which was found to produce more A-form consistent helices). While these attractive interactions minimally affect inter-helical stacking across a junction, on some rare occasions two helices would form non-physical interactions between their major grooves, mediated by extensive favorable B-B pseudo-atom interactions. Selective increases in the R_{ij}^{min} of interactions between M atoms and the S and P pseudo-atoms of base-paired nucleotides to effective radii of ~5 Å and ~6.5 Å, respectively, and increases of the R_{ij}^{min} between two M pseudo-atoms to ~3 Å, successfully prevented the formation of these conformations.

Dynamics simulations were performed by assigning pseudo-atom masses to the sum of the represented moiety's heavy atom masses and using the Langevin equation integrated with a 20 fs time step and 5 ps^{-1} friction coefficient.

3.2.2 Benchmarking simulations of TOPRNA helical parameterization

12-nucleotide (nt) random sequence hairpins were constructed such that stem base pairs had a 20% and 80% probability of being a GU or WC pair, respectively. Initial coordinates were obtained by first initializing the sequences as linear single-stranded chains containing both 5' and 3' P pseudo-atoms. A short simulation was then run in the presence of backbone dihedral and distance restraints that forced the stem nucleotides to adopt a helix-like conformation. This was followed by a removal of the restraints, addition of the base-pair bonds, M pseudo-atoms, and other associated pairing potentials, and then minimization of the system. Simulations of HIV-1 and HIV-2 TAR molecules were initialized from coordinates built from the first member of the NMR-MD ensembles (22). The A22•U40 base pair at the top of TAR helix 1, which is unstable in the NMR-MD ensemble, was excluded from later analyses. Both hairpin and TAR dynamics simulations were performed for 100 ns at 300 K, recording conformations every 200 ps.

The RMSD of the generated helices from idealized A-form structure was computed by aligning base-paired S and P pseudo-atoms to a TOPRNA representation of an idealized A-form helix (23), excluding the 5'-most P pseudo-atom of each stem strand. The helical twist of TOPRNA base-pair steps was determined by adapting a previously developed all-atom procedure (24): for two sequential base pairs, *i* and *i+1*, the rotation transform that brings the B and S pseudo-atom coordinates of *i* into concordance with the B and S pseudo-atoms of *i+1* was computed and related to the base-pair steps' local parameters of twist (ω), bending (Γ), and bending phase-angle (ϕ) through

$$\mathbf{R}_z(\omega/2-\phi)\mathbf{R}_y(\Gamma)\mathbf{R}_z(\omega/2+\phi). \qquad \textbf{[3.3]}$$

The local parameters were then used to determine the helical twist Ω_h through

$$2\cos(\Omega_h) = \cos(\omega)(1 + \cos(\Gamma)) - (1 - \cos(\Gamma)). \quad \textbf{[3.4]}$$

$\mathbf{R_z}$ and $\mathbf{R_y}$ are rotations about the x and y axes respectively.

3.2.3 Bulge simulations

Bulge motifs were constructed from the random sequences specified in **Table B.1** and initialized as single linear chains with a 3-nt linker connecting the two strands. Similar to the hairpins above, temporary restraints were used to fold the chains into helical structures, after which appropriate base-pair bonds, M pseudo-atoms, and other associated potentials were added and the linker nucleotides deleted, leaving trailing 3' and 5' P pseudo-atoms on both strands for symmetry. Initial coordinates for the no-connectivity systems were obtained by deleting the B, S, and P pseudo-atoms of the bulged nucleotide of the 1-nt systems. Two independent temperature replica exchange (TREX) simulations, each comprising 100,000 exchanges, were run using three replicas spanning 300 to 400 K with 1,000 steps of dynamics separating exchange attempts. The low temperature conformations of each simulation were combined to achieve a total of 200,000 exchanges of sampling for each sequence, and 2,000,000 exchanges for each bulge. TREX simulations were performed through the MMTSB replica exchange server (25), with 100,000 cycles of TREX requiring ~10 wall and ~30 CPU hours.

Protocols described in **Chapters 1-2** were used to measure Euler angles, (α_h, β_h, γ_h), that describe the inter-helical angle of each bulge conformation (23). Alignments were done to a TOPRNA representation of an idealized helix (23) using the P and S pseudo-atoms of the bulge-helices' three base pairs, excluding the 5'-most P pseudo-atoms of each strand; these 5' P pseudo-atoms lack dihedral potentials and are not confined to A-form-consistent conformations.

Measured angles were rounded to their nearest 10° grid increment and the fraction of total angles that were sampled was computed as previously described (23). Comparisons to our prior models were done by only considering the rigid-body predicted angles that were increments of 10° (26).

Free energies (ΔG^{topo}) of different inter-helical conformations were computed from our simulations using the equation

$$\Delta G_i^{topo} = k_B T \ln\left(P_{\max}/P_i\right). \quad \textbf{[3.5]}$$

P_i is the probability of the inter-helical angle *i* being sampled for a given a bulge type, and was computed by counting number times the angle was sampled over 2,000,000 REX cycles at 300 K and then dividing by 2,000,000. P_{max} is the maximum over all P_i of a given bulge and T is the temperature, set to 300 K.

3.2.4 Comparisons to PDB bulges

The set of inter-helical angles observed in the PDB was obtained by querying the RNA FRABASE (20) for all X-ray and NMR RNA structures containing two helices of at least 3-bp connected by 1 to 6-nt bulges. The searches were performed on September 20, 2012 and done in accordance with our earlier procedures (26). Each bulge was converted to a TOPRNA representation and its inter-helical orientation measured as described above. Conformations were counted as 'sampled' if they were ≤10° from the nearest TOPRNA sampled grid point (23).

ΔG^{topo} estimates for PDB inter-helical conformations were obtained by averaging the TOPRNA sampling probabilities (P_i in equation 5) of all grid points within a 10° radius of the PDB-measured angle. Grid points not sampled by TOPRNA were included in the average by setting their $P_i = 0$. This average, $\langle P_i \rangle$, was then substituted for P_i equation 5. This averaging was done to account for the ~10° errors associated with measuring inter-helical angles (23), and for

the steep changes in ΔG^{topo} across grid points. Analysis of the 525 PDB bulges with $\Delta G^{topo} > 2.5$ kcal/mol was done by clustering the bulges according to sequence and similar inter-helical angles and then manually examining several representatives from each cluster for RNA tertiary, protein, or crystal contacts. Contacts were defined as a heavy atom distance ≤3.5 Å.

3.2.5 Energy Analysis of Rigid-Body Model Predicted Conformations

2-nt, 3-nt, and 4-nt TOPRNA bulge systems were constructed according to same methods as that used for the equilibrium simulations, using sequences of 5'GCG$(U)^X$CGC and 5'GCGCGC for the two strands. These systems share the same base pairs as the idealized helix used in our previous rigid-body calculations, with $(U)^X$ indicating a poly-U bulge of X-nts. Molecular scaffolds possessing each of the rigid-body predicted (α_h, β_h, γ_h) conformations were also built for each bulge system by performing the necessary rotations to 6-bps of a TOPRNA representation of an idealized A-form helix, as described previously (26,27). The bulge systems were then targeted to the scaffolds by applying RMSD restraints to the P and S atoms of the base-paired residues with a force constant of 200 kcal/mol/$Å^2$ and performing 2 ns of dynamics at 300 K. After minimization, the energy of the bulge systems was evaluated, excluding the energetic contributions of the harmonic restraints. We note that (α_h, β_h, γ_h) of the rigid-body topologically allowed spaces that were added as error padding or for the intrinsic flexibility of A-form helices were not considered in this analysis given their speculative nature (26).

3.2.6 Measuring TOPRNA-predicted RDCs

Residual dipolar couplings (RDCs) report on the average orientation of individual RNA bond vectors with respect to the NMR external magnetic field (28,29). Given an ensemble of

structures *{X},* the average RDC $\langle D_{nm} \rangle$ of the inter-nuclear bond vector between atoms *n* and *m* can be calculated as

$$\langle D_{nm} \rangle = \frac{\mu_0 \gamma_n \gamma_m h}{8\pi^3 \langle r_{nm}^3 \rangle_{\{X\}}} \left\langle \sum_{kl=\{x,y,z\}} S_{kl} \cos(\alpha_{k,nm}) \cos(\alpha_{l,nm}) \right\rangle_{\{X\}}, \quad \textbf{[3.6]}$$

where γ_n is the gyromagnetic ratio of nucleus *n*, r_{nm} is the inter-nuclear distance, S_{kl} are the order tensor elements describing the global alignment of the molecule, and α_k is the angle of the bond vector with respect to the *k*-th axis of the molecular frame (30). S_{kl} were set equal to the experimental values obtained for either the HI or HII elongated helices of domain-elongated TAR and assumed to be independent of molecular conformation (31,32). *{X}* was obtained by performing the necessary inter-helical rotations (26) to two idealized helices of the same sequence as TAR for all inter-helical angles sampled by TOPRNA. The subsequent ensemble averages were weighted according to the TOPRNA populations.

3.3 Results and discussion

3.3.1 TOPRNA accurately models A-form helical structure and dynamics

To verify that the TOPRNA force field generates helices possessing correct A-form structure we performed simulations of 400 different 4-base-pair (bp) helices capped by 4-nucleotide (nt) hairpin loops and compared the structural characteristics of these helices to that observed in the PDB (**Figure 3.2**). Both the RMSDs of the helices from idealized A-form structure and helical twists fall within the range of values observed in the PDB (**Figure 3.2**). The few helices possessing mean RMSDs greater than one standard deviation above the PDB mean contained at least one GU base pair, consistent with these pairs' ability to induce helical distortions (33,34).

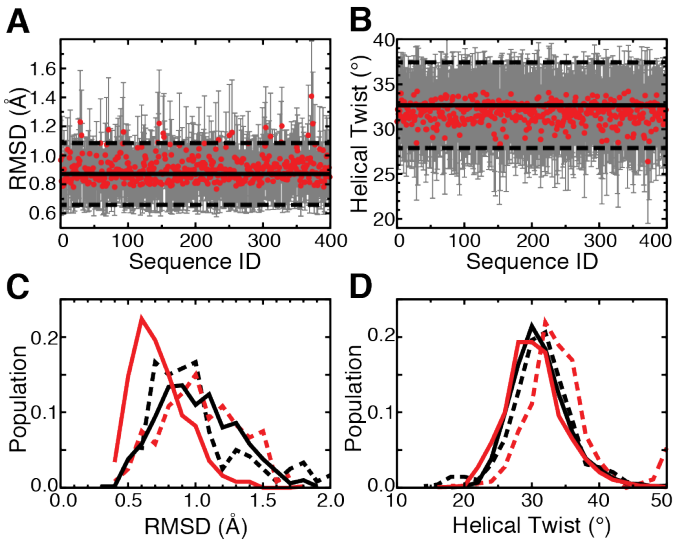

Figure 3.2: Simulations confirm TOPRNA's ability to reproduce A-form helical behavior
(A, B) The means and standard deviations of helical parameters measured for simulations of 400 random-sequence 4-bp hairpins are shown as red points and grey bars. Solid and dashed horizontal lines indicate the means and standard deviations of these parameters measured from our database of PDB helices. The helical twist values in (**B**) have been averaged over the three hairpin base-pair steps. **(C, D)** The distribution of helical parameters measured from TOPRNA simulations (solid lines) and the NMR-MD dynamic ensemble of HIV-1 TAR (dashed lines) (22). Parameters were measured for the first five base pairs of the lower helix (black) and the four base pairs of the upper helix (red), and the populations of the helical twist parameters represent distributions over both different conformations and base-pair steps.

We also examined the ability of TOPRNA to reproduce the dynamic behavior of helices in solution by comparing TOPRNA simulations of HIV-1 and HIV-2 TAR RNAs to previously constructed NMR-MD dynamic ensembles of these molecules (22). Built by using residual dipolar coupling (RDC) NMR measurements to select high-confidence structures from all-atom molecular dynamics (MD) simulations, these ensembles provide the most accurate available picture of the thermodynamic ensemble of states populated by RNA helices (22). Shown in **Figures 3.2** and **B.1**, TOPRNA helices are slightly more idealized over those observed in the NMR-MD ensembles, but overall exhibit close agreement in helical twist and idealized A-form RMSD distributions. Combined, these results demonstrate that TOPRNA accurately captures both the structure and dynamics of RNA helices.

3.3.2 TOPRNA analysis of topological constraints in bulge junctions

As the first application of TOPRNA we set out to further characterize the topological constraints that govern two-way junction bulge motifs. In **Chapter 2** using our heuristic idealized rigid-body rotation models, we demonstrated that these motifs are strongly topologically confined to a small subset of inter-helical conformations (26,27). However, necessary approximations made by these heuristic models, such as neglecting the volume-excluding properties of single-stranded bulge nucleotides, have obscured a complete understanding of the role that topological constraints play in dictating bulge conformation.

Through extensive simulations of model junctions comprising two, 3-bp helices connected by bulges of 1 to 4 nts in length, we utilized TOPRNA to achieve unprecedented sampling of the set of interhelical conformations accessible to different bulge motifs (**Figure 3.1A**). For completeness, each of the different bulge types was simulated using 10 different randomized sequences for 200,000 cycles of temperature replica exchange, generating a total of 2,000,000 different conformations per bulge. Then, employing the (α_h, β_h, γ_h) Euler angle convention (**Figure 3.3A**), where α_h and γ_h represent the twists of the two helices and β_h the inter-helical bend (23), we quantified the sampled inter-helical conformations and compared them to our prior results.

As shown in **Figures 3.3** and **B.2**, these simulations reveal strong agreement with the heuristic models. Not only do the sampled regions largely overlap (**Table 3.1**), but also the finer contours of these regions, such as the linear correlation among α_h and γ_h (26), are duplicated in both models (**Figure 3.3, B.2**). Both models also sample similar magnitudes of the inter-helical conformational space (**Table 3.1**). Thus, TOPRNA provides an independent corroboration of the significance of topological constraints in RNA structure.

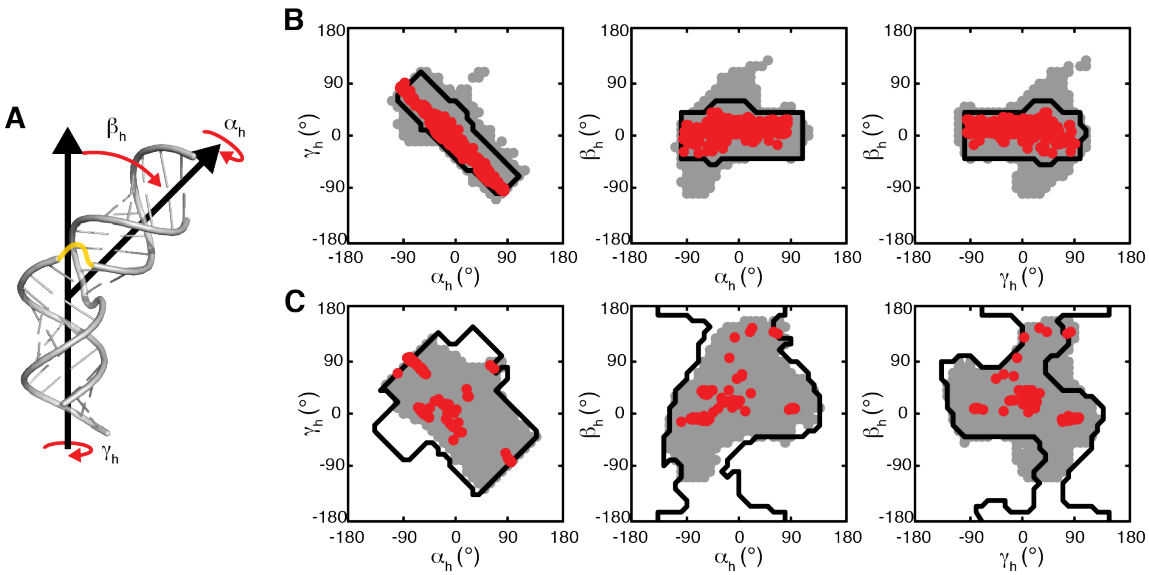

Figure 3.3: TOPRNA simulations of bulges reproduce the topologically allowed space.
(**A**) Cartoon of the (α_h, β_h, γ_h) convention used to quantify interhelical conformations, shown using an artificial junction between two idealized helices. A possible path of the single-stranded bulge is drawn in yellow. (**B, C**) 2D projections of the (α_h, β_h, γ_h) sampled by TOPRNA (grey), observed in the PDB (red), and predicted to be allowed by the **Chapter 2** heuristic model (black outlines), are shown for 1-nt and 3-nt bulges, respectively.

Bulge	Fraction of total (α_h, β_h, γ_h) sampled by TOPRNA (heuristic)[a]	Fraction of TOPRNA (α_h, β_h, γ_h) overlap with heuristic[b]	Fraction of heuristic (α_h, β_h, γ_h) overlap with TOPRNA[c]	Fraction of PDB conformations sampled by TOPRNA (heuristic)[d]
1-nt	0.075 (0.053)	0.61	1.0	1.0 (1.0)
2-nt	0.11 (0.097)	0.72	0.95	0.99 (0.79)
3-nt	0.15 (0.20)	0.81	0.80	1.0 (0.99)
4-nt	0.20 (0.38)	0.81	0.53	1.0 (0.93)

Table 3.1: Comparison between TOPRNA and heuristic model topologically allowed spaces
[a]Due to rounding (α_h, β_h, γ_h) to a 10° vs. 5° grid the heuristic fractions differ slightly from that reported in Chapter 2.
[b]The fraction of TOPRNA sampled conformations that are found within the heuristic idealized-helix topologically allowed space.
[c]The fraction of heuristic idealized-helix topologically allowed conformations sampled by TOPRNA, excluding conformations that were added to the heuristic-predicted allowable conformations as estimates of error padding or "intrinsic helical degrees of freedom".
[d]Comparisons were done to a total of 1853, 705, 347, and 30 PDB structures of 1, 2, 3, and 4-nt bulge systems, respectively.

3.3.3 Non-idealized helical behavior allows sampling of new inter-helical conformations

Despite the overlap between the two models, important differences do exist. 19-39% of the TOPRNA sampled conformations correspond to 'new' states that were not predicted by the heuristic models, with the 1 and 2-nt TOPRNA bulges sampling 40% and 10% more of the (α_h, β_h, γ_h) conformational space, respectively (**Figures 3.3, B.2, Table 3.1**). In contrast, the 3 and 4-nt simulations sample 25% and 50% fewer conformations. To resolve the physical significance of these differences we examined the ability of the models to capture the distribution of (α_h, β_h, γ_h) observed among the bulge junctions in the PDB. Strikingly, TOPRNA samples 99.9% of the PDB (α_h, β_h, γ_h) compared to 94.8% achieved by the heuristic models (**Figure 3.3**, **B.2**, **Table 3.1**). Furthermore, the single PDB conformation not sampled by TOPRNA, the 2-nt bulge of PDB 4ERD (35), is only 11° from the nearest TOPRNA sampled angle (distances of ≤10° count as 'sampled'), whereas it is 31° outside of the rigid-body topologically allowed space.

The increased sampling of PDB-observed conformations indicates that the new conformations sampled by TOPRNA are physically relevant. In **Chapter 2** we found that deviations in helical structure, particularly those associated with bulges possessing GU closing base pairs, could modify the steric interactions of a junction and thus make new inter-helical conformations accessible. Indeed, the 5.2% of PDB junctions not captured by the rigid-body models either possess GU closing pairs or correspond to the 4ERD or four earlier identified NMR outliers in **Chapter 2**. Hypothesizing that TOPRNA is capturing such deviations in helical structure, we compared the newly TOPRNA sampled (α_h, β_h, γ_h) to topologically allowed spaces built in **Chapter 2** from non-idealized helices; ~70% corresponded to conformations also found in these 'non-ideal' allowed spaces. Analysis of the variances in sampling exhibited by the different TOPRNA simulated sequences further demonstrates that these arise from sequence-

dependent variations in helical structure; each TOPRNA sequence samples 100% of the PDB conformations that share its same junction-closing base pairs (excluding the 4ERD outlier), but only 88-100% of the PDB conformations possessing different closing pairs. Therefore, even though the composition of the (α_h, β_h, γ_h) sampled by different sequences differs by only ~14% on average, with the varying angles typically no more than 12° outside of the set of (α_h, β_h, γ_h) sampled by other sequences, these differences can be important in shaping a bulge's topologically allowed space.

In **Chapter 2** we also noted that translational motions between the helices of a junction, which are not captured by the rigid-body model, had the possibility of dramatically increasing the number of conformations accessible to bulge motifs. Comparison between these previously predicted translation-mediated conformations and the newly TOPRNA-sampled (α_h, β_h, γ_h) revealed that translations explain an additional 7% of the new sampling, with these conformations typically no more than ~20° outside of the allowed spaces predicted by the heuristic model. Thus, while TOPRNA does capture such motions, they appear to be relatively insignificant.

Together, helix non-ideality and translational motions explain ~80% of the new conformations sampled by TOPRNA, demonstrating that the model captures these degrees of conformational freedom. On average, the other 20% are within ~24° of the rigid-body predictions. Determining the physicality of these unexplained conformations requires a fully atomistic model and is outside the scope of this work. We note that even if all are non-physical, the small number of these conformations (<10% of the total sampled by TOPRNA) makes it unlikely that they will introduce significant errors to our analysis.

3.3.4 Topological constraints on bulges exceed prior estimates

As mentioned above, the 3 and 4-nt TOPRNA bulge simulations sample significantly fewer overall inter-helical conformations than predicted by the heuristic models, yet still capture 100% of the conformations found in the PDB. This indicates that the constraints on these systems are much greater than previously estimated. As TOPRNA explicitly models the geometry and sterics of bulge-comprising single-stranded nucleotides whereas the rigid-body models ignored these constraints (26,27), this result is not unexpected. Determination of the TOPRNA energies of the rigid-body-predicted but 'TOPRNA-unsampled' (α_h, β_h, γ_h) conformations confirmed that these states are precluded due to elevated energies that primarily arise from bulge-nucleotides (see **Section 3.2.5**, **Figure B.3**).

An important consequence of these greater constraints is that a bulge's topologically allowed space continues to increase in size across a broad range of bulge-lengths. In **Chapter 2**, we had predicted that a 5-nt bulge had sufficient length to enable sampling of all sterically possible (α_h, β_h, γ_h) conformations. However, TOPRNA simulations of 5, 6, and 7-nt bulges using the same procedures as described above revealed this prediction to be incorrect. The additional steric and stereochemical constraints of the bulge-nucleotides in TOPRNA limit even 7-nt bulges to ~2/3 of the (α_h, β_h, γ_h) space that is accessible to junctions that do not have 'bulge'-strand connectivity (**Figure 3.4**). Thus, though few if any RNAs in the PDB possess junctions containing bulges of length >6 nt (20), in theory such longer bulges could be important in allowing RNAs to access conformations that would be forbidden to shorter bulges.

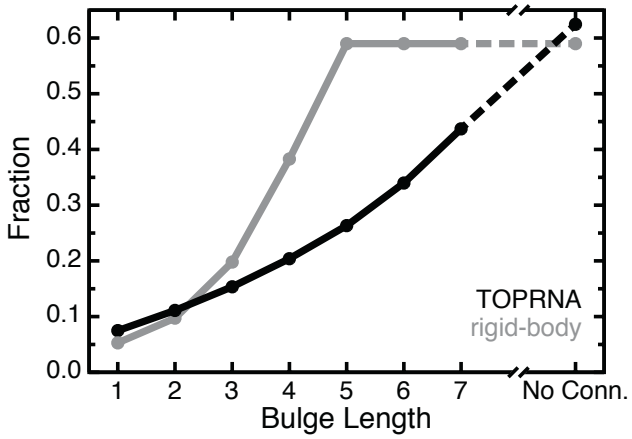

Figure 3.4: The fraction of interhelical space sampled by TOPRNA bulge simulations
The fraction sampled by TOPRNA and the **Chapter 2** rigid-body heuristic model is shown in black and gray, respectively. "No conn." denotes a bulge-like junction that lacks bulge-strand connectivity.

3.3.5 Topological constraints explain the distribution of conformations sampled by polypyrimidine bulges at low salt concentrations

In **Chapter 2** we showed that ensemble averages over the topologically allowed space reproduce both the experimentally measured magnitude and directionality of bulge-induced bends, suggesting that topological constraints may be responsible for the behavior of these systems. Evidenced by the similar anisotropies of the TOPRNA and rigid-body topologically allowed spaces (**Figure 3.3**, **B.2**), TOPRNA also captures the directionality of bulge-induced bends. Ensemble averages over the length of our simulations also reproduce the experimentally measured magnitude of these bends, matching, to approximately within experimental error, the mean bend measured for flexible polyU bulge systems in the absence of Mg^{2+} (**Figure 3.5A**) (36). Furthermore, our simulations reproduce the experimental observation that the magnitude of bulge-induced bends plateaus and then decreases as the bulge increases past 6-nts in size (**Figure 3.5A**) (37). However, this agreement does not hold for 1-nt bulges or for polyU bulges in the presence of Mg^{2+} (**Figure 3.5A**). PolyA bulges also exhibit larger bends than polyU and TOPRNA bulges, and have different Mg^{2+}-dependent behavior than polyU bulges (36).

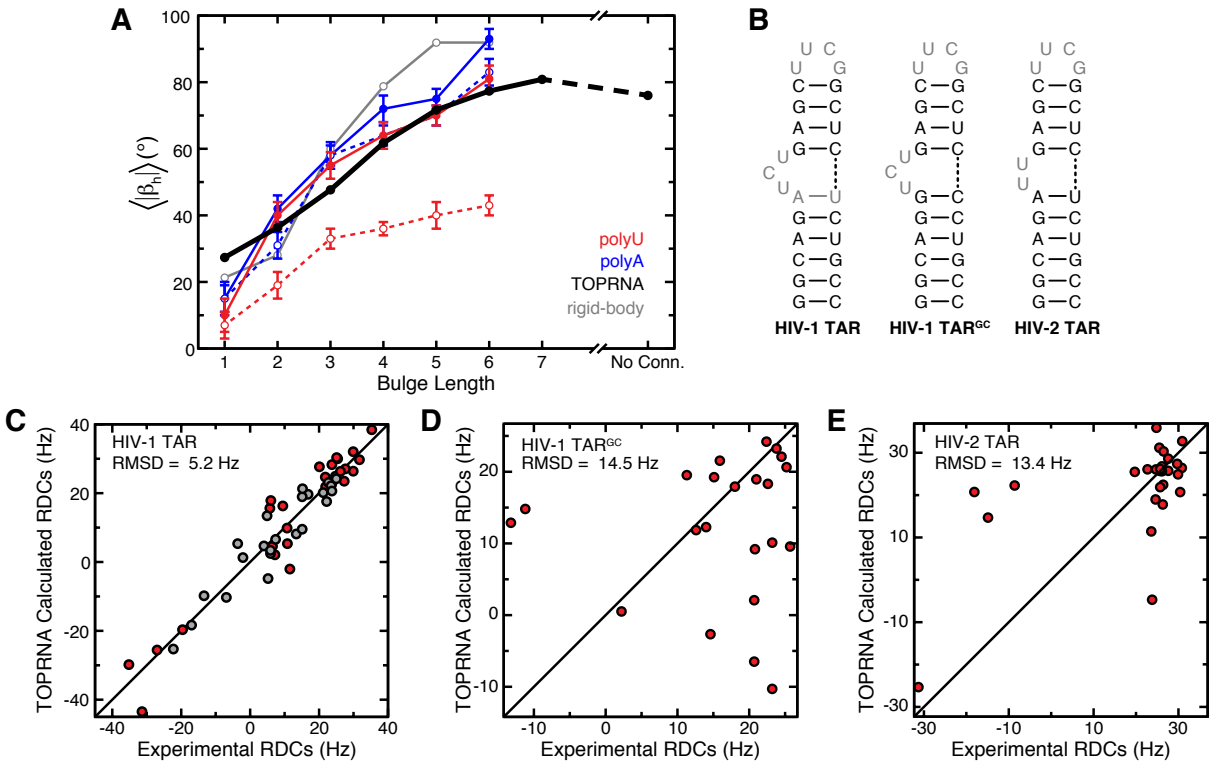

Figure 3.5: TOPRNA bulge simulations reproduce experimental properties of bulge motifs.
(A) $|\beta_h|$ averaged over TOPRNA simulations (black) and over the heuristic model predicted allowed spaces (gray) for different bulge motifs is plotted next to the bend angles measured by transient electronic birefringence for polyU (red) and polyA (blue) bulges in the absence (solid line, filled circles) and presence (dashed line, open circles) of Mg^{2+} (36). Error bars denote experimental error (36). "No conn." denotes a bulge-like junction that lacks connectivity in its bulge-strand. **(B)** HIV-1 TAR, HIV-1 TAR^{GC}, and HIV-2 TAR secondary structures. RDCs from nucleotides shown in gray were excluded from the RDC analysis. (**C, D, E**) HIV-1 TAR, HIV-1 TAR^{GC}, and HIV-2 TAR experimental RDCs versus values computed from either 3-nt or 2-nt TOPRNA bulge simulations (31,32). In **(C)**, red and gray points correspond to RDCs from helix I and helix II TAR elongations, respectively. N-H bond vectors were excluded from all RDC analysis.

We suggest that the above observations can be explained through the following model. At low salt concentrations, electrostatic repulsion between helices cancels out otherwise favorable inter-helical stacking interactions. For polyU bulges, which lack strong intra-bulge stacking, this results in a highly dynamic state that is largely governed by topological constraints and approximated by TOPRNA. Higher salt concentrations screen this repulsion and thus promote inter-helical stacking that is ignored by our simulations. The lower conformational entropy of the unstacked state of 1-nt bulges also stabilizes stacking at low salt. By contrast, the alternative

behavior of polyA bulges arises because stronger stacking interactions between the bulged adenines stabilize highly bent conformations at all salt concentrations (36).

Significantly, the above model is also consistent with the behavior of HIV-1 TAR, a biologically important 3-nt polypyrimidine bulge that has been extensively characterized by NMR and other methods (**Figure 3.5B**) (22,31,38-41). These studies have shown that TAR exists in an equilibrium between a dynamic unstacked state that populates a broad range of inter-helical conformations and a coaxially stacked state. At low Na^+ and Mg^{2+} concentrations TAR is predominantly unstacked, with increases in salt progressively stabilizing the stacked state but not altering the nature of the unstacked ensemble (40). Small molecule binding or selective mutations to the closing base pairs of the junction can also stabilize stacking (31,32). To further test our hypothesis that TOPRNA approximates the behavior of polypyrimidine bulges at low salt we thus assessed the ability of our simulations to reproduce atomic-level NMR RDC measurements made on TAR at such conditions (31). We note that RDCs are ideal for such a test as they depend strongly on the entire distribution of populated (α_h, β_h, γ_h) conformations (31,42). Remarkably, ensemble averages over the (α_h, β_h, γ_h) sampled by our 3-nt bulge simulations yield RDCs that match the experimental values to a root-mean-square-difference (RMSD) of 5.2 Hz (**Figure 3.5C**; see **Section 3.2.6**). This value is comparable both to the uncertainty associated with the experimental RDCs (~4 Hz) and to that of the all-atom NMR-MD ensemble mentioned earlier that was specifically optimized for its agreement with a superset of these RDCs (RMSD=4.8 Hz) (22). It is also substantially better than the ~15 Hz RMSD obtained when averaging over the non-optimized 80 ns MD-simulation source of the NMR-MD ensemble (22). By contrast, TOPRNA does a poor job of approximating the RDCs measured on TAR^{GC}, a

mutant with strengthened inter-helical stacking interactions that stacks even at low salt (RMSD = 14.5 Hz; **Figure 3.5B, D**) (32).

We also tested the ability of TOPRNA simulations to reproduce the RDCs of HIV-2 TAR, which contains a 2-nt polyU bulge (**Figure 3.5B**) (31). Here, we also found poor agreement due to inter-helical stacking effects neglected by TOPRNA (RMSD = 13.4 Hz; **Figure 3.5E**). For example, the NMR-MD ensemble constructed for HIV-2 TAR is dominated by coaxially stacked ($|\beta_h|$ <15°) conformations and has no conformations possessing $|\beta_h|$>30°.(22) This stacked conformation closely resembles the low-bend Mg^{2+}-present state of polyU bulges observed by Zacharias and Hagerman (**Figure 3.5A**) (36), and can be explained by the higher Na^+ concentrations used in the NMR experiments (31).

Taken together, these results support a model where the distribution of inter-helical conformations populated by polypyrimidine bulges is governed by an interplay of inter-helical stacking, electrostatic repulsion, and topological constraints. Whereas stacking predominates for short bulges and at high salt conditions, topological constraints govern the behavior of long polypyrimidine bulges at low salt. While the extent to which this holds true for non-polypyrimidine bulges is unclear, this finding nevertheless indicates topological constraints to be a highly significant driver of bulge conformation. It is worth emphasizing that it would be difficult to draw our conclusions regarding the role of topological constraints from models that include all RNA forces. We also note that while our simulations highlight the importance of inter-helical stacking to bulge conformation, increasing the strength of these interactions beyond the minor attraction already present between paired B pseudo-atoms is counter to the topological constraint focus of the model.

3.3.6 Topological constraints encode conformational free energies

The agreement between ensemble averages over our simulations and experiments on polypyrimidine bulges at low salt implies that the free energy landscape explored by TOPRNA mirrors that of real RNAs. We directly computed these energy landscapes from our simulations by converting the sampling probability of each (α_h, β_h, γ_h) angle into a free energy cost, ΔG^{Topo}, relative to a given bulge's highest probability (α_h, β_h, γ_h) conformation (equation **[3.5]**). ΔG^{Topo} reflects both the entropic and internal energy costs imposed by topological constraints. It captures, for example, whether an inter-helical orientation requires its bulge-nucleotides to adopt a strained conformation, or alternatively, if an orientation is entropically favored because it preserves the conformational freedom of its bulge-nucleotides. Note that the small attractive interaction experienced between paired B pseudo-atoms also provides a slight favorable contribution to the ΔG^{Topo} of coaxially stacked conformations. Interestingly, these calculations reveal that the ΔG^{Topo} of different (α_h, β_h, γ_h) conformations is on average ~3 kcal/mol and can be as large 6 kcal/mol (**Figure 3.6**), indicating that topological constraints strongly favor some bulge conformations over others.

To further evaluate the significance of these energy landscapes we again utilized comparisons to the set of (α_h, β_h, γ_h) observed among bulges in the PDB. As the conformations captured by crystallography or NMR primarily reflect 'folded' states that are stabilized by attractive interactions ignored by TOPRNA, we do not expect exact correspondence between ΔG^{Topo} and the distribution of (α_h, β_h, γ_h) in the PDB. However, given that intra-junction attractive interactions are typically ≤2.5 kcal/mol (43,44), if ΔG^{Topo} is significant then we should observe few if any PDB bulges with ΔG^{Topo} greater than this threshold. Remarkably, we find that this is indeed the case, with 82% of bulges possessing $\Delta G^{Topo} \leq 2.5$ kcal/mol (**Figure 3.6C**).

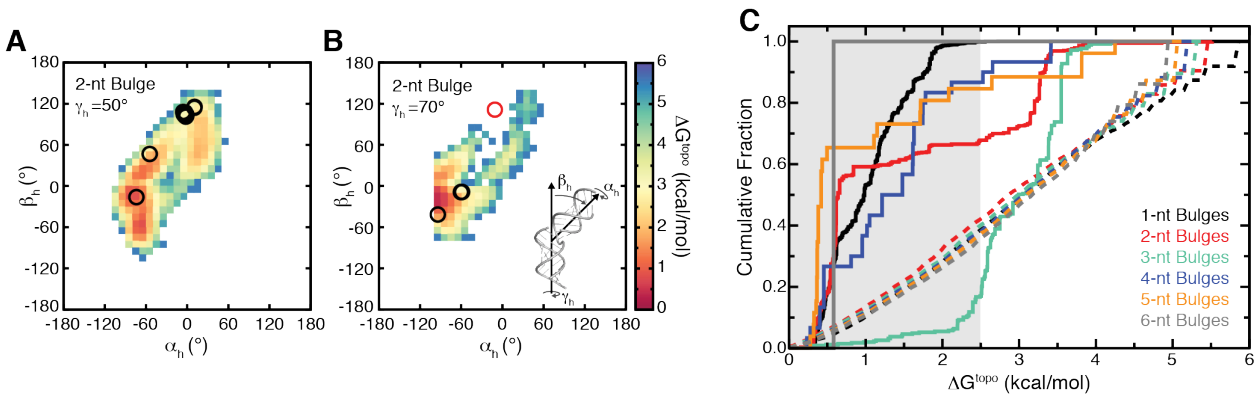

Figure 3.6: PDB junction conformations coincide with the TOPRNA free-energy landscape. Representative 2D slices of the 2-nt bulge free energy landscape (T=300 K) are shown for constant γ_h = 50° **(A)** and γ_h = 70° **(B)**. Open circles denote conformations observed in the PDB, with the red circle in **(B)** highlighting the 4ERD outlier. Inset in **(B)** is a cartoon illustrating the physical meaning of the (α_h, β_h, γ_h) angles. **(C)** Cumulative distribution functions showing the fraction of (α_h, β_h, γ_h) conformations with ΔG^{topo} less than a given cutoff value. The distributions for (α_h, β_h, γ_h) conformations observed in the PDB are shown as solid lines. The distributions for (α_h, β_h, γ_h) in the TOPRNA topologically allowed spaces are shown as dashed lines. The gray background highlights ΔG^{topo} values <2.5 kcal/mol. 1853, 705, 347, 30, 26 and 1 structures are represented by the PDB curves of 1-nt, 2-nt, 3-nt, 4-nt, 5-nt, and 6-nt bulges, respectively.

Moreover, analysis of the bulges possessing $\Delta G^{Topo} \geq 2.5$ kcal/mol revealed that 98% either participate in protein or RNA tertiary interactions, possess such interactions directly up/down-stream of the bulge, or possess crystal packing interactions. For example, the apparently different behavior of 3-nt bulges in **Figure 3.6C** arises because two motifs, Helix 12 of the 16S rRNA and Helix 96 of the 23S rRNA, collectively comprise ~80% of the 3-nt bulges in our database; both of these motifs participate in either RNA-RNA or protein-RNA tertiary interactions.

The above observations suggest that bulges with high ΔG^{topo} are only observed because they are stabilized by auxiliary interactions. For several bulges, this claim is directly supported by the existence of alternative lower-energy structures. Notably, in two different crystals (45,46) the 5-nt bulge of the HCV IRES IIa domain adopts conformations possessing ΔG^{topo}=3.8-4.3 kcal/mol. However, in an alternative ligand-bound crystal structure (47) and in solution NMR structures (48,49) this bulge is limited to conformations having ΔG^{topo}=1.1-2.1 kcal/mol. Similarly, of nine structures of HIV-1 TAR in our database (50-57), eight possess ΔG^{topo}<1.7

kcal/mol, with only the crystal-contact stabilized 397D (51) structure exhibiting a high-energy conformation (ΔG^{topo} = 3.1 kcal/mol). The protein-bound 2-nt bulge of the 4ERD (35) outlier mentioned earlier (ΔG^{topo} > 5.5 kcal/mol), and the 2-nt bulge of Helix III of the 5S rRNA (ΔG^{topo} = 1.5-2.6 kcal/mol in various ribosome structures; see for example (58,59)), have also been shown by NMR (60,61) to adopt lower-energy *apo* conformations possessing ΔG^{topo} = 2.0 and ΔG^{topo} = 0.6 kcal/mol, respectively.

The 2% of $\Delta G^{topo} \geq 2.5$ kcal/mol bulges that are not explained by auxiliary stabilizing interactions correspond to NMR structures. Of the 10 total, 7 were solved without RDC restraints (62-65), suggesting that their global structures may be unreliable (66), and one is an averaged structure of the HCV IRES derived from an ensemble whose best representative conformation has ΔG^{topo} = 1.7 kcal/mol (48). The final two are from the unpublished NMR structure 1U3K, for which refinement details were not available.

Together, these results strongly indicate that topological constraints play an important role in defining the RNA free energy landscape. In the absence of external stabilizing interactions, bulges are largely limited to low ΔG^{topo} conformations that are encoded by secondary structure. In addition, as evidenced by the experimental agreement above, these topological-constraint-encoded energy landscapes provide a reasonable estimate of the ensemble of structures sampled by dynamic polypyrimidine bulges at low salt. We note that electrostatics, preferred backbone rotameric states, and attractive interactions involving single-stranded nucleotides, all of which are ignored by the TOPRNA force field, are also key drivers of final RNA 3D structure. Indeed, detailed models that include these forces should be expected (and have been shown (9,11,13,67-69)) to achieve superior experimental agreement given sufficient sampling. However, the success of TOPRNA indicates that for a bulge with stable secondary

structure, these other energetic terms serve to fine-tune the energy landscape predefined by topological constraints. This finding emphasizes the continuing need for a more holistic understanding of the forces that dictate RNA 3D structure. We suggest that while not suitable by itself for high-resolution structure prediction, this finding may help guide further improvements in dedicated structure prediction and design methods. NMR structure refinement protocols may also benefit by considering the energetic contributions of topological constraints.

We also note the striking parallels between this finding and that obtained by Herschlag and coworkers (2) from their studies on simplified junction mimics. In these systems, topological constraints were shown to destabilize certain junction conformations by as much as ~5 kcal/mol and were hypothesized to play a potential role in encoding specificity of RNA 3D structure. By establishing that such topological-constraint-defined energies also exist in biologically relevant junctions, our analysis provides a tantalizing clue that secondary structure, independent of exact sequence, may indeed be encoding the 3D structure of many RNAs.

3.4 Conclusions

In this Chapter, we have developed the coarse-grained molecular dynamics model TOPRNA that is optimized to explore the contributions of topological constraints to the folding and dynamics of complex RNA systems. TOPRNA simulations of two-way bulge junctions recapitulate our findings in **Chapter 2** that topological constraints are significant determinants of bulge 3D structure. In particular, these basic constraints limit the set of inter-helical conformations accessible to 1- to 4-nt bulges to 7-20% of the total theoretical space. With the greater physical accuracy afforded by TOPRNA, we show that deviations in idealized A-form helix structure and stereochemical constraints posed by bulge-linking nucleotides play a critical role in defining the

set of allowed conformations. However, inter-helical translations play a relatively insignificant role.

Strikingly, in addition to defining a limited range of allowed conformations, our simulations demonstrate that topological constraints contribute as much as 6 kcal/mol to the free energy of different bulge conformations. The majority of bulges in the PDB adopt conformations with low ΔG^{topo}, and the few bulges with high ΔG^{topo} appear to be stabilized by interactions with external partners such as proteins. The surprising ability of our simulations to reproduce experimental measurements made on polypyrimidine bulges at low salt concentrations suggests that in some cases bulge conformational free energy is primary determined by ΔG^{topo} alone. However, for sequences that more strongly stack, or at higher salt concentrations, other energetic terms such as electrostatics and attractive interactions are needed to explain RNA behavior. Nevertheless, as a whole our results strongly support that topological constraints are a significant determinant of bulge conformational free energy. The generality of both the TOPRNA force field and the analysis methods developed in this Chapter provide a strong foundation for studies of topological constraints in a wide variety of RNAs.

3.5 References

1. Bailor, M.H., Sun, X. and Al-Hashimi, H.M. (2010) Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science*, **327**, 202-206.
2. Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S. and Herschlag, D. (2009) Do conformational biases of simple helical junctions influence RNA folding stability and specificity? *RNA*, **15**, 2195-2205.
3. Denning, E.J., Priyakumar, U.D., Nilsson, L. and Mackerell, A.D., Jr. (2011) Impact of 2'-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA. *J Comput Chem*, **32**, 1929-1943.
4. Zgarbova, M., Otyepka, M., Sponer, J., Mladek, A., Banas, P., Cheatham, T.E., 3rd and Jurecka, P. (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput*, **7**, 2886-2902.

5. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA,* **15**, 189-199.
6. Whitford, P.C., Schug, A., Saunders, J., Hennelly, S.P., Onuchic, J.N. and Sanbonmatsu, K.Y. (2009) Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function. *Biophys J*, **96**, L7-9.
7. Hyeon, C. and Thirumalai, D. (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat Commun*, **2**, 487.
8. Tan, R.K.Z., Petrov, A.S. and Harvey, S.C. (2006) YUP: A molecular simulation program for coarse-grained and multiscaled models. *J Chem Theory Comput*, **2**, 529-540.
9. Hyeon, C. and Thirumalai, D. (2005) Mechanical unfolding of RNA hairpins. *Proc Natl Acad Sci USA*, **102**, 6789-6794.
10. Denesyuk, N.A. and Thirumalai, D. (2013) Coarse-grained model for predicting RNA folding thermodynamics. *J Phys Chem B*, **117**, 4901-4911.
11. Ding, F., Sharma, S., Chalasani, P., Demidov, V.V., Broude, N.E. and Dokholyan, N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164-1173.
12. Sharma, S., Ding, F. and Dokholyan, N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951-1952.
13. Cao, S. and Chen, S.J. (2011) Physics-based de novo prediction of RNA 3D structures. *J Phys Chem B*, **115**, 4216-4226.
14. Bernauer, J., Huang, X.H., Sim, A.Y.L. and Levitt, M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *Rna*, **17**, 1066-1075.
15. Xia, Z., Gardner, D.P., Gutell, R.R. and Ren, P. (2010) Coarse-grained model for simulation of RNA three-dimensional structures. *J Phys Chem B*, **114**, 13497-13506.
16. Pasquali, S. and Derreumaux, P. (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. *J Phys Chem B*, **114**, 11957-11966.
17. Brooks, B.R., III, C.L.B., Jr, A.D.M., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. *et al.* (2009) CHARMM: The biomolecular simulation program. *J Comput Chem*, **30**, 1545-1614.
18. Knotts, T.A., Rathore, N., Schwartz, D.C. and de Pablo, J.J. (2007) A coarse grain model for DNA. *J Chem Phys*, **126**.
19. Murray, L.J., Arendall, W.B., 3rd, Richardson, D.C. and Richardson, J.S. (2003) RNA backbone is rotameric. *Proc Natl Acad Sci USA*, **100**, 13904-13909.
20. Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E.K., Blazewicz, J. and Adamiak, R.W. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, **11**, 231.
21. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA*, **101**, 7287-7292.
22. Frank, A.T., Stelzer, A.C., Al-Hashimi, H.M. and Andricioaei, I. (2009) Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs:

new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res*, **37**, 3670-3679.
23. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) 3D maps of RNA interhelical junctions. *Nat Protoc*, **6**, 1536-1545.
24. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, **31**, 5108-5121.
25. Feig, M., Karanicolas, J. and Brooks, C.L.I. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model*, **22**, 377-395.
26. Mustoe, A.M., Bailor, M.H., Teixeira, R.M., Brooks, C.L., III and Al-Hashimi, H.M. (2012) New insights into the fundamental role of topological constraints as a determinant of two-way junction conformation. *Nucleic Acids Res*, **40**, 892-904.
27. Bailor, M.H., Sun, X.Y. and Al-Hashimi, H.M. (2010) Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation. *Science*, **327**, 202-206.
28. Tjandra, N. and Bax, A. (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, **278**, 1111-1114.
29. Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc Natl Acad Sci USA*, **92**, 9279-9283.
30. Losonczi, J.A., Andrec, M., Fischer, M.W. and Prestegard, J.H. (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson*, **138**, 334-342.
31. Zhang, Q., Stelzer, A.C., Fisher, C.K. and Al-Hashimi, H.M. (2007) Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, **450**, 1263-1267.
32. Stelzer, A.C., Kratz, J.D., Zhang, Q. and Al-Hashimi, H.M. (2010) RNA dynamics by design: biasing ensembles towards the ligand-bound state. *Angew Chem Int Ed Engl*, **49**, 5731-5733.
33. Masquida, B. and Westhof, E. (2000) On the wobble GoU and related pairs. *Rna*, **6**, 9-15.
34. Varani, G. and McClain, W.H. (2000) The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, **1**, 18-23.
35. Singh, M., Wang, Z., Koo, B.K., Patel, A., Cascio, D., Collins, K. and Feigon, J. (2012) Structural basis for telomerase RNA recognition and RNP assembly by the holoenzyme La family protein p65. *Mol Cell*, **47**, 16-26.
36. Zacharias, M. and Hagerman, P.J. (1995) Bulge-Induced Bends in Rna - Quantification by Transient Electric Birefringence. *J Mol Biol*, **247**, 486-500.
37. Gohlke, C., Murchie, A.I.H., Lilley, D.M.J. and Clegg, R.M. (1994) Kinking of DNA and RNA Helices by Bulged Nucleotides Observed by Fluorescence Resonance Energy-Transfer. *Proc Natl Acad Sci USA*, **91**, 11660-11664.
38. Olsen, G.L., Bardaro, M.F., Jr., Echodu, D.C., Drobny, G.P. and Varani, G. (2010) Intermediate rate atomic trajectories of RNA by solid-state NMR spectroscopy. *J Am Chem Soc*, **132**, 303-308.
39. Zacharias, M. and Hagerman, P.J. (1995) The Bend in Rna Created by the Transactivation Response Element Bulge of Human-Immunodeficiency-Virus Is

Straightened by Arginine and by Tat-Derived Peptide. *Proc Natl Acad Sci USA*, **92**, 6052-6056.
40. Casiano-Negroni, A., Sun, X. and Al-Hashimi, H.M. (2007) Probing Na(+)-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry*, **46**, 6525-6535.
41. Lu, J., Kadakkuzha, B.M., Zhao, L., Fan, M., Qi, X. and Xia, T. (2011) Dynamic ensemble view of the conformational landscape of HIV-1 TAR RNA and allosteric recognition. *Biochemistry,* **50**, 5042-5057.
42. Bailor, M.H., Musselman, C., Hansen, A.L., Gulati, K., Patel, D.J. and Al-Hashimi, H.M. (2007) Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat Protoc*, **2**, 1536-1546.
43. Walter, A.E., Turner, D.H., Kim, J., Lyttle, M.H., Muller, P., Mathews, D.H. and Zuker, M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci USA*, **91**, 9218-9222.
44. Tyagi, R. and Mathews, D.H. (2007) Predicting helical coaxial stacking in RNA multibranch loops. *RNA*, **13**, 939-951.
45. Zhao, Q., Han, Q., Kissinger, C.R., Hermann, T. and Thompson, P.A. (2008) Structure of hepatitis C virus IRES subdomain IIa. *Acta Crystallogr D Biol Crystallogr*, **64**, 436-443.
46. Dibrov, S.M., Johnston-Cox, H., Weng, Y.H. and Hermann, T. (2007) Functional architecture of HCV IRES domain II stabilized by divalent metal ions in the crystal and in solution. *Angew Chem Int Ed Engl*, **46**, 226-229.
47. Dibrov, S.M., Ding, K., Brunn, N.D., Parker, M.A., Bergdahl, B.M., Wyles, D.L. and Hermann, T. (2012) Structure of a hepatitis C virus RNA domain in complex with a translation inhibitor reveals a binding mode reminiscent of riboswitches. *Proc Natl Acad Sci USA*, **109**, 5223-5228.
48. Lukavsky, P.J., Kim, I., Otto, G.A. and Puglisi, J.D. (2003) Structure of HCV IRES domain II determined by NMR. *Nat Struct Biol*, **10**, 1033-1038.
49. Paulsen, R.B., Seth, P.P., Swayze, E.E., Griffey, R.H., Skalicky, J.J., Cheatham, T.E., 3rd and Davis, D.R. (2010) Inhibitor-induced structural change in the HCV IRES domain IIa RNA. *Proc Natl Acad Sci USA*, **107**, 7263-7268.
50. Du, Z., Lind, K.E. and James, T.L. (2002) Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem Biol*, **9**, 707-712.
51. Ippolito, J.A. and Steitz, T.A. (1998) A 1.3-A resolution crystal structure of the HIV-1 trans-activation response region RNA stem reveals a metal ion-dependent bulge conformation. *Proc Natl Acad Sci USA*, **95**, 9819-9824.
52. Davis, B., Afshar, M., Varani, G., Murchie, A.I., Karn, J., Lentzen, G., Drysdale, M., Bower, J., Potter, A.J., Starkey, I.D. *et al.* (2004) Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic "hot spots". *J Mol Biol*, **336**, 343-356.
53. Murchie, A.I., Davis, B., Isel, C., Afshar, M., Drysdale, M.J., Bower, J., Potter, A.J., Starkey, I.D., Swarbrick, T.M., Mirza, S. *et al.* (2004) Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J Mol Biol*, **336**, 625-638.
54. Davidson, A., Leeper, T.C., Athanassiou, Z., Patora-Komisarska, K., Karn, J., Robinson, J.A. and Varani, G. (2009) Simultaneous recognition of HIV-1 TAR RNA bulge and loop

sequences by cyclic peptide mimics of Tat protein. *Proc Natl Acad Sci USA*, **106**, 11931-11936.
55. Aboul-ela, F., Karn, J. and Varani, G. (1995) The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J Mol Biol*, **253**, 313-332.
56. Davidson, A., Begley, D.W., Lau, C. and Varani, G. (2011) A small-molecule probe induces a conformation in HIV TAR RNA capable of binding drug-like fragments. *J Mol Biol*, **410**, 984-996.
57. Davidson, A., Patora-Komisarska, K., Robinson, J.A. and Varani, G. (2011) Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res*, **39**, 248-256.
58. Harms, J.M., Wilson, D.N., Schluenzen, F., Connell, S.R., Stachelhaus, T., Zaborowska, Z., Spahn, C.M. and Fucini, P. (2008) Translational regulation via L11: molecular switches on the ribosome turned on and off by thiostrepton and micrococcin. *Mol Cell*, **30**, 26-38.
59. Voorhees, R.M., Weixlbaumer, A., Loakes, D., Kelley, A.C. and Ramakrishnan, V. (2009) Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. *Nat Struct Mol Biol*, **16**, 528-533.
60. Chen, Y., Fender, J., Legassie, J.D., Jarstfer, M.B., Bryan, T.M. and Varani, G. (2006) Structure of stem-loop IV of Tetrahymena telomerase RNA. *EMBO J*, **25**, 3156-3166.
61. Huber, P.W., Rife, J.P. and Moore, P.B. (2001) The structure of helix III in Xenopus oocyte 5 S rRNA: an RNA stem containing a two-nucleotide bulge. *J Mol Biol*, **312**, 823-832.
62. Schmitz, M. (2004) Change of RNase P RNA function by single base mutation correlates with perturbation of metal ion binding in P4 as determined by NMR spectroscopy. *Nucleic Acids Res*, **32**, 6358-6366.
63. Diener, J.L. and Moore, P.B. (1998) Solution structure of a substrate for the archaeal pre-tRNA splicing endonucleases: the bulge-helix-bulge motif. *Mol Cell*, **1**, 883-894.
64. Houck-Loomis, B., Durney, M.A., Salguero, C., Shankar, N., Nagle, J.M., Goff, S.P. and D'Souza, V.M. (2011) An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature*, **480**, 561-564.
65. Gaudin, C., Mazauric, M.H., Traikia, M., Guittet, E., Yoshizawa, S. and Fourmy, D. (2005) Structure of the RNA signal essential for translational frameshifting in HIV-1. *J Mol Biol*, **349**, 1024-1035.
66. Tolbert, B.S., Miyazaki, Y., Barton, S., Kinde, B., Starck, P., Singh, R., Bax, A., Case, D.A. and Summers, M.F. (2010) Major groove width variations in RNA structures determined by NMR and impact of 13C residual chemical shift anisotropy and 1H-13C residual dipolar coupling on refinement. *J Biomol NMR*, **47**, 205-219.
67. Das, R., Karanicolas, J. and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods*, **7**, 291-294.
68. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51-55.
69. Seetin, M.G. and Mathews, D.H. (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J Comput Chem*, **32**, 2232-2244.

Chapter 4: The Role of Topological Constraints in tRNA Folding and Dynamics

This chapter has been adapted from the following publication:

Mustoe, A. M., Brooks, C. L. III, and Al-Hashimi, H. M. (2014) Topological constraints are major determinants of tRNA tertiary structure and dynamics and provide basis for tertiary folding cooperativity, *Nucleic Acids Research*, in press.

4.1 Introduction

In **Chapters 2** and **3** we extensively characterized the topological constraints of isolated two-way junction motifs. These studies demonstrated that excluded volume and connectivity constraints limit two-way junctions to only 7-26% of their theoretically possible interhelical conformations. Furthermore, topological constraints define anisotropic free energy landscapes that quantitatively approximate experimentally measured structural and dynamic properties of RNA bulges (1-4). These studies indicate the importance of topological constraints as a determinant of RNA 3D conformation and dynamics. However, while two-way junctions are fundamental building blocks of RNA architecture, many biologically important RNAs are organized around higher-order junctions such as three- and four-way junctions. The influence of topological constraints on the structure and dynamics of these more complex junctions is poorly understood.

Studies of isolated two-way junctions are also unable to address the potential roles that topological constraints may play in RNA tertiary folding. In particular, biological function requires many RNA molecules to fold into specific, highly complex 3D structures that are stabilized by tertiary interactions (5-7). Since RNA 3D folding largely takes place from a state possessing prefolded secondary structure, conformational biases arising from the topological constraints of RNA junctions could be important in guiding the tertiary folding outcome. Indeed, from their studies of PEG-linked DNA helices, Herschlag and coworkers (3) hypothesized that topological constraints could contribute to RNA folding specificity by posing large energetic penalties on forming non-native tertiary interactions. Such a mechanism could explain how RNAs overcome the limited specificity of their four-nucleotide alphabets and the resulting inherent promiscuity of tertiary interactions.

Supporting the notion that junctions may influence tertiary folding, fragment assembly studies of tRNA and the adenine riboswitch suggest that inter-helical linkers bias these RNAs towards native-like conformations (8). Structural surveys have also identified correlations between the length of single-strands in higher-order junctions and their folded conformations (9,10). However, these studies were unable to distinguish whether these observations were due to topological constraints or more complex factors such as sequence-specific base stacking. Other proposed roles for topological constraints in RNA folding, such as their potential to contribute to RNA folding cooperativity (11), have yet to be tested. Systematic analyses of the thermodynamic contributions of topological constraints to higher-order junction conformation are needed.

Until now, studying topological constraints in higher-order junctions has proven challenging. The heuristic models used by Bailor *et al.* (1) and **Chapter 2** rely on the many simplifications afforded by two-way junctions. In bulges, the two helices are adjoined at one

strand by a relatively stationary pivot. The same is true for internal loop motifs because bases tend to maximize formation of non-canonical base pairs, resulting in a bulge junction topology. This pivot-like connectivity allowed translations of the helices to be largely ignored and the finite length of the bulge linker to be modeled as a simple distance constraint. By comparison, higher-order junctions lack well-defined pivots, contain multiple single-stranded loops whose behavior is difficult to model *a priori*, and possess many more degrees of freedom due to the larger number of involved helices.

The development of TOPRNA in **Chapter 3** resolves these challenges. In this Chapter, we utilize TOPRNA to characterize for the first time the topological constraints of four-way junctions and their potential contributions to the structure and dynamics of tRNA, which has long served as a paradigm for understanding RNA tertiary folding. Our results expose new features of topological constraints that are unique to higher-order junctions and indicate that topological constraints are harnessed by tRNA to specify its tertiary structure, dynamics, and folding cooperativity.

4.2 Materials and methods

4.2.1 Simulation methods

Initial coordinates were derived from the 6TNA crystal structure, changing modified residues to their unmodified analogs. Dihedral potentials and base pair bonds were added based on each molecule's secondary structure as described previously (4). The cut A/D-loop and cut V-loop tRNAs were obtained by removing the bond between U8(S) and A9(P) and G46(S) and U47(P), respectively. Mutant tRNAs were constructed by shifting the 6TNA residue numbering and using CHARMM to build the subsequent 'missing' nucleotides (sequences are shown in **Figure C.9**).

For the VL-1 variant, coordinates for nt 46 were deleted and the upstream residue numbers shifted down. After building, all residues except for those immediately neighboring the mutation site were harmonically restrained and the system minimized. For the VS mutants, an additional round of building and minimization was used to add the V-stem and/or G26•U44 pairs after the inserted VL nts were initialized.

Restrained simulations were started from the same initial coordinates. Tertiary pairs were enforced using NOE restraints between B beads with r_{min}=5.5 Å and r_{max}=7.5 Å and between the associated S beads with r_{min}=11 Å and r_{max}=14 Å. Base triples were enforced by placing B-B and S-S NOE restraints between the tertiary nucleotide and each of the two helical nucleotides, with r_{min} and r_{max} set to ±1 Å and ±1.5 Å of the approximate B-B and S-S distances found in the crystal structure, respectively. Force constants for all NOE restraints were set to k_{min}=k_{max}=2.0 kcal/mol/Å^2, with maximum force asymptotes of 2.0 kcal/mol/Å. Several force field modifications were also made when restraining the tertiary G26•A44 pair, consistent with its role in extending the AC-stem. The backbone dihedrals of G26 and A44 were given potentials 1/4 the height of those used for WC-paired residues to favor A-form structure. The B beads of both G26 and A44 were also parameterized to experience a small attractive interaction to other paired B beads, as described for canonical pairs above. Finally, an M bead was added to the B bead of G26 to fill steric gaps that would otherwise exist between G26 and A44.

Both unrestrained and restrained simulations were performed using temperature replica exchange Langevin dynamics simulations with eight exponentially spaced temperature windows from 300 K to 450 K. Simulations were performed in CHARMM with a 20 fs timestep and 5 ps^{-1} friction coefficient through the aarex.pl package of the MMTSB toolset (4,12). Exchange attempts were separated by 2,000 dynamics steps, with acceptance ratios varying between 35-

45%. A total of 10^9 and 10^8 dynamics steps per replica were performed for the unrestrained and restrained simulations, respectively, with the first 2×10^6 steps of each simulation treated as equilibration and excluded from analysis. For the unrestrained simulations, this equilibration time was sufficient for the molecule to lose all memory of the starting crystal structure configuration (i.e. completely unfold). Analysis was performed on conformations recorded every 2,000 dynamics steps at 300 K. Convergence was confirmed by comparing ΔG^{topo} values computed from the first 10^8 steps of the unrestrained WT simulation to those obtained from the full 10^9 steps; ΔG^{topo} values varied by less than 0.5 k_BT for long-range contacts that form with $\Delta G^{topo}<9$ k_BT.

4.2.2 Measuring interhelical Euler angles

Euler angles describing the orientation between pairs of helices and the total fraction of these angles that were sampled were computed according to previously described conventions using a bin size of 10° (4,13). The H1 helix used for each pair of helices is always listed first in the text (13). The various crystal structure conformations of tRNA were obtained by searching the RNA FRABASE (14) for all 3.5 Å resolution or better X-ray structures with strand1 = '(((..(((', strand2 = ')))).(((', strand3 = ')))).....(((', and strand4 = ')))))))'.

The fraction of global junction conformations sampled by each simulation was computed by discretizing the measured $3\times(\alpha_h, \beta_h, \gamma_h)$ angles onto a 60° 9D grid and dividing the number of sampled grid points by 108^3 (108 is the number of non-degenerate $(\alpha_h, \beta_h, \gamma_h)$ between two helices on a 60° grid). While the number of possible grid points is significantly larger than the length of our simulations, the 499,000 snapshots of WT tRNA only sampled ~77,000 unique 9D

angles. It is likely that on a finer grid tRNA's global conformation would be substantially more constrained than estimated here.

4.2.3 Measuring mutual information

Mutual information (MI) provides a general measure of correlation derived from information theory that measures the extent to which the probability distributions of the two variables are independent of one another (15). MI ranges from 0 if the distributions are completely independent, to the value of the individual distribution's Shannon entropy if they are completely dependent. The mutual information between the Euler angles of two different helices was computed as

$$MI(X,Y) = \sum_{(X_i,Y_i)} P(X_i,Y_i) * \log\left(\frac{P(X_i,Y_i)}{P(X_i) * P(Y_i)}\right). \quad \textbf{[4.1]}$$

X and Y are the orientations of two helices measured with respect to a common H1 helix, and $P(X_i,Y_i)$, $P(X_i)$, and $P(Y_i)$ are the joint and individual probabilities of the two helices adopting the specific (α_h, β_h, γ_h) conformations X_i and Y_i. Probabilities were computed using histograms with 45° bin widths. The ratio of the number of populated bins to data points ranged between 10 and 20 for the 6D histograms used to compute joint probabilities. MI overestimation due to sample size finiteness was corrected for according to (15).

4.2.4 Measuring correlations between the centers of mass of different helices

To correlate the centers of mass (COM) of two helices, H_i and H_j, a reference helix H_r ($H_r \neq H_i \neq H_j$) was chosen as an internal reference frame. Global translations and rotations of the molecule were removed by superimposing the three junction-closing base pairs of H_r. H_i and H_j COM time series were then computed using the S and P pseudo-atoms of the three junction-

closing base pairs of each helix, and the canonical correlations of these time series were computed using the CCA library of R. The mean of the three orthogonal canonical correlation coefficients was used as the correlation expected along an arbitrary direction in 3D space.(16) This was repeated for both possible choices of H_r, as different H_r's remove global rotations and translations of the molecule with varied efficacy. The minimum of the two obtained mean correlations was chosen as the best representative. Time-randomizing the two COM variables and repeating the correlation analysis confirmed significance; in all cases, these time-randomized COM variables had $R \leq 0.005$.

4.2.5 Computing tertiary and coaxial stacking contacts

Two residues were considered to be in contact if the distance between their S beads was <14 Å and the residues were ≥5 apart in sequence number. Two loops were considered to be in contact if there was at least one residue-residue contact between the loops. The A/D-loop or V-loop was considered to be in contact with the D-stem if there was at least one residue-residue contact between the loop and the stem.

Coaxial stacking contacts between the D- and AC-stems, the T- and A-stems, and, if applicable, the V- and AC-stems, were determined using a set of criteria similar to that developed for all-atom structures (17). Our criteria are loosened to accommodate TOPRNA's coarse representation. The criteria were as follows: 1) The cosine of the angle between the base-pair-plane-normal vectors of the helix closing base-pairs is ≥0.7, where the base-pair-plane was determined as the least squares fit of all of the beads of the two paired nucleotides. 2) The distance between the centers-of-mass of the two closing base pairs is ≤9 Å for directly linked helices, or ≤14 Å for helices separated by one single-stranded nt. The centers of mass were

computed using only B beads. 3) The angles between each of the base-pair-plane normal vectors and the vector connecting the two pairs' centers of mass are both <60°.

The energetic cost of forming a residue-residue, loop-loop, or stacking contact, *c*, was computed as

$$\Delta G^{topo}(c) = -k_B T \ln\left(\frac{P(c)}{1 - P(c)}\right). \quad \textbf{[4.2]}$$

P(c) is the probability of observing the contact in a given simulation. The cost of forming joint contacts $c_1, c_2, \ldots, c_n$ was computed by substituting the joint probability $P(c_1, c_2, \ldots, c_n)$ into **[4.2]** above. The cooperativity *C* among a group of contacts was computed as

$$C(c_1,\ldots,c_n) = \frac{P(c_1,\ldots,c_n)}{\prod_n P(c_i)}. \quad \textbf{[4.3]}$$

4.2.6 Identification and analysis of best-packed conformers

Best-packed conformers from each simulation were identified by minimizing the energy function $E = n_l\varepsilon_l + n_s\varepsilon_s$, where n_l is the total number of residue-residue contacts between loops, n_s is the total number of coaxial stacks, and ε_l and ε_s are arbitrary scaling parameters. Unless otherwise indicated, ε_l was set to -0.6 k_BT and ε_s set to -3.5 k_BT. These values were chosen based loosely on the relative interaction energies expected to be contributed by non-specific residue-residue contacts and inter-helical stacking contacts. Choice of alternative ε_l and ε_s values had minimal effect on the identified conformers (**Figure C.4**).

The entropies of the 500 conformers with lowest *E* for each simulation were computed using an approach developed to estimate protein loop entropies (18). Conformers that are close

in structure to other conformers within this sub-ensemble will have higher entropies and thus indicate enriched regions of phase-space. The entropy S_i of conformer i is computed as

$$S_i = k_B \ln\left[1 + \sum_{j \neq i} \exp\left(\frac{-rmsd_{i,j}^{\ 3}}{(10\ \text{Å})^3}\right)\right]. \quad \textbf{[4.4]}$$

The sum is done over all conformers $j \neq i$, with $rmsd_{i,j}$ the RMSD in Å between i and j using all P beads unless otherwise noted. This functional form provides a smooth measure of the number of conformers within a ~10 Å volume of conformer i. The 10 Å radius was chosen based on its status as the P=0.01 cutoff for structural similarity in tRNA (19).

The overall specificity of the 500 most compact conformers of each tRNA species was quantified as the entropy-weighted fraction of compact conformers that are native-like,

$$\langle N \rangle = \frac{\sum_i N_i \exp\left(S_i / k_B\right)}{\sum_i \exp\left(S_i / k_B\right)}. \quad \textbf{[4.5]}$$

Here, S_i is the entropy of conformer i and N_i is 0 or 1 depending on whether i possesses contacts between the D and T loops and does not possess contacts that preclude a native-like 3D structure. These native-inconsistent contacts were defined as contacts between any two residues of the (A/D, AC), (A/D, ACCA), (D, AC), (D, ACCA), (D/AC, ACCA), (AC, V), (AC, T), (AC, ACCA), (V, ACCA), or (T, ACCA) loops. To maintain consistency when comparing across species, residues of the A/D, V, and ACCA loop residues were excluded from the RMSD calculations used to compute conformer entropies.

4.2.7 Folding tRNA with four restraints from an initially unfolded conformation

The potential for G26•A44, U8•A14, G15•C48, and G19•C56 tertiary restraints to determine tRNA's native 3D structure from unfolded conformations was probed through simulations that

hierarchically 'folded' the molecule from random configurations of the secondary structure. The restraints were enforced hierarchically in the order listed above with r_{min} and r_{max} identical to the values used for other restrained simulations. Force constants were initially set to kmax=kmin=0.1 kcal/mol/Å^2 and fmax=0.5 kcal/mol/Å, with the system allowed to equilibrate over 2×10^6 steps of Langevin dynamics at 300 K after the addition of each restraint. After all restraints were added, the system was allowed to further equilibrate over 10^5 steps of dynamics performed at each of 500 K, 450 K, 400 K, 350 K, and 300 K. Restraint force constants were then increased to kmax=kmin=4.0 kcal/mol/Å^2 and fmax=4.0 kcal/mol/Å and the system allowed to equilibrate over an additional 10^5 steps of dynamics at 300 K. Production simulations were performed at 300 K for a total of 10^6 dynamics steps, recording conformations every 10^3 steps, and the simulation-average structure computed.

In total, this procedure was performed for 200 different initial configurations, yielding 200 average structures. The generated average structures were then clustered according to the similarity of their measured $3\times(\alpha_h, \beta_h, \gamma_h)$ angles using the fixed-radius ART-2' clustering algorithm (20), modified to use cluster medoids instead of cluster means. Distances between two conformations *i* and *j* were computed as

$$d_{ij} = \frac{1}{3}\sum_{n=1}^{3}\left(\theta_{ij}\right)_n, \qquad \textbf{[4.6]}$$

where $(\theta_{ij})_n$ is the single-axis rotation amplitude (13) needed to bring the *n-th* $(\alpha_h, \beta_h, \gamma_h)$ of conformation *i* into concordance with the *n-th* $(\alpha_h, \beta_h, \gamma_h)$ of conformation *j*, and the clustering radius was set to 60°. This radius was chosen due to the observation that two different pairs of coaxially stacked helices, sourced from our previously constructed database of coaxially stacked helices within the PDB (2), may possess $(\alpha_h, \beta_h, \gamma_h)$ differing by as much as θ_{ij}=60°. Thus, a choice of 60° should allow two such equivalent conformations to cluster together.

Multiple independent clusterings computed from different initial orderings of the structures consistently yielded 7 clusters, with the most populous cluster (35%-41% of the population) always comprised of native conformations. The next most populous cluster contained no more than ~26% of the average structures. The members of non-native clusters represent different topoisomers that arise from the non-equilibrium folding methodology. While able to fulfill the four applied distance restraints, these structures were clearly not feasible RNA folds. For example, only structures in the native cluster were able to form any type of inter-helical stacking interactions.

4.3 Results

4.3.1 Topological constraints restrict tRNA global conformation and give rise to inter-helical correlations

We explored the topological constraints posed by the secondary structure of $tRNA^{Phe}$ (hereafter referred to as WT tRNA) using temperature replica exchange molecular dynamics simulations of the TOPRNA coarse-grained model (**Figure 4.1A**). These simulations showed good convergence after 10^8 steps of dynamics per replica, but were extended to a total of 10^9 steps to achieve the best sampling possible. To analyze the global conformations sampled by our simulations we use Euler angles, (α_h, β_h, γ_h), to describe the relative orientation between pairs of tRNA helices (1,2,13). Using the AC- and T-stems as an example, Euler angles specify the twist angles α_h around the T-stem and γ_h around the AC-stem, while β_h specifies the bend angle between them (**Figure 4.1B**). While a total of six sets of pairwise Euler angles exist between tRNA's four helices, only three are needed to uniquely define the orientation of three helices relative to a reference helix, arbitrarily chosen here to be the AC-stem (13).

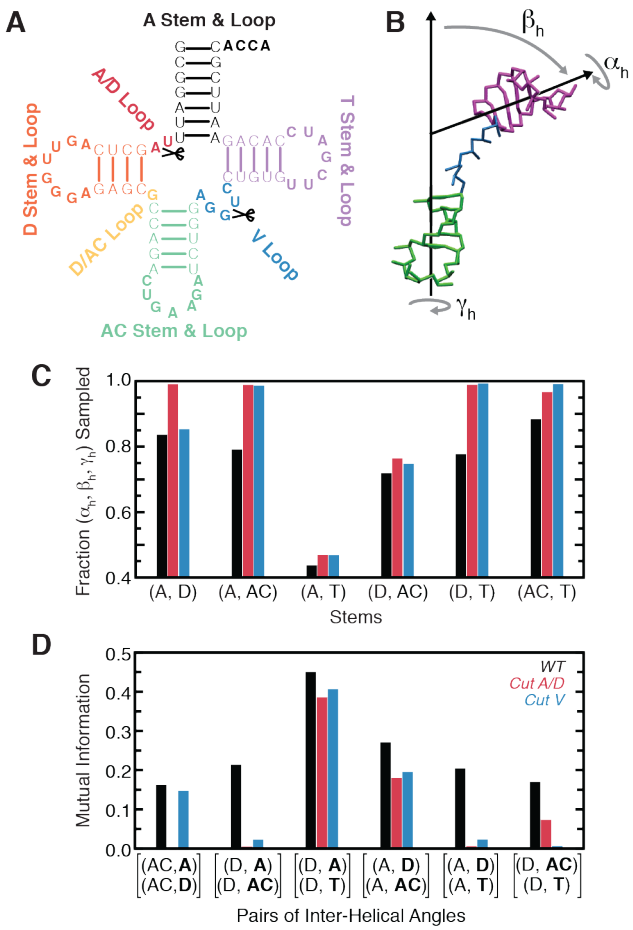

Figure 4.1: Secondary structure limits the set of global conformations accessible to tRNA
(A) Secondary structure and labeling scheme of $tRNA^{Phe}$. Loop residues are bolded and cut locations marked. **(B)** The Euler angle convention used to describe the relative 3D orientation of RNA helices. Shown is a representative TOPRNA snapshot of the AC- and T-stems, colored as in **(A)**, with the A- and D-stems and connecting loops not shown for clarity. **(C)** Fraction of possible (α_h, β_h, γ_h) angles sampled between pairs of tRNA helices by the WT (black), cut A/D-loop (red), and cut V-loop (blue) simulations. **(D)** The mutual information (MI) between pairs of inter-helical Euler angles measured with respect to a common reference helix. The two helices whose orientations are being correlated are bolded.

In **Chapter 3** we showed that topological constraints restrict the relative orientations of helices to as little as 7% of the total theoretical (α_h, β_h, γ_h) space for 1-nt bulges and as much as 62% for an infinitely long bulge (1,2,4). The connectivity constraints posed by single-stranded bulge linkers and the pivoted topologies of these motifs also give rise to correlations between the helical twist angles α_h and γ_h (1,2). In WT tRNA, the four-way junction constrains the relative orientation between pairs of helices to a lesser extent (43%-88%) than two-way junctions (7-

62%) (**Figure 4.1C**). Interestingly, these reduced constraints allow helices to sample orientations that are forbidden in two-way junctions. For example, ~7% of the (α_h, β_h, γ_h) sampled between the AC- and T-stems would be inaccessible to any type of two-way junction (1,2). These unique orientations become accessible because helices in higher-order junctions are no longer necessarily translationally constrained by well-defined pivots, effectively reducing their steric constraints.

As expected, the range of orientations sampled by helices increases with the length of the adjoining linker strands (**Figure 4.1C**). Thus, the A- and D-stems, which are separated by two nts, sample 82% of the (α_h, β_h, γ_h) space, whereas the A- and T-stems, which are linked by a pivot, are limited to only 43% of their possible relative orientations. Helices that are only indirectly linked, such as the D- and T-stems, are also constrained and sample <80% of their possible relative orientations. Comparisons to simulations of tRNA with cut A/D- or V-loops (**Figure 4.1A, C**) show that helices linked by two or fewer single-stranded nts are primarily constrained by local sterics; the short linkers translationally restrain the helices such that they cannot diffuse away and are unaffected by distal cuts of the junction (**Figure 4.1C**). By contrast, the constraints on all other pairs of helices depend on junction connectivity, with ~100% of possible conformations sampled upon junction cutting.

Although individual pairs of helices in WT tRNA are less constrained than in two-way junctions, the system as a whole is more constrained than is apparent from the pairwise analysis. As noted above, the global conformation of tRNA is described by three joint sets of Euler angles. If the helices behaved independently, the fraction of conformations sampled within this 9D angular space would equal the product of the fractions sampled by each of the three helix pairs individually. Instead, the ratio of these quantities is $r_{9D/3\cdot3D}\approx0.06$, indicating that the orientations

of tRNA's helices are coupled together. Indeed, based on mutual information (MI) measures, we observe small to moderate correlations (MI>0) between all pairs of inter-helical (α_h, β_h, γ_h) angles (**Figure 4.1D**). These correlations are also apparent in coordinate-space, with the centers of mass of different helices correlated with R≈0.2 to R≈0.6 (**Figure C.1**). Thus, once the orientation of two helices is defined, it poses constraints on the orientations of other helices either due to long-range steric effects and/or conformational restriction of the linker single-strands. For example, coaxial stacking of the D- and AC-stems anchors the termini of the A/D- and V-loops, shortening the effective linker between the A- and T-stems and thereby limiting their conformational freedom (**Figure C.2**). When stacked atop the AC-stem, the excluded volume of the D-stem also precludes twisted orientations of the A-stem relative to the AC-stem (**Figure C.2**). These correlations are significantly diminished in the cut tRNAs; the $r_{9D/3\text{-}3D}$ ratio increases to ~0.15 in both, and we observe dramatically reduced MI between different inter-helical angles as well as reduced correlations between the centers of mass of helices (**Figure 4.1D**, **C.1**).

Together, these correlations serve to constrain the relative orientations of WT tRNA's helices to ~6% of the theoretically possible conformations in 9D Euler space. Thus, the ability of topological constraints to induce long-range correlations between all four helices helps compensate for the decreased constraints experienced between pairs of helices.

4.3.2 Topological constraints prevent tRNA from forming non-native well-packed conformations

Given that there are no attractive interactions in TOPRNA, the simulation of WT tRNA spends the majority of its time in entropically favored extended conformations. However, conformations

that form long-range inter-loop contacts are transiently sampled, including native-like conformations (5×10^{-4} of conformations are <10 Å root-mean-square-deviation (RMSD) from the crystal structure). From their studies of helices linked by polyethylene glycol tethers, Herschlag and coworkers proposed that topological constraints could prevent an RNA from forming non-native tertiary contacts and thus contribute to the specificity of tertiary folding (3). To test this possibility in WT tRNA, we computed the probability $P(r_i,r_j)$ at which two loop residues r_i and r_j come within a S-S bead distance cutoff of 14 Å, the approximate distance between two Watson-Crick (WC) paired residues. $P(r_i,r_j)$ was then converted to a free energy $\Delta G^{topo}(r_i,r_j)$ that defines the energetic cost of forming the distance-dependent contact. As shown in **Figure 4.2A**, topological constraints pose a penalty as large as ~8 k_BT (~5 kcal/mol at 300 K) for bringing different regions of WT tRNA into proximity. Strikingly, native tertiary contacts (outlined in black in **Figure 4.2A**) are specifically topologically favored, forming with the smallest ΔG^{topo} penalty (2-6 k_BT). By contrast, non-native contacts are typically discriminated against via large ΔG^{topo} penalties (>7 k_BT). It is worth noting that this several k_BT difference is similar in magnitude to the -1 to -5 k_BT stability of WC base pairs (21).

Comparisons to the cut tRNAs reveal that the specificity for native contacts is a unique property of WT topological constraints (**Figure 4.2B**, **C.3**). Notably, the ΔG^{topo} penalty for forming non-native contacts decreases by as much as -3 k_BT in the cut tRNAs. By contrast, the ΔG^{topo} of forming native 'core' contacts between the A/D-, V-, and D-loops is increased by 1-2 k_BT, and remains roughly unchanged for other native contacts. Thus, the ability of topological constraints to discriminate against non-native contacts is substantially reduced.

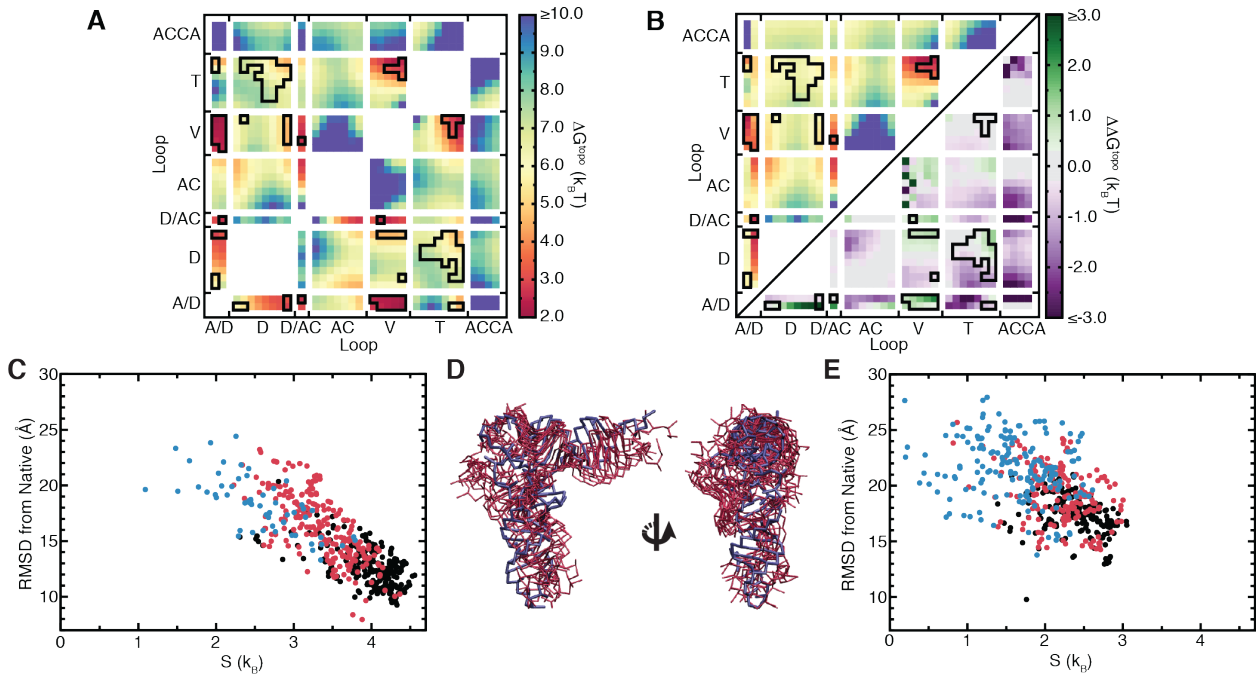

Figure 4.2: Secondary structure prevents tRNA from forming non-native tertiary contacts
(A) Free energy cost of forming different inter-loop residue-residue contacts in WT tRNA. Contacts observed in the crystal structure are outlined in black. **(B)** The free energy cost of forming different contacts upon cutting the A/D-loop. The ΔG^{topo} is shown in the upper left triangle using the same color scale as **(A)**. In the lower triangle, the $\Delta\Delta G^{topo}$ between the cut A/D-loop relative to WT tRNA is shown, with the color scale to the right. **(C)** Entropies and all P-bead RMSDs of the 500 best-packed conformers sampled by WT tRNA. Conformations that possess only native-consistent contacts and have D-T loop-loop contacts are colored black, those that possess only native-consistent contacts but lack D-T loop-loop contacts colored red, and those that possess native-inconsistent contacts are colored blue (see methods). Note that high entropies indicate conformers that are thermodynamically favored. **(D)** Superposition of the crystal structure (blue) and the five highest entropy best-packed conformers from the WT tRNA simulation (red). **(E)** Entropies and all-P bead RMSDs of the 500 best-packed conformers sampled by cut A/D-loop tRNA. The color scheme is the same as in **(C)**.

While the above results indicate that individual native tertiary contacts are specifically topologically favored in WT tRNA, they do not necessarily imply that topological constraints favor native over non-native 3D conformations. Instead, one must consider the penalty of forming many tertiary contacts *simultaneously*. To explore this question, we identified the 500 WT tRNA conformers that have the maximal number of long-range contacts between single-stranded loops and inter-helical stacking interactions, which we term the 'best-packed' conformers (see **Section 4.2.6**). It is important to emphasize that this procedure equally weights native and non-native contacts, and ignores both sequence and local geometry. Remarkably,

despite this naïve identification procedure, we find that this ensemble of best-packed conformers is highly enriched in native-like conformers (**Figure C.4**). The extent to which conformers are close in structure to other conformers in this best-packed subset also indicates whether they can be readily accessed, and thus are entropically favored, or conversely whether they are rarely accessed and thus entropically disfavored (18). Computing these entropies reveals that topological constraints strongly funnel WT tRNA's free energy landscape towards native-like conformations (**Figure 4.2C**). Conformers that have comparatively low RMSD from the native structure and have only native-consistent contacts (for example no contacts between the D- and AC-loops; see **Section 4.2.6**) have significantly higher entropies. In fact, the five conformers with highest entropy have native-state RMSDs ranging from 10.9 Å to 13 Å (**Figure 4.2D**). This approaches the 10 Å RMSD threshold that is a significant prediction of tRNA 3D structure (19), despite our model treating loops as freely rotatable chains and completely ignoring sequence.

By contrast, applying the same procedure to the cut tRNAs shows that best-packed conformations are significantly less enriched in native-like conformations and are less funneled towards the native state (**Figure 4.2E, C.5**). In combination with **Figures 4.2C** and **C.3**, these findings lead to the prediction that cutting one strand in tRNA should decrease thermodynamic stability. Significantly, this agrees with experiments showing that cuts anywhere within the A/D-loop or at the V-loop termini catastrophically disrupt the ability of tRNA to fold (22). The same experiments also showed that cuts to D/AC-loop and T/A-linker also reduced folding to a lesser degree; while not simulated here, this is consistent with our expectation that such cuts should similarly disrupt WT tRNA's topological constraints. However, in what is likely a result of stabilization afforded by tertiary base-triples, cuts to the interior of the V-loop did not significantly affect folding even though we predict they should comparably disrupt topological

constraints (22). Taken together, these observations support that topological constraints constitute an important component of the tRNA folding landscape, emphasizing that tertiary interactions and electrostatics also play key roles. Notably, the particular severity of A/D-loop cuts is likely due to the >3 k_BT increase in the ΔG^{topo} of forming the crucial U8•A14 tertiary interaction in addition to the overall decrease in topological constraints (**Figure 4.2B**).

4.3.3 Topological constraints render some tertiary interactions redundant and help direct tRNA dynamics along specific pathways

We next examined the consequences of combining the specific conserved tertiary interactions of tRNA with topological constraints (**Figure 4.3**). Still treating loop residues as freely rotatable chains, we performed a TOPRNA simulation of WT tRNA restrained by these nine tertiary interactions using simple residue-residue distance restraints ($tRNA_{9R}$). $tRNA_{9R}$ is effectively constrained to only native-like global conformations (**Figure 4.3B**); the average structure is ~8 Å RMSD from the crystal structure (**Figure 4.3C**), and conformations with high native RMSDs can be attributed to global twisting and bending motions of helices around this average structure (**Figure C.6**). Compared to the distribution of best-packed conformations in the unrestrained simulation of WT tRNA (**Figure C.4**), non-native conformations are eliminated and the population of low RMSD conformations is increased in $tRNA_{9R}$. Thus, native tertiary interactions effectively stabilize the set of best-packed conformations favored by topological constraints and further funnel them towards the native conformation. The need for only native secondary structure and several tertiary interactions to define macroscopic structure in this manner may explain why there are comparatively few sequence constraints on other regions of tRNA (**Figure 4.3A**) (23,24).

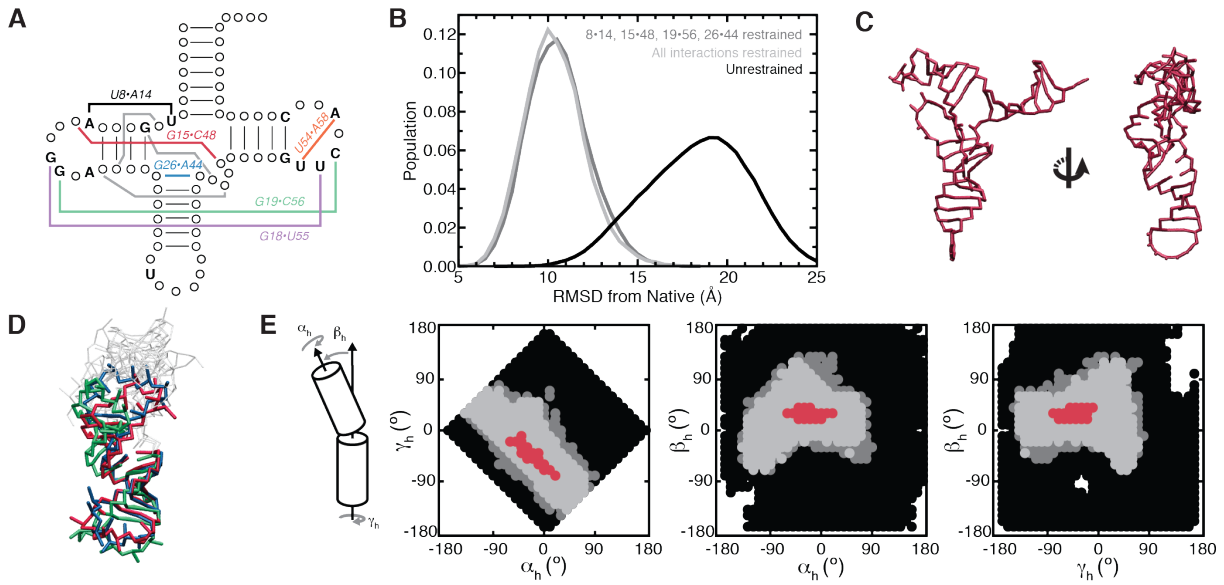

Figure 4.3: Tertiary interactions confine tRNA to native-like conformations
(A) Diagram of conserved tRNA residues and tertiary interactions. Residues with <90% conservation are indicated by circles (23). Conserved tertiary interactions are labeled and semi-conserved base triples are drawn as gray lines. **(B)** RMSD distributions of simulations of unrestrained WT tRNA (black), WT tRNA restrained by all nine conserved tertiary interactions ($tRNA_{9R}$; light gray), and WT tRNA restrained by the four non-redundant interactions ($tRNA_{4R}$; dark gray). **(C)** The average structure of $tRNA_{9R}$. **(D)** Three representative structures from the $tRNA_{9R}$ simulation illustrating the orientations sampled between the D- and AC-stems. Structures are superimposed by the AC-stem. Residues of the A- and T-stems and connecting loops are not colored for clarity. **(E)** 2D projections of the (α_h, β_h, γ_h) angles sampled between the AC- and D-stems. Angles only sampled by unrestrained tRNA are shown in black; angles sampled by both unrestrained and $tRNA_{4R}$ are shown in dark gray; and angles sampled by $tRNA_{9R}$, $tRNA_{4R}$, and unrestrained WT tRNA are shown in light gray. Red points correspond to angles measured from 109 different tRNA crystal structures. A reference cartoon of the three angles is shown at left. Note that as discussed in the text, examining only one pairwise set of (α_h, β_h, γ_h) angles provides an incomplete picture of the extent to which topological constraints confine tRNA conformation.

We also examined whether the full set of tertiary interactions is needed to restrict WT tRNA to native conformations. Additional restrained simulations indicated that the base-triple, U54•A58, and either the G18•U55 or G19•C56 restraints could all be removed without an increase in the mean native RMSD ($tRNA_{4R}$; **Figure 4.3B**). Of the remaining 'non-redundant' U8•A14, G15•C48, G26•A44, and G19•C56 restraints, all but the U8•A14 restraint could be singly removed with the mean native RMSD increasing by only 1-2 Å (not shown).

These results indicate that, at the coarse level of our models, topological constraints render some of tRNA's tertiary interactions structurally redundant. While the favorable

interaction energies contributed by the 'non-redundant' contacts surely play an important role in stabilizing the native fold, they are unnecessary for defining tRNA global architecture. A caveat is that the above simulations were begun from the native conformation, and thus the full set of interactions could still be necessary to specify the native state from unfolded conformations. We tested this by hierarchically 'folding' tRNA molecules from random initial configurations using only the four non-redundant tertiary restraints (see **Section 4.2.7**). Repeating this 200 times and clustering the product 'folds' revealed that the most populous cluster was native (**Figure C.7**). Thus, coupled with the entropic bias of topological constraints, the non-redundant interactions are sufficient to specify global native structure. Notably, the apparent redundancy in tRNA's tertiary interaction network is consistent with evidence that not all conserved tertiary pairs are required for folding. Some cytosolic and many mitochondrial tRNAs lack a subset of conserved pairs (24,25), and individual ablations (24,26-28) or complete reengineering (29) of the tertiary interactions of canonical tRNAs do not inhibit function. Our results also complement prior studies that showed that tRNA 3D structure can be predicted based on these non-redundant interactions (30-34).

As noted above, a marked feature of both the $tRNA_{4R}$ and $tRNA_{9R}$ simulations is that tRNA retains significant structural flexibility. This flexibility would likely be reduced upon inclusion of energy terms beyond topological constraints. Nevertheless, it is interesting to note that this flexibility appears to be directed along specific motional modes that are qualitatively similar to those identified by more rigorous studies of tRNA dynamics (35-38). These include global bending and twisting motions of the two arms of the tRNA 'L', and bending and twisting of the D- and AC-stems relative to one another (**Figures 4.3D, 4.3E, C.6**, **C.8**). It is well known that such motions are integral to tRNA function, including recognition by aminoacyl synthetases

and initial selection and translocation on the ribosome (38,39). Thus, similar to how topological constraints and a few tertiary interaction constraints are sufficient to encode macroscopic structure, these constraints are also sufficient to macroscopically define biologically important dynamics.

4.3.4 Topological constraints give rise to folding cooperativity

The apparent correlations between the orientation of WT tRNA helices (**Figure 4.1D**, **C.1**) together with the redundancy in tRNA's tertiary interaction network (**Figure 4.3B**) suggests that topological constraints may provide a basis for folding cooperativity. Here, formation of a subset of native tertiary interactions is accompanied by greater structural confinement and a reduction in the entropic penalty for forming remaining native tertiary interactions. This mechanism could help explain the tertiary folding cooperativity exhibited by tRNA (40) and other RNAs (11,41), and parallels the explanation given for folding cooperativity in proteins (42).

We quantified this cooperativity from our unrestrained simulations by computing the ratio of the joint probability $P([l_i, l_j], [l_k,l_m], s_i, \ldots)$ of forming contacts between different pairs of loops $[l_i, l_j]$ and $[l_k,l_m]$ and inter-helical stacks s_i, to the product of the individual probabilities (see methods). This coarse analysis confirms that, on average, different contacts form cooperatively, and that cooperativity increases as more contacts are formed (**Figure 4.4A**). Notably, combinations of six native contacts form on average with a ~200× greater probability than if they were independent. The cooperativity among native contacts is also as much as 6× higher than between combinations of non-native contacts. Thus, topological constraints also help prevent non-native contacts from forming cooperatively. By contrast, the cut tRNAs exhibit an

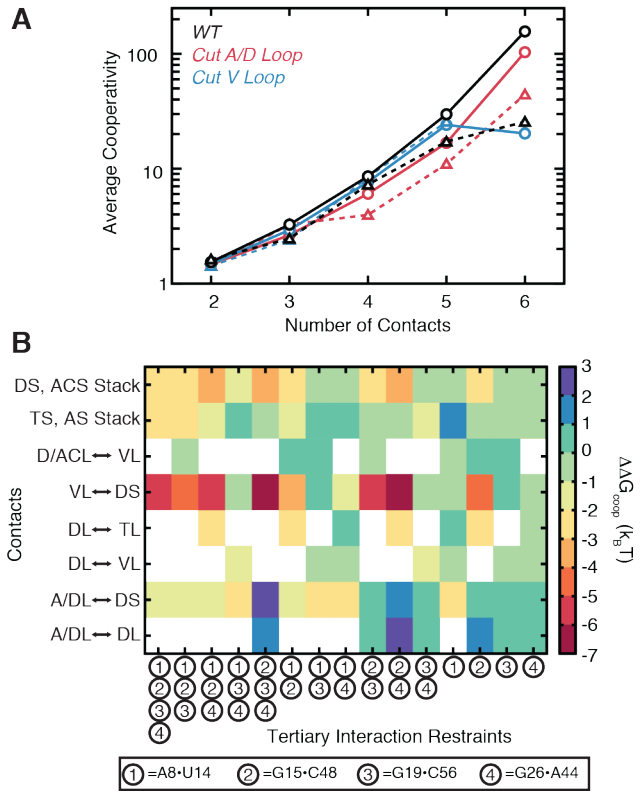

Figure 4.4: tRNA's tertiary interaction network is cooperative
(A) The mean cooperativity of jointly forming *n* number of loop-loop or stacking contacts, averaged over all combinations of native contacts (solid line, circles) or combinations containing at least one non-native contact (dashed line, triangles). Results are shown for the WT (black), cut A/D-loop (red), and cut V-loop (blue) simulations. Combinations that were observed ≤10 times were excluded from the averages. **(B)** The thermodynamic cooperativity among tRNA's tertiary interactions, computed with equation **[4.7]**. Restrained tertiary interactions are numbered on the x-axis according to the shown key. Loop-loop, loop-stem, and stacking contacts for which cooperativities were measured are shown along the y-axis. Cooperativities are not computed for loops that have an active restraint placed between them.

---

overall decrease in cooperativity, and perhaps more importantly, a narrower difference in the cooperativities of native vs. non-native contacts (**Figure 4.4A**).

We also computed the energetic consequences of cooperativity on native folding in WT tRNA by taking the difference between the ΔG^{topo} of forming a set of contacts $\{TC\}$ in a simulation restrained by tertiary interactions $\{TI\}$, relative to the ΔG^{topo} of forming $\{TC\}$ without restraints:

$$\Delta\Delta G_{coop}\left(\{TC\},\{TI\}\right) = \Delta G^{topo}\left(\{TC\}\right)_{\{TI\}restrained} - \Delta G^{topo}\left(\{TC\}\right)_{unrestrained}. \qquad \textbf{[4.7]}$$

In agreement with our analysis of the unrestrained simulations, single interactions by themselves only weakly influence the stability of other contacts (**Figure 4.4B**). However, as more interactions are restrained, formation of other contacts becomes increasingly energetically favored. Particularly notable is that the G15•C48 tertiary pair contributes up to -7 k_BT to the stability of contacts between the V-loop and the D-stem, potentially explaining why the base triples that form between these regions have few sequence constraints (43), and consistent with the above observed redundancy of these triples. The G15•C48 pair also promotes inter-helical stacking, both individually (-3 $k_BT < \Delta\Delta G_{coop} < 0$ k_BT) and more strongly in conjunction with other tertiary pairs (-4 $k_BT < \Delta\Delta G_{coop} < -2$ k_BT). By contrast, the U8•A14 pair generally disfavors formation of additional contacts, and vice versa, other tertiary restraints disfavor formation of A/D- to D-loop contacts. This anti-cooperativity arises because of the large increase in steric constraints associated with bringing the A- and D-stems into close proximity; tRNA conformations that have both the U8•A14 pair and other native contacts are thus entropically disfavored. The inability of U8•A14 to be stabilized by other interactions, coupled with its importance for confining tRNA to its native-state (see above), may explain its particularly strong evolutionary conservation (23,24).

4.3.5 Topological constraints are conversed across diverse tRNA secondary structures

While the large majority of tRNA species share the classic cloverleaf secondary structure explored above, there are several commonly observed variations (24). The most common variation involves a decrease in the length of the V-loop from 5-nt to 4-nt (**Figure 4.5A**). Not too surprisingly, TOPRNA simulations of a tRNA with 4-nt V-loop (VL4) reveals that this molecule

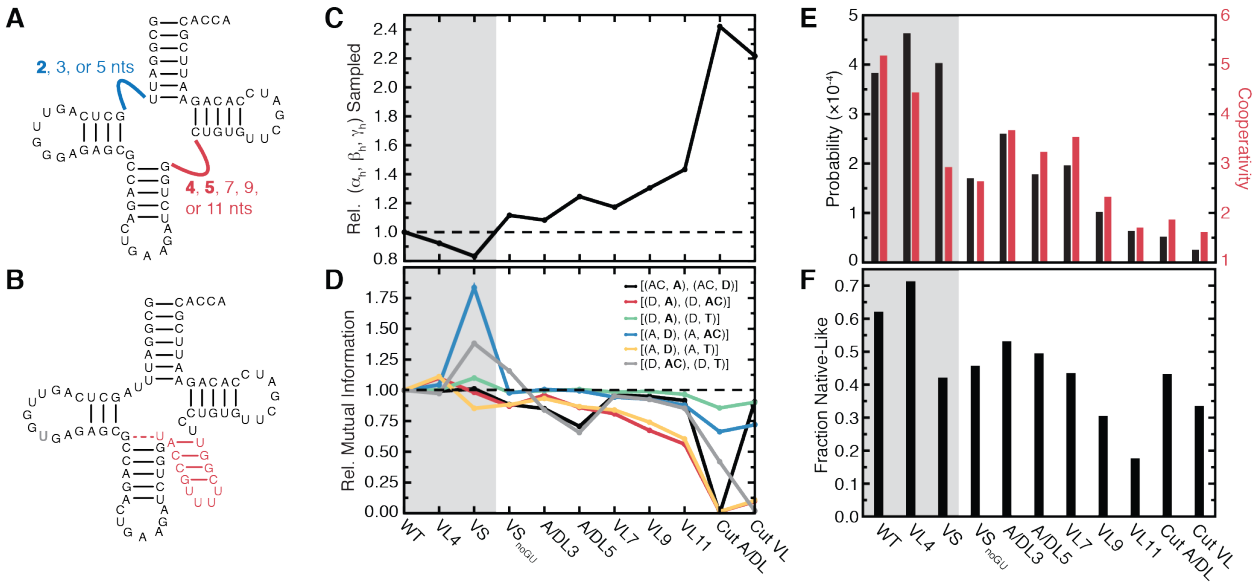

Figure 4.5: Naturally occurring tRNA secondary structures conserve topological constraints
(A) Class I tRNAs. The A/D and V-loop are shown as blue and red lines respectively, with tested length variations labeled. Naturally observed lengths are bolded. Full sequences are shown in **Figure C.9**. **(B)** Example Class II tRNA. Inserted V-stem is shown in red, and G26•U44 pair shown by a dashed line. Note the additional D-loop nt (gray) and 3-bp D-stem. **(C)** Relative fraction of 3×(α_h, β_h, γ_h) inter-helical conformations sampled by different tRNAs compared to WT. **(D)** Mutual information between different pairs of inter-helical Euler angles relative to WT tRNA. The helices whose orientations are being correlated are bolded in the key. **(E)** The probability and cooperativity of jointly forming loop-loop contacts between the D and T loops and both native inter-helical stacks. **(F)** The fraction of 500 best-packed conformers that possess native-like folds, weighted by entropy as described in methods. The gray background in **C**-**F** is used to highlight natural tRNA variants.

---

is similarly if not even more confined by topological constraints, and has similar native-state specificity (**Figure 4.5**).

A more dramatic variation is found in so-called Class II tRNAs, where the V-loop is replaced by a several base-pair-long V-stem (**Figure 4.5B**). This is normally accompanied by replacement of the tertiary G26•A44 pair atop the AC-stem with a more stable G26•U44 pair, as well as several changes to the D-loop (**Figure 4.5B**) (23). Despite these secondary structure differences, Class II tRNAs fold to a common 3D structure through poorly understood mechanisms. Strikingly, TOPRNA simulations of a Class II tRNA (VS) reveal that topological constraints are conserved, with the overall number of interhelical conformations sampled decreasing by ~17% and the probability of jointly forming stacking interactions and D- and T-

loop contacts increasing by ~5% (**Figure 4.5C, E**). However, native contacts form less cooperatively and the best-packed conformations of VS are substantially less native-specific, suggesting that the folding landscape of this tRNA may be more complex than Class I species (**Figure 4.5**).

Inversely, we find that changes in tRNA secondary structure that would be expected to disrupt topological constraints are evolutionarily disfavored. While an entire stem can replace the V-loop in Class II tRNAs, it is very rare to observe tRNAs with single-stranded V-loops longer than 5-nts (**Figure 4.5A**) (23,24). Consistent with the negative selection of such secondary structures, simulations of Class I tRNAs with 7-nt (VL7), 9-nt (VL9), and 11-nt (VL11) long V-loops indicate that these changes decrease topological constraints (decrease by 17-43%; **Figure 4.5C**) and reduce native-state specificity (decrease by 33-75%; **Figure 4.5E, F**). By comparison, Class II tRNAs preserve WT-like topological constraints by sequestering additional V-loop nucleotides into a hairpin.

Topological constraints also help explain why the non-canonical G26•A44 pair of Class I tRNAs is replaced in most Class II tRNAs by a more stable G26•U44 pair (**Figure 4.5B**) (23). Removing the G26•U44 base-pair from VS (VS_{noGU}) reduces topological constraints by ~30% compared to VS and significantly reduces the likelihood of forming native tertiary contacts (**Figure 4.5**). This result also helps explains why U44 is often 2'-O-methylated in Class II tRNAs that contain G26•U44 pairs. Significantly, although this modification is thought to function by locally stabilizing the G26•U44 pair, its absence has been shown to globally destabilize Class II tRNAs (44,45).

Finally, we explored the effects of lengthening the A/D-loop, which is universally conserved in all tRNAs to be ≤2 nt long (**Figure 4.5A**) (23,24). Simulations of Class I tRNAs

with 3-nt (A/DL3) and 5-nt (A/DL5) A/D-loops reveal that topological constraints and native specificity are modestly decreased in these molecules (8-50%, depending on tRNA and property; **Figure 4.5**). While consistent with the evolutionary preference for 2-nt or less A/D-loops, these modest decreases are likely insufficient to explain why such secondary structures are never observed in nature. Among many possible explanations, additional nucleotides may disrupt the native tertiary interactions that form between the A/D-loop and the D-loop and stem, or lead to misfolded secondary structures.

4.4 Discussion

Our results show that topological constraints encoded at the secondary structure level provide a robust strategy for encoding macroscopic properties of RNA 3D structure and dynamics. In tRNA, topological constraints serve as a source of negative design (46,47) by imposing significant penalties on the formation of non-native tertiary contacts. This is in strong agreement with the hypothesis of Herschlag and colleagues that topological constraints provide a mechanism for circumventing the limited specificity of RNA's nucleotide alphabet (3). By coupling the orientation of helices together over long length scales, topological constraints also allow a tertiary interaction formed in one region of an RNA to influence the likelihood of forming additional distant tertiary interactions. This provides a source of folding cooperativity that helps stabilize tRNA's tertiary structure. It is important to emphasize that topological constraints are largely sequence independent. Thus, for tRNA, these properties are inherited by any species that maintains an appropriate secondary structure. This may help free the primary sequence of tRNA to vary according to other functions orthogonal to folding (39).

As discussed above, our findings are consistent with, and can help explain, many prior experiments on tRNA. It is particularly satisfying that our results help explain why evolution has conserved the Class I and Class II isoforms of tRNA secondary structure, but strongly selects against secondary structures with V-loops containing more than five single-stranded residues. Our results may also explain the thermodynamic coupling between secondary and tertiary structure folding that is observed in some tRNAs (48-53). Experimental (54) and computational (55) studies of tRNA suggests that folding of native secondary structure nucleates tertiary structure folding. While not true for all RNAs (56,57), this is consistent with our hypothesis that secondary structure exerts a powerful influence on tertiary structure stability. This is further supported by experiments showing that stabilizing secondary structure can rescue folding of tRNA mutants with disrupted tertiary interactions (53), and by the clear implication that increased GC sequence content stabilizes thermophilic tRNA species (24,58).

A growing body of literature has suggested the importance of junction secondary structure to the folding of other RNAs. Correlations between the length of single-strands in junctions and their folded conformation have been identified and used with some success to predict RNA 3D conformation (9,10,59-62). Paralleling our results, Sim and Levitt found that fragment assembly models built from secondary structure were biased towards the native conformation (8). Experiments have also shown that the junction of the hairpin ribozyme modulates the thermodynamics of tertiary folding, primarily by altering the entropic cost of folding (63-66). However, the physical basis for these observations has remained unclear. Combined with prior studies of two-way junctions (1-4), our work indicates that topological constraints provide a free energy based framework for understanding the link between secondary structure and 3D conformation.

The ability of topological constraints to discriminate against non-native tertiary interactions and encode cooperativity may be particularly important for large RNAs. Notably, cooperativity similar to what we find in tRNA has been shown to be critical to the tertiary structure stability of large RNAs (11,41,67). These RNAs also often utilize multiple identical tertiary interaction motifs to stabilize their 3D folds, implying that factors beyond sequence code for their specificity (68,69). Large RNAs also fold through native-like compact intermediates that lack fully formed tertiary interactions (70), and are stabilized by molecular crowders that non-specifically favor compact 3D conformations (71,72). In a unique case, a segment of the HIV-1 genome RNA was shown to adopt a well-defined solution structure despite not having well-defined tertiary contacts (73). These observations are consistent with secondary structure providing an inherent source of 3D folding specificity. Although not explored here, we note that topological constraints could also play important kinetic roles in RNA folding. For example, progressive formation of helices and accompanying topological constraints could bias folding along specific pathways (74,75).

Clearly, topological constraints operate only on a coarse level. Other forces, including electrostatics and sequence-specific attractive interactions, must be considered in order to achieve an atomistic level of understanding of RNA structure and dynamics. The assumption of fixed RNA secondary structure used in our study also makes it difficult to explore thermodynamic coupling between secondary structure and tertiary structure folding. However, the additivity of free energy ensures that the specificity and cooperativity encoded by topological constraints will translate to real RNAs. For promiscuous tertiary interactions, the differential energetic costs posed by topological constraints could be a primary factor in determining the folding outcome.

4.5 Conclusions

In this Chapter, we used TOPRNA to probe the thermodynamic contributions of topological constraints towards defining the 3D architecture and dynamics of tRNA. We demonstrated that topological constraints strongly limit tRNA's four-way junction to ~6% of its theoretically possible interhelical conformations. Furthermore, topological constraints impose up to ~6 kcal/mol energetic penalties on forming different tertiary contacts. Strikingly, these variable free energy penalties appear to be harnessed by tRNA to discriminate against formation of non-native tertiary contacts, providing a sequence-independent source of folding specificity. Indeed, these penalties make it such that tRNA global 3D structure can be predicted simply by maximizing long-range interloop packing.

We also demonstrate that topological constraints give rise to long-range correlations between the relative orientation of tRNA's helices, which in turn provides a mechanism for encoding thermodynamic cooperativity between distinct tertiary interactions. Such cooperativity has long been observed and known to be important to RNA tertiary folding, but its basis unclear. The correlations between tRNA helices also make it such that only several tertiary interactions are needed to confine tRNA to its native global structure and specify functionally important 3D dynamics.

Finally, in agreement with prior experimental data, we showed that disrupting topological constraints should destabilize tRNA molecules. Moreover, we showed that tRNA junction architectures that have reduced topological constraints are evolutionarily disfavored, whereas naturally occurring variations of tRNA secondary structure preserve

topological constraints. This suggests that topological constraints underlie the evolutionary conservation of tRNA secondary structure.

Taken together, our results demonstrate that secondary-structure-encoded topological constraints play a central role in stabilizing and specifying RNA tertiary structure and dynamics.

4.6 References

1. Bailor, M.H., Sun, X.Y. and Al-Hashimi, H.M. (2010) Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation. *Science*, **327**, 202-206.
2. Mustoe, A.M., Bailor, M.H., Teixeira, R.M., Brooks, C.L., III and Al-Hashimi, H.M. (2012) New insights into the fundamental role of topological constraints as a determinant of two-way junction conformation. *Nucleic Acids Res*, **40**, 892-904.
3. Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S. and Herschlag, D. (2009) Do conformational biases of simple helical junctions influence RNA folding stability and specificity? *RNA*, **15**, 2195-2205.
4. Mustoe, A.M., Al-Hashimi, H.M. and Brooks, C.L, III (2014) Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges. *J Phys Chem B*, **118**, 2615–2627.
5. Gesteland, R.F., Cech, T.R. and Atkins, J.F. (2006) *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor Laboratory Press.
6. Dethoff, E.A., Chugh, J., Mustoe, A.M. and Al-Hashimi, H.M. (2012) Functional complexity and regulation through RNA dynamics. *Nature*, **482**, 322-330.
7. Cruz, J.A. and Westhof, E. (2009) The dynamic landscapes of RNA architecture. *Cell*, **136**, 604-609.
8. Sim, A.Y.L. and Levitt, M. (2011) Clustering to identify RNA conformations constrained by secondary structure. *Proc Natl Acad Sci USA*, **108**, 3590-3595.
9. Laing, C. and Schlick, T. (2009) Analysis of four-way junctions in RNA structures. *J Mol Biol*, **390**, 547-559.
10. Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83-93.
11. Behrouzi, R., Roh, J.H., Kilburn, D., Briber, R.M. and Woodson, S.A. (2012) Cooperative tertiary interaction network guides RNA folding. *Cell*, **149**, 348-357.
12. Feig, M., Karanicolas, J. and Brooks, C.L.I. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model*, **22**, 377-395.
13. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) 3D maps of RNA interhelical junctions. *Nat Protoc*, **6**, 1536-1545.

14. Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E.K., Blazewicz, J. and Adamiak, R.W. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, **11**, 231.
15. Steuer, R., Kurths, J., Daub, C.O., Weise, J. and Selbig, J. (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18 Suppl 2**, S231-240.
16. Briki, F. and Genest, D. (1994) Canonical analysis of correlated atomic motions in DNA from molecular dynamics simulation. *Biophys Chem*, **52**, 35-43.
17. Tyagi, R. and Mathews, D.H. (2007) Predicting helical coaxial stacking in RNA multibranch loops. *RNA*, **13**, 939-951.
18. Xiang, Z., Soto, C.S. and Honig, B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci USA*, **99**, 7432-7437.
19. Hajdin, C.E., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2010) On the significance of an RNA tertiary structure prediction. *RNA*, **16**, 1340-1349.
20. Karpen, M.E., Tobias, D.J. and Brooks, C.L., III. (1993) Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry*, **32**, 412-420.
21. Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) RNA structure prediction. *Annu Rev Biophys Biophys Chem*, **17**, 167-192.
22. Pan, T., Gutell, R.R. and Uhlenbeck, O.C. (1991) Folding of circularly permuted transfer RNAs. *Science*, **254**, 1361-1364.
23. Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res*, **37**, D159-162.
24. Marck, C. and Grosjean, H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189-1232.
25. Helm, M., Brule, H., Friede, D., Giege, R., Putz, D. and Florentz, C. (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA*, **6**, 1356-1379.
26. Nazarenko, I.A., Harrington, K.M. and Uhlenbeck, O.C. (1994) Many of the conserved nucleotides of tRNA(Phe) are not essential for ternary complex formation and peptide elongation. *EMBO J*, **13**, 2464-2471.
27. Peterson, E.T., Blank, J., Sprinzl, M. and Uhlenbeck, O.C. (1993) Selection for active E. coli tRNA(Phe) variants from a randomized library using two proteins. *EMBO J*, **12**, 2959-2967.
28. Hou, Y.M. and Schimmel, P. (1992) Novel transfer RNAs that are active in Escherichia coli. *Biochemistry*, **31**, 4157-4160.
29. Kotlova, N., Ishii, T.M., Zagryadskaya, E.I. and Steinberg, S.V. (2007) Active suppressor tRNAs with a double helix between the D- and T-loops. *J Mol Biol*, **373**, 462-475.
30. Lavender, C.A., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2010) Robust and Generic RNA Modeling Using Inferred Constraints: A Structure for the Hepatitis C Virus IRES Pseudoknot Domain. *Biochemistry*, **49**, 4931-4933.

31. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189-199.
32. Malhotra, A., Tan, R.K.Z. and Harvey, S.C. (1994) Modeling Large Rnas and Ribonucleoprotein-Particles Using Molecular Mechanics Techniques. *Biophys J*, **66**, 1777-1795.
33. Seetin, M.G. and Mathews, D.H. (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J Comput Chem*, **32**, 2232-2244.
34. Major, F., Gautheret, D. and Cedergren, R. (1993) Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc Natl Acad Sci USA*, **90**, 9408-9412.
35. Nakamura, S. and Doi, J. (1994) Dynamics of transfer RNAs analyzed by normal mode calculation. *Nucleic Acids Res*, **22**, 514-521.
36. Li, W. and Frank, J. (2007) Transfer RNA in the hybrid P/E state: correlating molecular dynamics simulations with cryo-EM data. *Proc Natl Acad Sci USA*, **104**, 16540-16545.
37. Bahar, I. and Jernigan, R.L. (1998) Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. *J Mol Biol*, **281**, 871-884.
38. Alexander, R.W., Eargle, J. and Luthey-Schulten, Z. (2010) Experimental and computational determination of tRNA dynamics. *FEBS Lett*, **584**, 376-386.
39. Giege, R., Juhling, F., Putz, J., Stadler, P., Sauter, C. and Florentz, C. (2012) Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip Rev RNA*, **3**, 37-61.
40. Cole, P.E., Crothers, D.M. and Yang, S.K. (1972) Conformational-Changes of Transfer Ribonucleic-Acid - Equilibrium Phase-Diagrams. *Biochemistry*, **11**, 4358-&.
41. Sattin, B.D., Zhao, W., Travers, K., Chu, S. and Herschlag, D. (2008) Direct measurement of tertiary contact cooperativity in RNA folding. *J Am Chem Soc*, **130**, 6085-6087.
42. Dill, K.A., Fiebig, K.M. and Chan, H.S. (1993) Cooperativity in Protein-Folding Kinetics. *Proc Natl Acad Sci USA*, **90**, 1942-1946.
43. Gautheret, D., Damberger, S.H. and Gutell, R.R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J Mol Biol*, **248**, 27-43.
44. Kotelawala, L., Grayhack, E.J. and Phizicky, E.M. (2008) Identification of yeast tRNA Um(44) 2 '-O-methyltransferase (Trm44) and demonstration of a Trm44 role in sustaining levels of specific tRNA(Ser) species. *RNA*, **14**, 158-169.
45. Kawai, G., Yamamoto, Y., Kamimura, T., Masegi, T., Sekine, M., Hata, T., Iimori, T., Watanabe, T., Miyazawa, T. and Yokoyama, S. (1992) Conformational rigidity of specific pyrimidine residues in tRNA arises from posttranscriptional modifications that enhance steric interaction between the base and the 2'-hydroxyl group. *Biochemistry*, **31**, 1040-1046.
46. Hecht, M.H., Richardson, J.S., Richardson, D.C. and Ogden, R.C. (1990) De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science*, **249**, 884-891.
47. Richardson, J.S. and Richardson, D.C. (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA*, **99**, 2754-2759.
48. Wilkinson, K.A., Merino, E.J. and Weeks, K.M. (2005) RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J Am Chem Soc*, **127**, 4659-4667.

49. Privalov, P.L. and Filimonov, V.V. (1978) Thermodynamic analysis of transfer RNA unfolding. *J Mol Biol*, **122**, 447-464.
50. Shelton, V.M., Sosnick, T.R. and Pan, T. (2001) Altering the intermediate in the equilibrium folding of unmodified yeast tRNAPhe with monovalent and divalent cations. *Biochemistry*, **40**, 3629-3638.
51. Crothers, D.M., Cole, P.E., Hilbers, C.W. and Shulman, R.G. (1974) The molecular mechanism of thermal unfolding of Escherichia coli formylmethionine transfer RNA. *J Mol Biol*, **87**, 63-88.
52. Sorin, E.J., Nakatani, B.J., Rhee, Y.M., Jayachandran, G., Vishal, V. and Pande, V.S. (2004) Does native state topology determine the RNA folding mechanism? *J Mol Biol*, **337**, 789-797.
53. Puglisi, J.D., Putz, J., Florentz, C. and Giege, R. (1993) Influence of tRNA tertiary structure and stability on aminoacylation by yeast aspartyl-tRNA synthetase. *Nucleic Acids Res*, **21**, 41-49.
54. Stein, A. and Crothers, D.M. (1976) Conformational changes of transfer RNA. The role of magnesium(II). *Biochemistry*, **15**, 160-168.
55. Li, R., Ge, H.W. and Cho, S.S. (2013) Sequence-Dependent Base-Stacking Stabilities Guide tRNA Folding Energy Landscapes. *J Phys Chem B*, **117**, 12943-12952.
56. Wu, M. and Tinoco, I., Jr. (1998) RNA folding causes secondary structure rearrangement. *Proc Natl Acad Sci U S A*, **95**, 11555-11560.
57. Koculi, E., Cho, S.S., Desai, R., Thirumalai, D. and Woodson, S.A. (2012) Folding path of P5abc RNA involves direct coupling of secondary and tertiary structures. *Nucleic Acids Res*, **40**, 8011-8020.
58. Galtier, N. and Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*, **44**, 632-636.
59. Laing, C., Jung, S., Iqbal, A. and Schlick, T. (2009) Tertiary Motifs Revealed in Analyses of Higher-Order RNA Junctions. *J Mol Biol*, **393**, 67-82.
60. Lamiable, A., Barth, D., Denise, A., Quessette, F., Vial, S. and Westhof, E. (2012) Automated prediction of three-way junction topological families in RNA secondary structures. *Comput Biol Chem*, **37**, 1-5.
61. Kim, N., Laing, C., Elmetwaly, S., Jung, S., Curuksu, J. and Schlick, T. (2014) Graph-based sampling for approximating global helical topologies of RNA. *Proc Natl Acad Sci USA*, **111**, 4079-4084.
62. Bindewald, E., Hayes, R., Yingling, Y.G., Kasprzak, W. and Shapiro, B.A. (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res*, **36**, D392-397.
63. Walter, N.G., Burke, J.M. and Millar, D.P. (1999) Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. *Nat Struct Biol*, **6**, 544-549.
64. Zhao, Z.Y., Wilson, T.J., Maxwell, K. and Lilley, D.M. (2000) The folding of the hairpin ribozyme: dependence on the loops and the junction. *RNA*, **6**, 1833-1846.
65. Tan, E., Wilson, T.J., Nahas, M.K., Clegg, R.M., Lilley, D.M. and Ha, T. (2003) A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc Natl Acad Sci USA*, **100**, 9308-9313.

66. Klostermeier, D. and Millar, D.P. (2000) Helical junctions as determinants for RNA folding: origin of tertiary structure stability of the hairpin ribozyme. *Biochemistry*, **39**, 12970-12978.
67. Baird, N.J., Srividya, N., Krasilnikov, A.S., Mondragon, A., Sosnick, T.R. and Pan, T. (2006) Structural basis for altering the stability of homologous RNAs from a mesophilic and a thermophilic bacterium. *RNA*, **12**, 598-606.
68. Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B. and Steitz, T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci USA*, **98**, 4899-4903.
69. Costa, M. and Michel, F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*, **14**, 1276-1285.
70. Woodson, S.A. (2010) Compact intermediates in RNA folding. *Annu Rev Biophys*, **39**, 61-77.
71. Kilburn, D., Roh, J.H., Behrouzi, R., Briber, R.M. and Woodson, S.A. (2013) Crowders perturb the entropy of RNA energy landscapes to favor folding. *J Am Chem Soc*, **135**, 10055-10063.
72. Denesyuk, N.A. and Thirumalai, D. (2011) Crowding promotes the switch from hairpin to pseudoknot conformation in human telomerase RNA. *J Am Chem Soc*, **133**, 11858-11861.
73. Fang, X., Wang, J., O'Carroll, I.P., Mitchell, M., Zuo, X., Wang, Y., Yu, P., Liu, Y., Rausch, J.W., Dyba, M.A. *et al*. (2013) An unusual topological structure of the HIV-1 Rev response element. *Cell*, **155**, 594-605.
74. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Curr Opin Struct Biol*, **21**, 296-305.
75. Cho, S.S., Pincus, D.L. and Thirumalai, D. (2009) Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *Proc Natl Acad Sci USA*, **106**, 17349-17354.

Chapter 5: Evidence for a Central Role of Topological Constraints in a Pathogenic Mitochondrial tRNA Mutation

This chapter has been adapted from the publication:

Mustoe, A. M., Liu, X., Lin, P., Al-Hashimi, H. M., Fierke, C. A., and Brooks, C. L. III. (2014) Pathogenic insertion mutation destabilizes human mitochondrial $tRNA^{Ser(UCN)}$ by disrupting its topological constraints, *in preparation*.

Author contributions:

All authors helped design the research. A.M. Mustoe performed the simulations; A.M. Mustoe, X. Liu and P. Lin performed the melting experiments; and X. Liu and P. Lin made the RNA. A.M. Mustoe wrote the manuscript with contributions from X. Liu.

5.1 Introduction

In **Chapter 4** we demonstrated that topological constraints play a central role in encoding the specificity, stability, and cooperativity of tertiary folding in canonical cytosolic (cc-tRNA). Furthermore, we showed that small changes to the secondary structure of the tRNA four-way junction could significantly disrupt topological constraints and their favorable contributions to folding, which we predicted should destabilize such mutant tRNAs. Correlating with this observation, secondary structures that disrupt topological constraints are evolutionarily disfavored, whereas alternative junction architectures that conserve tRNA's topological constraints are evolutionarily allowed.

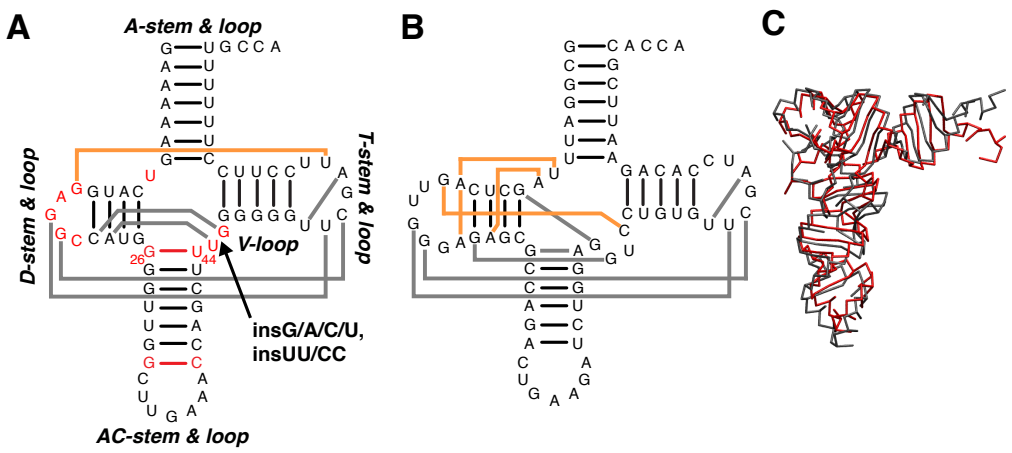

Figure 5.1: Structure of human mt-tRNA$^{Ser(UCN)}$
(A) Secondary structure of human mt-tRNA$^{Ser(UCN)}$ and the location of the pathogenic insG mutation. Secondary structure regions that differ from cc-tRNA are drawn in red. Gray and orange lines indicate tertiary interactions inferred from the tRNA^{Pyl} crystal structure. Tertiary interactions shared by cc-tRNA are shown in gray, and novel interactions are shown in orange. **(B)** Secondary structure of yeast tRNA^{Phe} (cc-tRNA). Conserved tertiary interactions missing in mt-tRNA^{Ser} are shown in orange. **(C)** Superposition of TOPRNA representations of the cc-tRNA (gray) and tRNA^{Pyl} (red) crystal structures.

While **Chapter 4** investigated the most common natural variants of tRNA secondary structure, mitochondrial tRNAs (mt-tRNAs) can exhibit more radical deviations from the canonical cloverleaf structure (1,2). mt-tRNAs are encoded by the mitochondrial DNA (mt-DNA), and along with the mitochondrial ribosome, are responsible for translating 13 mt-DNA-encoded proteins that are essential components of the electron transport chain (3). Compared to cc-tRNAs, mt-tRNAs often lack multiple universally conserved tertiary interactions and have secondary structures with shortened interhelical linkers (1,2). However, in what has been a long-standing puzzle in the RNA folding community, mt-tRNAs are still able fold to near-canonical 3D structures.

Our studies of cc-tRNA lead to the straightforward prediction that the reduced interhelical linker length of mt-tRNAs should increase their topological constraints. In turn, we hypothesize that greater topological constraints may more strongly bias mt-tRNAs towards their native 3D structures, compensating for a decreased number of tertiary interactions. An excellent

system for testing this hypothesis is the relatively well-studied mammalian mt-tRNA$^{Ser(UCN)}$ (hereafter referred to as mt-tRNA^{Ser}) (4-7), which is highly similar to the rare cytosolic pyrolysine tRNA (tRNA^{Pyl}) found in ~20 bacteria and archea species (8,9). These tRNAs differ in several significant respects from cc-tRNAs (**Figure 5.1**). The A/D-loop is only one nucleotide (nt) long, compared to 2-nts in all other tRNA species. The non-canonical base pair atop the AC-stem is also typically replaced with a canonical base pair, and the V-loop is shortened from four or five nucleotides to three (1,9,10). As revealed by the crystal structure of tRNA^{Pyl} (11), these unique secondary structure features preclude these tRNAs from forming several normally critical tertiary pairs and base triples (**Figure 5.1**). Nevertheless, biophysical studies of bovine mt-tRNA$^{Ser(UCN)}$ and archaeal tRNA^{Pyl} transcripts have shown that these molecules stably fold, albeit less stably than cc-tRNAs (5,8).

Intriguingly, a mutation that we predict should disrupt the topological constraints of mt-tRNA^{Ser} is also known to be pathogenic in humans. Formally termed the 7472insC mutation but hereafter referred to as the insG mutation, this mutation lengthens the polyG tract that terminates in the V-loop of human mt-tRNA^{Ser} (**Figure 5.1**), and has been shown to cause deafness, progressive encephalomyopathy, and MERFF (myoclonic epilepsy with ragged red fibers) (7,12-14). Consistent with a pathogenic mechanism involving destabilization of tertiary structure, prior studies showed that the insG mutation leads to impaired transcript processing and a decrease in mt-tRNA^{Ser} mitochondrial concentration by 65% (6,15,16). Enzymatic probing revealed no changes in the mutant's secondary structure (6), and it is thus hypothesized that the mutation destabilizes tertiary structure. One proposal is that the inserted G is accommodated into the T-loop, which in turn disrupts the T-loop's highly conserved local structure and correspondingly its ability to form long-range tertiary pairs with the D-loop (12). However, such

a disruption is likely to destabilize the mutant by more than 65% given that disrupts . The inserted G could be equally well accommodated into the V-loop, which would disrupt topological constraints and thus potentially destabilize 3D structure.

Here, we test these hypotheses using TOPRNA simulations and UV melting experiments. Our simulations indicate that the topological constraints of mt-tRNA^{Ser}'s unique secondary structure greatly reduce the entropic cost of tertiary folding, thereby stabilizing the molecule and compensating for its decreased number of tertiary interactions. Furthermore, experiments and simulations demonstrate in quantitative agreement that the pathogenic insG mutation destabilizes mt-tRNA^{Ser} tertiary structure by disrupting its topological constraints, thus providing the first example of a topological-constraint-based disease mechanism.

5.2 Materials and methods

5.2.1 TOPRNA Simulations

Simulations were performed using TOPRNA2, a modified version of the original TOPRNA force field that corrects a bias in the backbone dihedral applied to base-paired residues that favored undertwisting of helices (see **Appendix D.1**). Initial coordinates of WT, insN, and insNN mt-tRNA^{Ser} were obtained using the `-init` option of the *toprnaCreate.pl* software (brooks.chem.lsa.umich.edu). Langevin dynamics temperature replica exchange simulations were performed in CHARMM (17) using the MMTSB toolset (18) for a total of 10^9 dynamics steps, using a 20 fs timestep, 5 ps^{-1} friction coefficient, and 8 temperature windows spanning 300 to 450 K. The first 2×10^6 simulations steps were discarded as equilibration. Exchanges were attempted every 2000 dynamics steps, with exchange rates varying between 42-45% for all simulations. All other simulation parameters were set as previously described in prior Chapters.

5.2.2 Simulation Analysis

Analysis was restricted to conformations sampled at 300 K. Interhelical Euler angles were measured as described in previous chapters, using the three base pairs below the G26-U44 pair as the AC-stem reference to facilitate comparisons to cc-tRNA and mt-tRNA$^{Ser}_{G26U44}$ simulations. The fraction of interhelical conformations sampled was computed by binning on a 60° grid. Distances in Euler angle space were computed as the amplitude of the single axis rotation needed to convert a given interhelical (α_h, β_h, γ_h) angle to the value measured in the crystal structure, as described in **Section 1.4.3**. The D- and T-loops were considered to be in contact if the S-beads from at least one pair of residues were within 14 Å. Interhelical stacking was evaluated as described in **Section 4.2.5**. RMSDs were computed using P-beads, excluding the 3'-A/GCCA and V-loops.

Additional calculations of ΔG^{topo}_{fold} using orthogonal metrics ensured that our estimates were independent of the specific definition of native structure used (not shown). In place of the modified Tyagi and Mathews criteria (19) for defining stacking, we assessed whether the (α_h, β_h, γ_h) angles between the D- and AC-stems and T- and A-stems were within 30° single axis rotations of the values observed in the 6TNA and 2ZNI crystal structures. To assess whether the D- and T-loops could form contacts, we superimposed each TOPRNA conformation on the crystal structure using the D-stem and then computed the (x,y,z) coordinates of the closing base-pair of the T-stem. If these coordinates were within 15 Å of the coordinates observed in the crystal structure than we considered the TOPRNA conformation to possess D-to-T-loop contacts.

The 500 best-packed conformations of each molecule were identified as described in **Section 4.2.6**, with ε_s and ε_l scaling parameters set to -3.5 and -0.6 kcal/mol. Stacking interactions were evaluated between the A- and D-stems, D- and AC-stems, and T- and A-stems.

The native-specificity of the best-packed ensemble was similar computed as described in **Section 4.2.6.**

5.2.3 Preparation of human mt-tRNA$^{Ser(UCN)}$ constructs

The human mt-tRNA$^{Ser(UCN)}$ gene was subcloned into a pUC18 plasmid to prepare DNA templates for transcription by T7 RNA polymerase in two steps. First, the mt-tRNA$^{Ser(UCN)}$ gene containing the native sequence of the 5' leader, body, and 3' trailer was amplified by PCR using 0.04 unit/μL PfuTurbo® DNA polymerase (Stratagene), 0.2 mM dNTPs (Life Technologies), 4 ng/μL mitochondrial genomic DNA template extracted from HEK293 cells (mitochondrial DNA isolation kit, BioVision), and 0.5 μM primers (forward primer: 5'-TAA TAC GAC TCA CTA TAG ACA AAG TTA TGA AAT GGT TTT TCT AAT AC-3'; reverse primer: 5'-CAC CAT CTA GAC AAA AAA GGA AGG AAT CG-3'). The PCR reaction cycles were: (1) denaturation at 95 ˚C for 2 min, (2) 30 cycles of denaturation at 95 ˚C for 1 min, annealing at 55.5 ˚C for 30 sec, and elongation at 72 ˚C for 1 min, and (3) a final elongation step at 72 ˚C for 10 min. The PCR product was ligated into a pUC18 vector digested by SmaI restriction enzyme (New England Biolabs) and transformed into XL1Blue competent cells for blue-white selection using a LB-agar plate top dressed with Isopropyl β-D-1-thiogalactopyranoside (IPTG, Fisher Scientific) and 5-bromo-4-chloro-indolyl-β-D-galactopyranoside (X-gal, Fisher Scientific). White colonies were picked for extraction of plasmid DNA (QIAprep Spin Miniprep Kit, Qiagen) and the sequence of the yielding plasmid (pre-mt-tRNA36-5) was confirmed by sequencing using a M13-Forward primer (University of Michigan Sequencing Core). Second, the final a plasmid containing the human mt-tRNA$^{Ser(UCN)}$ gene and a T7 promoter sequence was constructed by amplification of the tRNA gene plus a CCA sequence (BstNI restriction site)

from the pre-mt-tRNA36-5 plasmid by PCR as described above using the forward primer 5'-TAA TAC GAC TCA CTA TAG TTT TGA AAA AG- CAT GGA GGC-3' and reverse primer 5'-CAA AAA AGG AAG GAA TCG AAC C-3', followed by blunt-end ligation into a SmaI digested pUC18 vector, blue-white selection, and sequencing by the M13-Forward primer (University of Michigan Sequencing Core).

Insertion mutant plasmids were obtained by site-directed mutagenesis of the WT plasmid using forward primers 5'-GGC TTG AAA CCA GCT TTX GGG GGG TTC GAT TC-3', and reverse primers 5'-GAA TCG AAC CCC CCX AAA GCT GGT TTC AAG CC-3', where X=A, C, G, U or CC, UU. The PCR cycles were: (1) denaturation at 95 ˚C for 2 min, (2) 18 cycles of denaturation at 95 ˚C for 1 min, annealing at 55 ˚C for 1 min and elongation at 68 ˚C for 6 min, and (3) a final elongation step at 68 ˚C for 10 min. The PCR product was treated with 0.4 unit/μL DpnI restriction enzyme (New England Biolabs) prior to transformation into XL1Blue cells for plasmid extractions. Insertion mutant plasmid sequences were confirmed by DNA sequencing using a M13-Forward primer (University of Michigan Sequencing Core).

The DNA templates for *in vitro* run-off transcription of WT and insertion mutants were prepared by digestion of the corresponding plasmids by the BstNI restriction enzyme (New England Biolabs). RNA was transcribed in the presence of 4 mM NTPs (Sigma), 0.1 μg/μL T7 RNA polymerase, 0.8-1 μg/μL linearized DNA template, 1 mM spermidine (Sigma), 5 mM DTT (Sigma), 2 μg/mL pyrophosphatase (Roche), 50 mM Tris–HCl (pH 8.0) and 20 mM $MgCl_2$, incubated at 37 °C for overnight, and purified by 10 or 12% denaturing polyacrylamide/bis gel containing 7 M urea. RNA was eluted from gel slices in TESN buffer (10 mM Tris, pH 8, 1 mM EDTA, 0.1% SDS, and 500 mM NaCl) at 4 °C overnight followed by buffer-exchange into TEN

buffer (10 mM Tris, pH 8, 1 mM EDTA, and 500 mM NaCl) by Amicon (10,000 MWCO, Millipore) at 4 °C and ethanol precipitated.

5.2.4 UV Melting and CD Experiments

Concentrated RNA was diluted to into a lightly buffered water solution containing ~0.01 mM EDTA and 0.1 mM Tris (pH 7.2), denaturated at 95° C for two minutes, refolded at room temperature for 15 minutes, followed by addition of 5× concentrated folding buffer and incubation at room temperature for 30 minutes. Final solution conditions consisted of 0.5 μM RNA, 0.01 mM EDTA, 0.1 mM Tris, 150 mM NaCl, 20 mM Na-cacodylate (pH 7.2), and either 2 mM or 5 mM $MgCl_2$. For melting experiments, the solution was topped with 50 μL of mineral oil to prevent evaporation. Melts were performed at a rate of 0.5°C/min from 15 to 90°C monitoring absorbance at 260 nm on a Cary 100Bio UV/Vis spectrophotometer. Derivatives were obtained by subtracting the buffer baseline, normalizing to the absorbance at 15°C, and computing the slope of the linear least-squares fit to all data points within a ±3° C temperature window (20). Experimental replicates showed good agreement (not shown).

Global Melt Fit (GMF) provided by D.E. Draper (Johns Hopkins) was used to obtain melting enthalpies and temperatures of the WT, insA, insU, insG, and insUU mt-tRNAs utilizing custom python wrapper scripts to interface with the GMF fitting routine (20). Curves were fit assuming four sequential transitions, which yielded a total ΔH of unfolding of 200-260 kcal/mol consistent with expectation based on nearest neighbor stacking rules. ΔH and ΔAbs of all transitions were required to be >0 for the fit to be considered successful. In multiple cases, including WT at 5 mM $MgCl_2$, the low temperature peak could be fit by either a single transition or two closely spaced transitions, with the χ^2 of the fit using two transitions for the low

temperature peak sometimes as much as fivefold better than when fit by a single transition. However, such fits were generally less reproducible. Moreover, the low T_m predicted by these fits was <30° C for all molecules, which contradicts NMR evidence collected on the highly similar bovine mt-tRNA^{Ser} that tertiary structure melts above 37° C at similar solution conditions (0.5 mM tRNA, 10 mM Na-cacodylate (pH 6.5), 150 mM NaCl, 10 mM $MgCl_2$) (5). Given that all curves could be satisfactorily fit using a single transition to model the first peak, and that the most parsimonious model is that all the tRNAs should have similar melting behavior, we thus excluded such "two first-transition" fits from our analysis. Note that with this choice, the parameters of the first transition were invariant if the melting curve was refit assuming only three sequential transitions instead of four (not shown).

5.3 Results

5.3.1 Topological constraints stabilize WT mt-tRNA$^{Ser(UCN)}$

To explore whether topological constraints of the unique secondary structure of WT mt-tRNA^{Ser} help stabilize tertiary structure relative to cc-tRNA, we performed extensive temperature replica exchange simulations of an updated version of the TOPRNA coarse-grained model (**Appendix D.1**, **Figure D.1**). As in prior Chapters, we analyzed the global conformations sampled by our simulations using three sets of Euler angles, (α_h, β_h, γ_h), to describe the simultaneous pairwise orientations between the A-, D-, and T-stems and the AC-stem (**Figure 5.2A**). Counting the number of unique 9D-angles sampled in turn allows us to estimate the degree to which topological constraints restrict the mt-tRNA^{Ser} conformational ensemble. Consistent with our expectations, the shorter A/D- and V-loops place significantly greater constraints on mt-tRNA^{Ser} conformation compared to cc-tRNA, with it sampling only 2.5% compared to 5.7% of possible

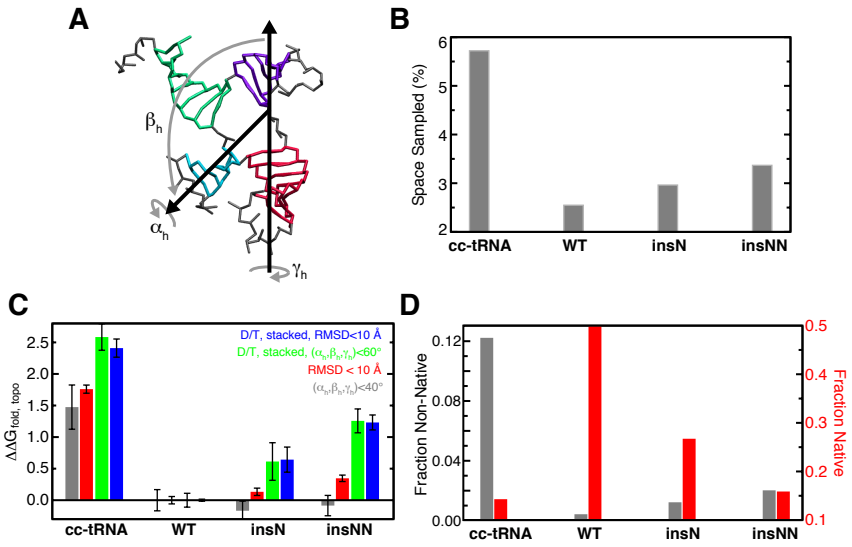

Figure 5.2: Greater topological constraints bias mt-tRNA^{Ser} towards native structure
(A) Representative snapshot from a TOPRNA simulation of mt-tRNA^{Ser} illustrating the Euler angle representation of the interhelical orientation between the AC-stem (red) and D-stem (blue). Similar angles can also be measured between the AC-stem and A-stem (green) and D-stem (purple) (not shown). Loops, which are treated as freely rotatable chains by TOPRNA, are colored gray. **(B)** The fraction of theoretically possible interhelical (α_h, β_h, γ_h) states sampled by TOPRNA simulations. **(C)** Differences in the topological constraint contribution to folding free energy $\Delta\Delta G_{fold,\ topo} = \Delta G_{fold,\ topo}(i) - \Delta G_{fold,\ topo}(WT)$. Values and error bars represent the mean and standard deviation of $\Delta\Delta G$ computed by block averaging over thirds of the simulations. **(D)** Fraction of the 500 best-packed conformations from each tRNA simulation that have non-native contacts (gray) or that have D-to-T-loop contacts and both interhelical stacks (red). The fraction of conformations with both native interactions is weighted by conformational entropy based on how many structurally similar conformations are in the best-packed ensemble.

---

interhelical conformations (**Figure 5.2B**). In **Chapter 4** we showed that one of the primary mechanisms through which topological constraints restrict cc-tRNA conformation is by coupling the orientation of all four tRNA helices together, giving rise to long-range correlated motions. Consistent with this explanation, we find that helices in mt-tRNA^{Ser} are 10-90% more correlated than in cc-tRNA (**Figure D.2**).

We next examined whether the greater topological constraints of WT mt-tRNA^{Ser} bias the molecule towards native-like conformations. In particular, we computed the probability P_{nat} of mt-tRNA^{Ser} and cc-tRNA sampling native-like 3D conformations in our simulations, which is directly related to the topological constraint contribution to the free energy of folding through

$$\Delta G_{fold}^{topo} = -k_B T \ln\left(\frac{P_{nat}}{1 - P_{nat}}\right) \quad \textbf{[5.1]}$$

As TOPRNA lacks attractive interactions, our simulations spend the majority of their time in extended conformations and P_{nat} is exceedingly small; depending on the metric used to define native-like structure, P_{nat} ranged from 10^{-2} to 10^{-5}. ΔG_{fold}^{topo} is hence large and positive for both molecules, reflecting the entropic penalty of folding that must be offset by favorable tertiary interactions. Strikingly, however, we find that the $\Delta\Delta G_{fold}^{topo}$(T = 300 K) between cc-tRNA and mt-tRNA^{Ser} is as large as 2.5 kcal/mol (**Figure 5.2C**, **D.3**), indicating that topological constraints stabilize mt-tRNA^{Ser} by this amount. Notably, this is comparable to the favorable free energy provided by one to two tertiary base pairs (21) and strongly supports the hypothesis that topological constraints help compensate for the decreased number of tertiary interactions in mt-tRNA^{Ser}.

In estimating P_{nat} above for mt-tRNA^{Ser} and cc-tRNA to obtain $\Delta\Delta G_{fold}^{topo}$=2.5 kcal/mol, we required that TOPRNA-sampled conformations satisfy both local and global features of tRNA native structure to count as native. Specifically, native-like conformations were defined as those having <10 Å RMSD from the crystal structure and also possessing contacts between the D- and T-loops and both native interhelical stacks. As an alternative measure of global structure in place of RMSD, we also computed P_{nat} requiring that all interhelical (α_h, β_h, γ_h) angles are within 60° of their crystal structure value. In theory, the $\Delta\Delta G_{fold}^{topo}$ should be independent of the metric used to compute P_{nat}, which we observe for the two different definitions of global structure. Interestingly, however, the $\Delta\Delta G_{fold}^{topo}$ between mt-tRNA^{Ser} and cc-tRNA decreases to ~1.5 kcal/mol if P_{nat} is computed using only a global definition of native structure, ignoring

whether tertiary interactions are formed (**Figure 5.2C**). This difference in $\Delta\Delta G_{fold}^{topo}$ depending on whether formation of tertiary interaction is included in the native definition indicates that topological constraints stabilize mt-tRNA^{Ser} by two mechanisms. First, greater topological constraints reduce the cost of forming native-like interhelical conformations relative to cc-tRNA. Second, once the helices have adopted near-native orientations, the shorter connecting loops of mt-tRNA^{Ser} restrict the translational freedom of the helices, reducing the entropic cost of bringing the D and T-loops into proximity. The shorter loops also reduce the cost of stacking by restricting the ability of mt-tRNA helices to adopt non-stacked conformations (**Figure D.3**).

The above analysis reports on the energetic cost of each tRNA forming the native structure. Another important aspect of the folding energy landscape is whether or not the tRNAs can form competing tertiary folds, or in other words, the specificity of the folding landscape. Indeed, we showed in **Chapter 4** that topological constraints contribute significantly to cc-tRNA folding specificity by preventing formation of non-native tertiary folds. We thus assessed whether the greater constraints of mt-tRNA^{Ser} also increase its folding specificity. As before, we identified the 500 conformers sampled by TOPRNA that have the best inter-loop packing (and thus potential to form putative tertiary interactions). Notably, only 0.4% of mt-tRNA^{Ser} compared to ~12% of cc-tRNA best-packed conformers possess non-native tertiary contacts (**Figure 5.2D**). Additionally, the degree to which individual conformers are close in structure to other best-packed conformers provides a relative measure of their conformational entropy. We find that near-native 3D conformations of mt-tRNA^{Ser} have significantly higher entropies than cc-tRNA (**Figure D.4**), which we summarize using the entropy-weighted fraction of best-packed conformers that possess both D-to-T-loop contacts and interhelical stacking interactions (**Figure**

5.2D). Thus, mt-tRNA^{Ser} topological constraints both lower the cost of forming native interactions (ΔG^{topo}_{fold}) and increase the energy gap between native and non-native folds.

The mt-tRNA^{Ser} simulation analyzed above treated the G26-U44 base pair as a permanent feature of secondary structure (**Figure 5.1A**). Although WC or GU pairing is preferred at this position in the majority of mt-tRNA^{Ser} and tRNA^{Pyl} species, A^{+}C, AA, or GA pairing also occurs frequently (1,9,10). To determine the contribution of pairing between residues 26 and 44 to our above results we performed an additional TOPRNA simulation of mt-tRNA^{Ser} where these residues were treated as single strands. Compared to when G26-U44 is paired, the mt-tRNA$^{Ser}_{noBP}$ simulation samples ~50% more interhelical conformations (**Figure D.5**). The ΔG^{topo}_{fold} of mt-tRNA$^{Ser}_{noBP}$ is also greater by ~1 kcal/mol, and folding specificity is reduced ~5% (**Figure D.5**). However, relative to cc-tRNA, mt-tRNA$^{Ser}_{noBP}$ is still stabilized by ~1.5 kcal/mol. These results support that the entire junction architecture is responsible for the topological constraint stabilization of mt-tRNA^{Ser}. The stabilization afforded by G26-U44 pairing is also consistent with the evolutionary preference for WC/GU pairs at this position.

5.3.2 The insG mutation destabilizes mt-tRNA^{Ser} by disrupting its topological constraints

We next explored whether the insG mutation disrupts the ability of topological constraints to stabilize mt-tRNA^{Ser}, thereby explaining its pathogenicity. Consistent with our hypothesis, a TOPRNA simulation of mt-tRNA^{Ser} with an insN mutation in the V-loop samples ~16% more interhelical conformations than the WT simulation (**Figure 5.2B**). (Note that we use the insN notation to emphasize that our simulations are independent of inserted nucleotide identity). Likewise, the interhelical correlations of the insN mutant are ~1-20% decreased compared to WT (**Figure D.2**).

Significantly, this reduction in topological constraints is accompanied by a ~0.6 kcal/mol increase in the ΔG_{fold}^{topo} of the insN mutant compared to WT, and a 50% decrease in folding specificity (**Figure 5.2**, **D.3**, **D.4**). This is a consequential destabilization of mt-tRNA^{Ser} given its expected instability (5). Indeed, destabilizing mt-tRNA^{Ser} by 0.6 kcal/mol should decrease its folded population by ~65%, which remarkably matches the 65% decrease in insG mutant concentration observed in cells (15). An analogous insA mutation was also shown to reduce the in-cell activity of the related tRNA^{Pyl} by 83% (22), consistent with our expectation that the insN mutation should destabilize such tRNAs in a sequence independent manner. Interestingly, this destabilization is due to an increased energetic cost of forming tertiary contacts, as near-native global 3D conformations are formed with similar energies (**Figure 5.2C**). Consistent with the insN mutant's decreased interhelical correlations (**Figure D.2**), ~0.2 kcal/mol of this increase is due to a loss of cooperativity among the interhelical stacks and D-to-T-loop contacts (**Figure D.3**). Thus, in part due to decoupling of the helices, greater conformational freedom near the native-state increases the entropic cost of forming tertiary interactions in the mutant. Simulations of the insN mutant without a G26-U44 base pair also demonstrated similar decreases in topological confinement, $\Delta\Delta G_{fold}^{topo}$ between mutant and WT, and folding specificity (**Figure D.4**).

To further establish the trend between topological constraints and folding, we also performed simulations of an insNN V-loop mutant of mt-tRNA^{Ser} (**Figure 5.1A**). As expected, these simulations revealed that the insNN mutant has further decreased topological constraints, energetic cost of folding, and folding specificity (**Figure 5.2**, **D.2**, **D.3**, **D.4**). Similar results were also observed for an insNN simulation performed without a G26-U44 base pair (**Figure D.5**).

5.3.3 Melting experiments reveal destabilization of insertion mutants in quantitative agreement with TOPRNA predictions

Our simulations make several strong predictions regarding mt-tRNA^{Ser} stability: (i) Single and double insertions in the V-loop should destabilize tertiary structure by ~0.6 kcal/mol and ~1.2 kcal/mol at 300 K, respectively. (ii) This destabilization should be independent of the identify of the inserted nucleotide(s), assuming that like the insG mutation (6), the insertion does not cause secondary structure misfolding. (iii) This destabilization should be directly related to an increase in the entropic cost of folding for the mutants. Notably, from this last prediction we can derive the expected decrease in melting temperature (T_m) of the mutants. As detailed in **Appendix D.2**, if the folding enthalpies of the mutant and WT are identical than the T_m of the two tRNAs can be related through

$$T_m^{WT}\Delta S_{fold,3Dconf}^{WT} \approx T_m^{mut}\Delta S_{fold,3Dconf}^{mut}. \quad \textbf{[5.2]}$$

$\Delta S_{fold,\ 3Dconf}$, the change in 3D conformational entropy upon folding, can be estimated directly from our simulations. For $T_m^{WT}\approx$45° C, these calculations predict that the T_m of the insN and insNN mutants should be decreased by 8° C and 15° C, respectively (**Figure D.6**).

To test these predictions, we used *in vitro* transcription to prepare WT, insG, insC, insU, insA, insCC, and insUU human mt-tRNA$^{Ser(UCN)}$ constructs and characterized their folding stabilities by UV melting experiments. Experiments were done at both near-physiological (2 mM $MgCl_2$; 150 mM NaCl) and 'stabilizing' (5 mM $MgCl_2$; 150 mM NaCl) solution conditions. At both conditions, the WT transcript exhibits a pronounced cooperative melting transition at 40 or 45°C followed by multiple subsequent broad transitions at higher temperatures (**Figure 5.3A, B**).

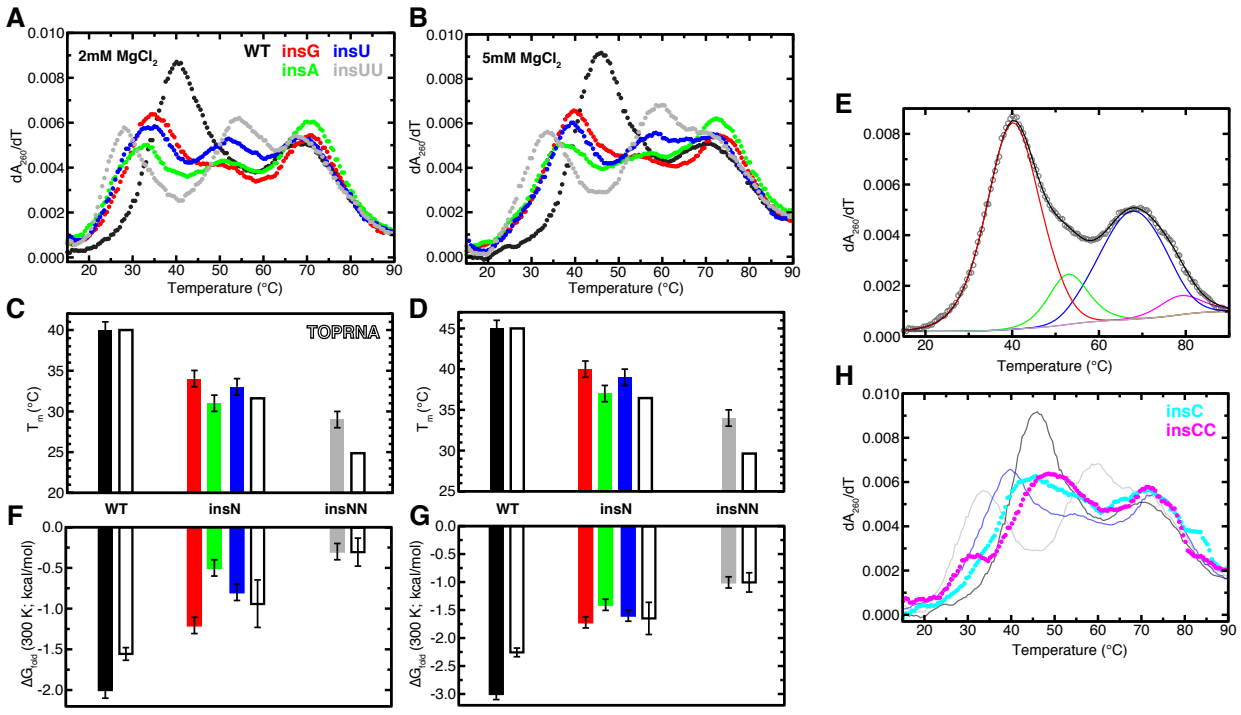

Figure 5.3: UV melting experiments verify that insertion mutants destabilize mt-tRNA^{Ser}
(A,B) Derivative of absorbance at 260 nm of mt-tRNA^{Ser} species in the presence of 2 mM $MgCl_2$ **(A, C, F)** or 5 mM $MgCl_2$ **(B, D, G)**. All melts were performed in a background of 20 mM sodium cacodylate (pH 7.2) and 150 mM NaCl. Curves of different molecules are colored according to the key in (A). **(C, D)** Tertiary structure melting temperatures determined from van't Hoff fits to melting curves. Error bars represent estimated 1° C error. Bars are colored by mutant according to (A). TOPRNA predicted values, in reference to the WT T_m, are shown with open bars and represent the mean of the different ΔT_m estimates in Figure S6. **(E)** Example van't Hoff fit to the WT mutant at 2mM $MgCl_2$. The different transitions are colored, the baseline in gray, and the overall fit in black. **(F, G)** ΔG_{fold} determined from van't Hoff fits extrapolated to 300 K. Color scheme is the same as in (C, D). TOPRNA predictions are referenced to the insUU ΔG_{fold}. **(H)** Melting curves of insC and insCC mutants at 5mM $MgCl_2$. WT, insG, and insUU melting curves are shown by lines colored according to the key in (A) for reference.

The increase in melting temperature with $MgCl_2$ concentration is consistent with the expected stabilizing effects of $MgCl_2$, and both the shape and T_m of these transitions are in good agreement with prior UV melting studies of the highly similar bovine mt-tRNA$^{Ser(UCN)}$ transcripts (5). NMR melting studies of the bovine transcript further demonstrated that the first transition corresponds to tertiary structure melting coupled with some loss of D-stem and T-stem structure (5). We thus also assign the low T_m of 40 and 45°C to tertiary structure melting.

Significantly, the insG, insA, and insU mutants exhibit similar melting curves, yet with the tertiary structure T_m decreased 6-9°C compared to WT, and decreased 11°C for the insUU

mutant (**Figure 5.3**). These values are in remarkable agreement with the 8°C and 15° C decreases estimated from our simulations (**Figure 5.3C, D**); given the extremely coarse-grained nature of TOPRNA and the many approximations used to derive the ΔT_m estimates, we consider differences of several degrees to be well within the expected error of our predictions. While the insG, insA, and insU melting curves differ slightly, their gross similarities strongly support that topological constraints are the primary driver of the mutants' instability. Interestingly, the insertion mutants also melt less cooperatively than the WT transcript (**Figure 5.3A, B**). While this is consistent with the observation above that the mutants should fold less cooperatively, we caution that the derivative of absorbance reflects on the total loss of base stacking versus tertiary structure melting *per se*. Equally plausible is that dA/dT is lower in the mutants because more residual base stacking is maintained at the lower T_m.

We further used a van't Hoff analysis to extract thermodynamic parameters of the tertiary structure melting transition for the WT, insA, insU, insG, and inUU species (20). The melting curves were reasonably fit by a four-state sequential unfolding model (see **Section 5.2.3**), with the ΔH of the first transition ranging between 40-54 kcal/mol for all tRNAs, consistent with the expectation that this transition represents melting of tertiary structure and some D-stem and/or T-stem base pairs (**Figure 5.3E**, **Table D.1, D.2**) (5). Strikingly, when used to derive the ΔG_{fold} of tertiary structure at 300 K, this analysis reveals that the insA/G/U mutants are destabilized 0.8-1.6 kcal/mol relative to WT, and the insUU mutant 1.7-2.0 kcal/mol (**Figure 5.3F, G**). Again, these values are remarkably close to our predictions that the $\Delta\Delta G_{fold}$ should be 0.6 kcal/mol and 1.2 kcal/mol, respectively. The $\Delta\Delta G_{fold} \approx 0.6$ kcal/mol between the insN and insUU mutants exactly matches our prediction. The larger than predicted $\Delta\Delta G_{fold}$ between the WT and insN mutants may be due to distortions of tertiary structure hydrogen bonding or Mg^{2+} interactions

caused by the inserted nucleotides. Similarly, the differences in ΔG_{fold} between the insA/G/U mutants can be easily explained by differences in the ability of the inserted nucleotide to be sterically accommodated and/or form hydrogen bonds in the folded state.

By contrast, the insC and insCC mutants lack a defined tertiary structure melting transition (**Figure 5.3H**). Both mutants appear to exhibit small shoulders near the predicted T_m's of insN and insNN mutants, but these shoulders were poorly fit by a van't Hoff analysis (not shown). We propose that this is due to secondary structure misfolding, perhaps involving mispairing between the inserted C and the complementary Gs atop the AC-stem (**Figure 5.1A**).

5.4 Discussion

How mt-tRNA$^{Ser(UCN)}$ and tRNA^{Pyl} can stably fold to near-canonical tRNA 3D structures despite their non-canonical architectures has long been poorly understood. Our results reveal that these tRNAs employ a unique mechanism where the entropic cost of folding is reduced by their more conformationally constrained secondary structure, thereby compensating for a fewer number of tertiary interactions. In particular, our simulations predict that increased topological constraints stabilize WT mt-tRNA^{Ser} 3D structure by as much as -2.5 kcal/mol and also substantially increase folding specificity compared to cc-tRNA. In the insA, insU, and insG mutants, lengthening the V-loop by a single nucleotide destabilizes mt-tRNA^{Ser} by 0.8-1.6 kcal/mol, leading to a 6-9° C decrease in melting temperature. A two nt insertion destabilizes the insUU mutant by 1.7-2.0 kcal/mol and decreases T_m by 11° C. The agreement between these experimentally measured values and those predicted by our simulations, as well as the sequence independence of the mutant destabilization, provides strong evidence that topological constraints are indeed responsible for these changes.

Importantly, our results provide new insight into why the insG mutation is pathogenic (**Figure 5.4**). Prior studies have found that mt-tRNA^{Ser} concentration is decreased 65% in mitochondria when all mt-DNA plasmids contain the insG mutation, which presumably disrupts mitochondrial protein synthesis and hence cellular respiration (15). This decrease is primarily due to fast degradation of unprocessed or misprocessed transcripts (6,16). However, why the mutant is specifically degraded was unclear. Our data show that unmodified WT mt-tRNA^{Ser} transcripts are only marginally stable, such that the destabilized insG mutant is primarily unfolded at physiological temperatures (T_m<37° C at near-physiological $MgCl_2$ concentrations). Remarkably, the 0.8 kcal/mol destabilization we observe should decrease the folded population of mt-tRNA^{Ser} by 65%, exactly the decrease observed in cells (15). Recent studies have shown that unstable cytosolic tRNAs are rapidly degraded by a rapid tRNA decay (RTD) surveillance pathway (23). We speculate that a similar mitochondrial pathway may be responsible for insG mt-tRNA^{Ser} degradation. Although mt-tRNA^{Ser} lacks the consecutive 5'-Gs shown to be important for cytosolic RTD, it possesses the requisite unstable acceptor stem (23). In support of an RTD-like mechanism, destabilizing mutations to the mt-tRNA^{Ser} acceptor stem lead to similar reductions in mt-tRNA^{Ser} mitochondrial concentration and pathogenic phenotypes (24,25). Finally, we note that the evolutionary appearance of insG and not insA/U/C mutations in the human population does not necessarily indicate that the former is more tolerable; insG mutations are naturally expected at a much higher frequency due to the adjacent homopolymeric G-tract (12,26).

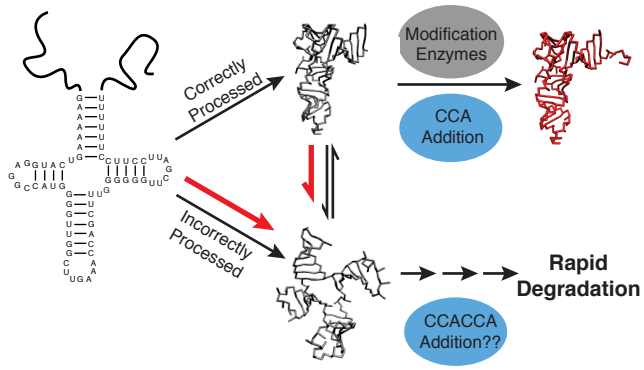

Figure 5.4: Proposed pathogenic mechanism of the 7472insC mutation
Maturation of the polycistronic mt-tRNA^{Ser} transcript includes processing steps such as 5'-leader and 3'-trailer cleavage, CCA addition, and post-transcriptional nucleotide modifications that stabilize tertiary structure (red). The unstable transcript exists in equilibrium between folded and unfolded tertiary conformations, with only the folded conformation likely efficiently processed. In competition with maturation is a rapid decay pathway that degrades unfolded and misprocessed tRNAs. Given the weak A-stem of mt-tRNA^{Ser}, we suggest that degradation may be promoted by 3'-CCACCA addition by the CCA-enzyme, as shown for cytosolic tRNAs (23). The insG mutation shifts the transcript equilibrium towards unfolded conformations (red arrows), leading to increased misprocessing and degradation (16). However, a fraction of mutants are fully processed, and once modified have similar stabilities as WT mt-tRNA^{Ser} (6).

We emphasize that the true measure of our simulations lies in their ability to predict overall destabilization trends rather than their exact quantitative agreement with experiment. Clearly, forces ignored by our simulations such as electrostatics and sequence specific attractive interactions play crucial roles in RNA tertiary folding (27,28), which is born out by the small sequence and $MgCl_2$ stability variations we observe experimentally. It is also worth noting that the inherent instability of mt-tRNA^{Ser} – T_m for most cc-tRNAs is >60° C compared to 45° C measured here (5) – makes it clear that greater topological constraints are insufficient to fully offset the significant loss of tertiary interactions in mt-tRNA^{Ser}. A full understanding of RNA tertiary folding will require accounting for all aspects of the RNA free energy landscape, including secondary structure misfolding which we propose occurs in the insC and insCC mutants. Given these facts, we note that our near quantitative prediction of the mutant ΔT_m is likely partially due to offsetting errors in our analysis, as we assumed that the TOPRNA

populations of different mt-tRNA conformations directly match the populations observed in the real system. Nevertheless, we underscore that our $\Delta\Delta G_{fold,\ topo}$ estimates are assumption free, and due to the additivity of free energy, reflect a real component of the RNA free energy landscape (29,30).

Our results add to the growing body of literature showing that secondary structure, and topological constraints in particular, play a central role in RNA 3D folding (29-37). Significantly, our study is the first to experimentally measure the contribution of topological constraint to the folding free energy of a complex RNA molecule. Indeed, to our knowledge, our study is the first to quantitatively predict *a priori* the destabilizing effect that a secondary structure mutation will have on RNA tertiary folding. Extensive work on the hairpin ribozyme has demonstrated that its evolutionarily selected secondary structure reduces the entropic cost of folding (38-41), although the origins of this entropic stabilization were unclear. Combined with our results, we suggest that entropic stabilization by topological constraints may be a general design principle of RNA folding. Such a mechanism may be the critical bridge towards understanding tertiary structure stability, specificity, and ultimately rationally designing novel RNA folds.

5.5 Conclusions

In this Chapter, we have demonstrated that increased topological constraints encoded by the unique secondary structure of mt-tRNA$^{Ser(UCN)}$ decrease the entropic cost of folding by ~2.5 kcal/mol compared to canonical tRNA, thereby compensating for mt-tRNA$^{Ser(UCN)}$'s reduced number of tertiary interactions. Additional simulations show that the pathogenic 7472insC mutant of mt-tRNA$^{Ser(UCN)}$, which has a one nucleotide longer V-loop, experiences reduced

topological constraints and as a consequence is destabilized by ~0.6 kcal/mol relative to wild-type. Strikingly, this decrease in stability is sufficient to fully account for the decrease in mutant mitochondrial concentration observed in cell-based assays, which is presumed to be the basis of pathogenesis. UV melting experiments further confirmed that insertion mutations lower mt-tRNA$^{Ser(UCN)}$ melting temperature by 6-9°C and increase the folding free energy by 0.8-1.5 kcal/mol in a largely sequence- and salt-independent manner, in quantitative agreement with our simulation predictions. Together, these results confirm that topological constraints make key contributions to the free energy of RNA tertiary folding and provide the first evidence of a pathogenic mutation that is due to disruption of topological constraints.

5.6 References

1. Helm, M., Brule, H., Friede, D., Giege, R., Putz, D. and Florentz, C. (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA*, **6**, 1356-1379.
2. Giege, R., Juhling, F., Putz, J., Stadler, P., Sauter, C. and Florentz, C. (2012) Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip Rev RNA*, **3**, 37-61.
3. Yarham, J.W., Elson, J.L., Blakely, E.L., McFarland, R. and Taylor, R.W. (2010) Mitochondrial tRNA mutations and disease. *Wiley Interdiscip Rev RNA*, **1**, 304-324.
4. Watanabe, Y., Kawai, G., Yokogawa, T., Hayashi, N., Kumazawa, Y., Ueda, T., Nishikawa, K., Hirao, I., Miura, K. and Watanabe, K. (1994) Higher-order structure of bovine mitochondrial tRNA(SerUGA): chemical modification and computer modeling. *Nucleic Acids Res*, **22**, 5378-5384.
5. Hayashi, I., Kawai, G. and Watanabe, K. (1998) Higher-order structure and thermal instability of bovine mitochondrial tRNA^{SerUGA} investigated by proton NMR spectroscopy. *J Mol Biol*, **284**, 57-69.
6. Toompuu, M., Yasukawa, T., Suzuki, T., Hakkinen, T., Spelbrink, J.N., Watanabe, K. and Jacobs, H.T. (2002) The 7472insC mitochondrial DNA mutation impairs the synthesis and extent of aminoacylation of tRNA$^{Ser(UCN)}$ but not its structure or rate of turnover. *J Biol Chem*, **277**, 22240-22250.
7. Yokogawa, T., Watanabe, Y., Kumazawa, Y., Ueda, T., Hirao, I., Miura, K. and Watanabe, K. (1991) A novel cloverleaf structure found in mammalian mitochondrial tRNA(Ser) (UCN). *Nucleic Acids Res*, **19**, 6101-6105.
8. Theobald-Dietrich, A., Frugier, M., Giege, R. and Rudinger-Thirion, J. (2004) Atypical archaeal tRNA pyrrolysine transcript behaves towards EF-Tu as a typical elongator tRNA. *Nucleic Acids Res*, **32**, 1091-1096.

9. Gaston, M.A., Jiang, R. and Krzycki, J.A. (2011) Functional context, biosynthesis, and genetic encoding of pyrrolysine. *Curr Opin Microbiol*, **14**, 342-349.
10. Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res*, **37**, D159-162.
11. Nozawa, K., O'Donoghue, P., Gundllapalli, S., Araiso, Y., Ishitani, R., Umehara, T., Soll, D. and Nureki, O. (2009) Pyrrolysyl-tRNA synthetase-tRNA(Pyl) structure reveals the molecular basis of orthogonality. *Nature*, **457**, 1163-1167.
12. Tiranti, V., Chariot, P., Carella, F., Toscano, A., Soliveri, P., Girlanda, P., Carrara, F., Fratta, G.M., Reid, F.M., Mariotti, C. *et al.* (1995) Maternally inherited hearing loss, ataxia and myoclonus associated with a novel point mutation in mitochondrial tRNASer(UCN) gene. *Hum Mol Genet*, **4**, 1421-1427.
13. Verhoeven, K., Ensink, R.J., Tiranti, V., Huygen, P.L., Johnson, D.F., Schatteman, I., Van Laer, L., Verstreken, M., Van de Heyning, P., Fischel-Ghodsian, N. *et al.* (1999) Hearing impairment and neurological dysfunction associated with a mutation in the mitochondrial tRNASer(UCN) gene. *Eur J Hum Genet*, **7**, 45-51.
14. Jaksch, M., Klopstock, T., Kurlemann, G., Dorner, M., Hofmann, S., Kleinle, S., Hegemann, S., Weissert, M., Muller-Hocker, J., Pongratz, D. *et al.* (1998) Progressive myoclonus epilepsy and mitochondrial myopathy associated with mutations in the tRNA(Ser(UCN)) gene. *Ann Neurol*, **44**, 635-640.
15. Toompuu, M., Tiranti, V., Zeviani, M. and Jacobs, H.T. (1999) Molecular phenotype of the np 7472 deafness-associated mitochondrial mutation in osteosarcoma cell cybrids. *Hum Mol Genet*, **8**, 2275-2283.
16. Toompuu, M., Levinger, L.L., Nadal, A., Gomez, J. and Jacobs, H.T. (2004) The 7472insC mtDNA mutation impairs 5' and 3' processing of tRNA(Ser(UCN)). *Biochem Biophys Res Commun*, **322**, 803-813.
17. Brooks, B.R., III, C.L.B., Jr, A.D.M., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. *et al.* (2009) CHARMM: The biomolecular simulation program. *J Comput Chem*, **30**, 1545-1614.
18. Feig, M., Karanicolas, J. and Brooks, C.L.I. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model*, **22**, 377-395.
19. Tyagi, R. and Mathews, D.H. (2007) Predicting helical coaxial stacking in RNA multibranch loops. *RNA*, **13**, 939-951.
20. Draper, D.E., Bukhman, Y.V. and Gluick, T.C. (2000) Thermal methods for the analysis of RNA folding pathways. *Curr Protoc Nucleic Acid Chem*, 11.13. 11-11.13. 13.
21. Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) RNA structure prediction. *Annu Rev Biophys Biophys Chem*, **17**, 167-192.
22. Ambrogelly, A., Gundllapalli, S., Herring, S., Polycarpo, C., Frauer, C. and Soll, D. (2007) Pyrrolysine is not hardwired for cotranslational insertion at UAG codons. *Proc Natl Acad Sci USA*, **104**, 3141-3146.
23. Wilusz, J.E., Whipple, J.M., Phizicky, E.M. and Sharp, P.A. (2011) tRNAs marked with CCACCA are targeted for degradation. *Science*, **334**, 817-821.
24. Mollers, M., Maniura-Weber, K., Kiseljakovic, E., Bust, M., Hayrapetyan, A., Jaksch, M., Helm, M., Wiesner, R.J. and von Kleist-Retzow, J.C. (2005) A new mechanism for mtDNA pathogenesis: impairment of post-transcriptional maturation leads to severe

depletion of mitochondrial $tRNA^{Ser(UCN)}$ caused by T7512C and G7497A point mutations. *Nucleic Acids Res*, **33**, 5647-5658.
25. Li, X., Fischel-Ghodsian, N., Schwartz, F., Yan, Q., Friedman, R.A. and Guan, M.X. (2004) Biochemical characterization of the mitochondrial tRNASer(UCN) T7511C mutation associated with nonsyndromic deafness. *Nucleic Acids Res*, **32**, 867-877.
26. Garcia-Diaz, M. and Kunkel, T.A. (2006) Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem Sci*, **31**, 206-214.
27. Butcher, S.E. and Pyle, A.M. (2011) The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc Chem Res*, **44**, 1302-1311.
28. Lipfert, J., Doniach, S., Das, R. and Herschlag, D. (2014) Understanding nucleic Acid-ion interactions. *Annu Rev Biochem*, **83**, 813-841.
29. Mustoe, A.M., Al-Hashimi, H.M. and Brooks, C.L., III. (2014) Coarse grained models reveal essential contributions of topological constraints to the conformational free energy of RNA bulges. *J Phys Chem B*, **118**, 2615-2627.
30. Mustoe, A.M., Brooks, C.L.I. and Al-Hashimi, H.M. (2014) Topological Constraints are Major Determinants of tRNA Tertiary Structure and Dynamics and Provide Basis for Tertiary Folding Cooperativity. *In Submission*.
31. Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S. and Herschlag, D. (2009) Do conformational biases of simple helical junctions influence RNA folding stability and specificity? *RNA*, **15**, 2195-2205.
32. Bailor, M.H., Sun, X.Y. and Al-Hashimi, H.M. (2010) Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation. *Science*, **327**, 202-206.
33. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) 3D maps of RNA interhelical junctions. *Nat Protoc*, **6**, 1536-1545.
34. Laing, C. and Schlick, T. (2009) Analysis of four-way junctions in RNA structures. *J Mol Biol*, **390**, 547-559.
35. Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83-93.
36. Sim, A.Y.L. and Levitt, M. (2011) Clustering to identify RNA conformations constrained by secondary structure. *Proc Natl Acad Sci USA*, **108**, 3590-3595.
37. Hajdin, C.E., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2010) On the significance of an RNA tertiary structure prediction. *RNA*, **16**, 1340-1349.
38. Walter, N.G., Burke, J.M. and Millar, D.P. (1999) Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. *Nat Struct Biol*, **6**, 544-549.
39. Zhao, Z.Y., Wilson, T.J., Maxwell, K. and Lilley, D.M. (2000) The folding of the hairpin ribozyme: dependence on the loops and the junction. *RNA*, **6**, 1833-1846.
40. Tan, E., Wilson, T.J., Nahas, M.K., Clegg, R.M., Lilley, D.M. and Ha, T. (2003) A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc Natl Acad Sci USA*, **100**, 9308-9313.
41. Klostermeier, D. and Millar, D.P. (2000) Helical junctions as determinants for RNA folding: origin of tertiary structure stability of the hairpin ribozyme. *Biochemistry*, **39**, 12970-12978.

Chapter 6: Topological Constraints Provide a Basis for Tertiary Specificity, Cooperativity, and Hierarchical Folding in a Large RNA

6.1 Introduction

In **Chapters 4-5**, we showed that topological constraints play a key role in stabilizing and specifying tRNA tertiary structure. Topological constraints likely play similarly important roles in the folding of larger, more complex RNAs. In particular, as discussed in **Chapters 1** and **4**, many large RNAs reuse multiple of the same tetraloop/tetraloop-receptor (TL/TLR) tertiary interaction motif to stabilize their folds (**Figure 6.1**) (1). The other primary stabilizing tertiary interaction motifs found in large RNAs, so-called A-minor motifs, involves formation of base-to-base and base-to-backbone hydrogen bonds between single-stranded A's and Watson-Crick (WC) base-pairs with limited sequence specificity (2-4). Thus, neither the sequences of TL/TLR nor A-minor motifs provide a significant source of tertiary folding specificity. Furthermore, many large RNAs possess strong thermodynamic cooperativity between distinct tertiary interactions (5,6), and are stabilized by the addition of molecular crowders that non-specifically favor compact conformations (7,8). Together, these observations are consistent with topological constraints providing an underlying source of folding specificity and cooperativity.

The principle system that has been used to study large RNA folding is the *Tetrahymena* Group I intron ribozyme, which was both the first catalytic RNA discovered and

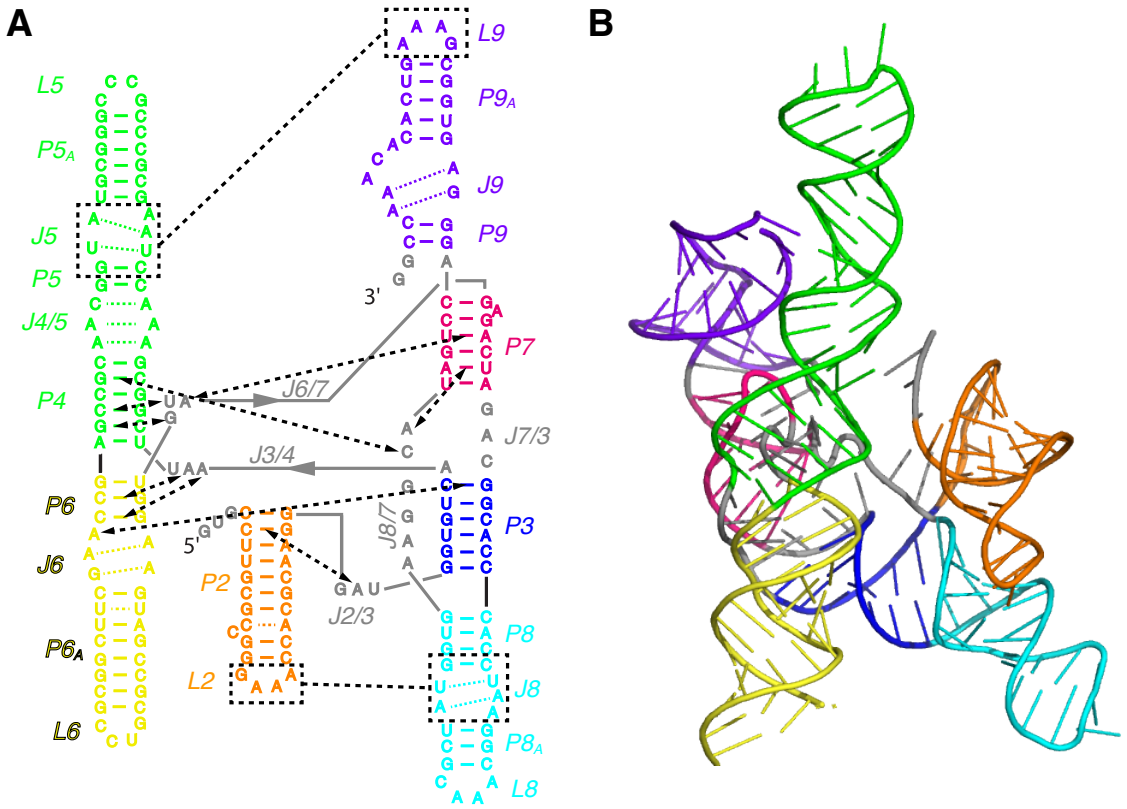

Figure 6.1: Secondary structure and 3D structure of the *Azoarcus* ribozyme
(**A**) Different secondary structure elements are colored and labeled. Important tertiary interactions are drawn as black dashed lines. (**B**) Crystal stucture of the *Azoarcus* ribozyme colored as in (**A**) (PDB 1U6B).

---

the catalyst for the subsequent revolution in understanding RNA's role in molecular biology (9). Other Group I intron species, in particular from *Twort* and *Azoarcus*, have also been extensively characterized (10). Group I introns as a class perform two consecutive transesterification reactions that require binding and alignment of the 5' exon in the intron active site, cleavage of the 5' exon by an exogenous gaunosine, conformational change that places the 3' exon in the active site, and subsequent attack by the 5' exon to ligate the 5' and 3' exon ends (11). These steps require Group I introns to adopt highly specific, complex 3D structures (**Figure 6.1**).

All group I introns possess a highly conserved core consisting of P3 through P9 paired (P) regions that self-organize into two separate domains: P4-P5-P6 and P3-P7-P9 (12,13). The *Azoarcus* ribozyme shown in **Figure 6.1** contains only these core elements and thus represents a

minimal group I ribozyme fold; in *Tetrahymena* and other species, additional 'peripheral' domains can extend from the core helices and buttress the core tertiary structure (14). The two core domains are topologically linked together by the central P3 pseudoknot, and the tertiary structure is stabilized by a network of base-triple, A-minor, and TL/TLR tertiary interactions **(Figure 6.1)** (12,14). The active site is located within the cleft between the two domains, centered at the top of P7, and the surrounding helices provide the necessary global scaffold to specifically recognize and align the exon substrates (12,14).

Extensive studies of the group I intron tertiary folding have revealed both common themes as well as species-specific idiosyncrasies (10,15). Following the general hierarchical model of RNA folding, secondary structure folds first, followed by counterion-mediated collapse to a native-like compact intermediate, and finally, slower reorganization of the catalytic core. For most species, a large fraction of molecules becomes trapped in stable kinetic intermediates, which is thought to be due to local secondary structure misfolding in and around the P3 pseudoknot (10,16). However, a reasonable proportion of the folding flux proceeds directly to the native state (10,15). From a topological constraint perspective, this suggests that native secondary structure, when correctly folded, helps prevent the formation of tertiary misfolded conformations. Further suggesting an important role for topological constraints in group I intron folding, studies of the *Azoarcus* ribozyme observed widespread thermodynamic cooperativity between distinct tertiary interactions (5). Interestingly, this cooperativity appears upon folding to native-like compact intermediates, before tertiary structure is fully formed. Topological constraints provide a logical basis for this observed cooperativity, and indeed Woodson and colleagues speculated as much (5). Finally, landmark studies by Michel and Westhof (12) observed that the secondary structure of the group I intron core posed strong constraints on

global 3D conformation, and in turn were able to use these constraints coupled with inferred tertiary interactions to accurately predict the intron 3D structure.

In this Chapter, we use TOPRNA simulations to explore the role of topological constraints in large RNA folding using the *Azoarcus* ribozyme as a model system. Consistent with our observations on smaller RNAs, we find that topological constraints strongly limit the ribozyme's 3D conformation, help encode tertiary structure specificity, and provide the basis of widespread thermodynamic cooperativity between tertiary interactions. We furthermore present evidence that topological constraints give rise to hierarchical folding behavior, which we speculate may be important for preventing formation of tertiary misfolded intermediates.

6.2 Methods

6.2.1 Azoarcus ribozyme simulation details

All TOPRNA simulations were performed using the updated TOPRNA2 force field (**Appendix D.1**). Temperature replica exchange simulations were performed using 8 temperature windows from 300 to 400 K using the REPD module of CHARMM (17), with exchanges attempted every 5000 dynamics steps. Exchange rates varied between 0.34 and 0.36. Unrestrained simulations were performed for a total of 5×10^9 dynamics steps and restrained simulations for 10^9 steps. All other simulation parameters were set as described in previous chapters.

Initial coordinates were obtained from chain B of PDB 1U6B (18) using the `-fromc` option of *toprnaCreate.pl* (brooks.chem.lsa.umich.edu), replacing the U1A protein binding site added for crystallization purposes with the natural tetraloop of the L-9 *Azoarcus* ribozyme. All base pairs drawn as solid colored lines in the secondary structure shown in **Figure 6.1** were enforced as standard TOPRNA base pairs.

As described in more detail in **Section 6.3.1**, we also applied non-canonical pairing restraints to a selection of internal loop residues following the insight described in **Chapter 2** that internal loop motifs adopt fully non-canonically paired conformations. These non-canonical base pairs were either (a) inferred from the crystal structure, or (b) predicted using MC-Fold (19). The MC-Fold predictions were accomplished by imbedding the internal loop sequence and junction-closing base pairs into a well-defined hairpin and then choosing the predicted pairing with lowest energy that did not disrupt the junction-closing base pairs. Non-canonical base pairs were enforced as described in the TOPRNA documentation (brooks.chem.lsa.umich.edu). Briefly, an M bead is added to one of the participating residues to ensure the base pair fills the appropriate steric volume. The B beads of both residues are modified to feel the small attractive interaction to other paired B beads that all other paired B beads experience. Non-canonically paired residues are also given backbone dihedral potentials with $K_{nc}=¼K_c$, where K_c and K_{nc} are the force constants of canonically and non-canonically paired residues. Finally, flat-bottom NOE restraints are placed between the paired B beads with r_{min}=5.5 Å and r_{max}=7.5 Å, and between the S beads of the same residues with r_{min}=11 Å and r_{max}=14 Å, with $f_{max}=k_{min}=k_{max}$= 2 kcal/mol/Å^2. The specific non-canonical base pairs enforced in the crystal-structure-inferred simulation are: C11-A21, A48-A78, A49-A77, A50-C76, U53-U73, A54-A72, A89-A112, G90-A111, U92-U109, U139-U153, A140-A152, G173-A193, and A174-A192. The non-canonical pairs enforced in the MC-Fold-inferred simulation are: C11-A21, A48-A78, A49-A77, A50-C76, A54-U73, A55-A72, A88-A112, G90-A111, U92-U109, A140-U153, A141-A152, G173-A193, and A174-A189.

We also performed a third simulation where the backbone dihedrals of non-canonically paired internal loop residues were restrained to their crystal structure values. As an exception, the

backbone of the J5 and J8 11-nt receptors were restrained to the crystal structure of the free 11-nt receptor (PDB 1TLR (20)), which adopts a different *apo* conformation in the absence of a docked GAAA tetraloop. The applied restraints were given force constants of K=50 kcal/mol and had periodicity of 1. We note that the backbone of unpaired residues in internal loops were left unrestrained, with the exception of the J5 and J8 receptors where all internal loop residues were restrained to maintain the continuously stacked 'zipper' *apo* conformation.

Tertiary restrained simulations were performed using similar procedures as described for restrained tRNA simulations in **Chapter 4**. L2/P8 interactions were enforced using NOE restraints between the S-beads of A16 to U139 and A152, and A18 to C138 and G154. L9/P5 interactions were enforced using NOE restraints between the S-beads of A181 to A72 and U53, and A183 to C52 and G74. The triple helix was enforced by restraining the two backbone dihedrals separating P4 and P6 to their crystal structure values with K=50 kcal/mol, and placing NOE restraints between the S-beads of A39 to C87 and G113; A40 to C86 and G114; G116 to C43 and G83; and U117 to G44 and C82. All NOE restraints used $f_{max}=k_{min}=k_{max}=2$ kcal/mol/Å^2, with r_{min} and r_{max} set to ±1 Å of the value measured in the crystal structure.

6.2.2 Analysis

Analysis was restricted to conformations sampled at 300 K, with the first 2.5×10^7 dynamics steps discarded as equilibration. Unless otherwise noted, RMSDs to the 1U6B crystal structure and radius of gyrations were computed using all P beads. Interhelical (α_h, β_h, γ_h) angles and mutual information were measured as described in **Chapter 4**.

Well-packed folds were identified using a similar procedure as described in **Chapter 4** using the energy function $E=n_s\varepsilon_s+ n_l\varepsilon_l$. n_s is the number of interhelical stacking interactions in a

given conformation, computed as described in **Chapter 4** between P4 and P6, P3 and P8, and P7 and P9. n_l is the total number of loop-residue to loop-residue and loop-residue to helix contacts in the molecule, where a contact is defined as an inter-S-bead distance ≤14 Å, excluding contacts between residues less than five nucleotides apart or within the same structural element (i.e. both in J5). The inclusion of loop-residue to helix contacts differs from the procedure used in **Chapter 4**, where n_l was computed using only loop-residue to loop-residue contacts. The choice was made to include loop-residue to helix contacts because, unlike tRNA, such contacts comprise the majority of tertiary interactions in the *Azoarcus* ribozyme. The scaling parameters were set to ε_s=-3.5 kcal/mol and ε_l=-0.1 kcal/mol, with ε_l adjusted relative to the value used for tRNA due to the change in how n_l was computed. For completeness, we also identified the best-packed folds using the same parameters as used for tRNA and observed no significant differences (not shown). The entropies of the best-packed folds were computed as described in equation **[4.4]**, replacing the 10 Å RMSD radius in the denominator with 23 Å. The 23 Å value was chosen based on its status as the P=0.01 cutoff for structural similarity in 197 nt RNA (21).

6.2.3 Isolated TL/TLR motif simulations

Temperature replica exchange simulations of the isolated TL/TLR motif shown in **Figure 6.7** were performed using the same parameters used for the *Azoarcus* ribozyme simulations. The exchange rate between all replicas was 0.54. Initial coordinates were obtained using the `-fromc` option of *toprnaCreate.pl*, using the P5 hairpin of the *Azoarcus* crystal structure as the initial scaffold with all other structural elements generated by the *toprnaCreate.pl* initialization procedure. For the production simulations, backbone dihedral restraints were used to restrain all residues of the 11-nt receptor motif to their crystal structure values, using force constants of 50

kcal/mol and periodicity of one. Analysis was performed identically as it was for the full ribozyme.

6.3 Results and Discussion

6.3.1 TOPRNA simulations generate an experimentally reasonable unfolded ensemble of the Azoarcus ribozyme

Our studies of bulges and tRNA molecules in **Chapters 2-5** indicate that topological constraints strongly limit the 3D conformation of RNAs containing single junctions. However, for larger RNAs that contain multiple junctions, and hence possess many more degrees of freedom, the significance of topological constraints in governing molecular conformation is unclear. To explore the role of topological constraints in large RNA folding, we performed TOPRNA simulations of the *Azoarcus* ribozyme using 5×10^9 steps of temperature replica exchange molecular dynamics simulations. This unprecedented conformational sampling allows us to confidently map the constraints on the ribozyme and measure free energy differences as large as 11 k_BT between different conformations.

Unlike the molecules simulated with TOPRNA in previous chapters, the *Azoarcus* ribozyme contains multiple internal-loop two-way junctions. In **Chapter 2**, we showed that internal loop motifs have a strong tendency to fully non-canonically pair, effectively reducing them to bulges from a topological constraint perspective. By default, TOPRNA treats all residues not participating in canonical bases pairs as freely rotatable chains. Thus, in order to properly represent the constraints imposed by internal loops, it is necessary to infer and enforce non-canonical pairing between internal loop residues. We performed two independent simulations either (i) enforcing internal loop non-canonical pairs observed in the crystal structure (18), or (ii)

enforcing non-canonical base pairs predicted by MC-Fold (19) from primary sequence. Additionally, a simulation was performed restraining backbone dihedrals of internal loops to their crystal structure values. In all three simulations, unpaired internal loop residues remained freely rotatable.

Comparisons between the three different simulations revealed that the internal loop pairing scheme had a negligible effect on all molecular properties measured below (not shown), and we therefore limit our discussion to the simulation performed using non-canonical pairs inferred from the crystal structure. Thus, while the microscopic pairing arrangement of internal loop residues does in theory affect topological constraints (**Chapter 2**), these effects appear to be insignificant from a global structure perspective. This is an important result as one considers using TOPRNA to simulate RNA molecules with unknown internal loop pairings. It is also worth noting that while the Group I ribozyme internal loops are highly conserved, this is due to their role as docking sites for long-range tertiary interaction partners (1,12); their insignificance from a topological constraint perspective is thus not necessarily surprising.

As expected, global analysis of our simulations revealed that that the ribozyme quickly 'unfolds' from its initial crystal structure conformation, equilibrating to a highly heterogeneous ensemble of extended conformations (**Figure 6.2**). On average, the ribozyme has a mean radius of gyration (R_g) of ~42 Å, which compares to the $R_g \approx 30$ Å of the crystal structure. Thus, as expected from polymer physics (22), chain connectivity and steric repulsions pose a large entropic cost on adopting compact conformations, causing the molecule to swell considerably in the absence of attractive interactions. Notably, the $R_g \approx 42$ Å of our simulations is in good agreement with the radius of hydration $R_H \approx 40$ Å measured by analytical centrifugation for the unfolded *Azoarcus* ribozyme at low solution salt concentrations (23) (**Figure 6.2**). This R_g value

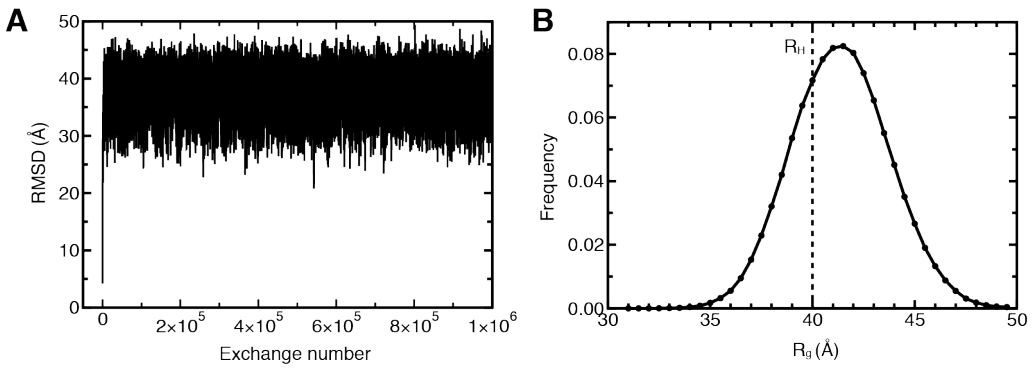

Figure 6.2: RMSD and R_g of the TOPRNA simulation of the *Azoarcus* ribozyme
(A) All P-bead RMSD to the 1U6B crystal structure as function of temperature replica exchange attempt. **(B)** Radius of gyration histogram. The radius of hydration measured by measured by analytical centrifugation in the presence of 100 mM KCl is drawn as a dashed line (23).

---

also agrees with that obtained from a previous coarse-grained study of the unfolded *Azoarcus* ribozyme (23). While it is difficult to assess with certainty, the agreement between our simulation and experiment suggests that TOPRNA provides a reasonable picture of the unfolded ribozyme ensemble. The extended nature of unfolded RNAs at low salt concentrations is typically thought to be a result of electrostatic repulsion between helices (10,24,25). Our simulations, along with the previous study by Chen *et al.* (23), indicate that the extended conformation of unfolded RNAs can be just as easily explained by the high entropic cost of closely packing the molecule; electrostatic repulsion at low salt concentrations may offset favorable tertiary interactions, resulting in the energetically neutral scenario modeled by our simulations. This explanation mirrors our rationalization in **Chapter 3** for why TOPRNA simulations reproduce experimental measurements made on bulges at low salt concentrations.

6.3.2 Topological constraints encode long-range interhelical correlations

To better characterize the effects of topological constraints on *Azoarcus* ribozyme conformation we followed the approach developed in prior chapters. Namely, we use Euler angles, (α_h, β_h, γ_h), to quantify the interhelical orientations sampled between pairs of helices across different

junctions. In addition to the ribozyme's much larger size, the interlocked three-way and four-way pseudoknotted junctions defined by P3, P4, P6, P7, and P8 represent a completely novel motif whose topological constraints have never before been quantified. Pseudoknots are important motifs in many RNA process, and are topologically distinct from other secondary structure elements; whereas strands enter and exit on the same end of a helix in a normal hairpin, in peusdoknots the strands enter and exit the helix on opposite ends (**Figure 6.1**).

The fraction of total interhelical conformations sampled between different pairs of helices is shown in **Figure 6.3A**. In general, this analysis reveals that helices are as constrained on a pairwise basis as observed in canonical tRNA (**Chapter 4**). Helices linked by intervening single strands are less constrained due to relaxed steric constraints. For example, P2 and P9, which are linked by only 1-3 single-stranded nucleotides to the rest the molecule, are largely unconstrained and sample 82-87% of their possible interhelical orientations relative to their nearest helix. Similarly, helices within the central four-way junction pseudoknot sample ~80% of possible orientations when linked by intervening single strands, and ~45% when directly linked. Notably, however, helices in the three-way pseudoknot are much more constrained. P3 and P7, and P7 and P8, only sample 45-55% of their possible relative orientations despite being linked by three and six single-stranded residues, respectively. Moreover, P3 and P8 are confined to only 25% their possible orientations, comparable to the constraints posed by the 3-nt bulge between P9 and P9A (**Figure 6.3A**).

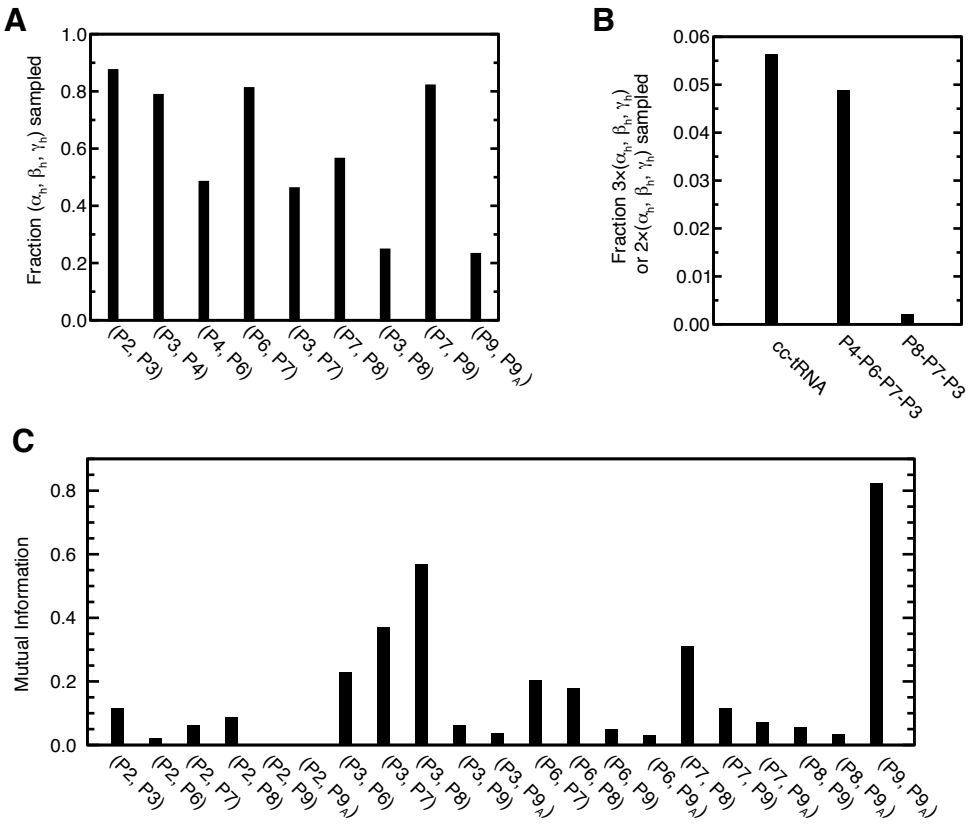

Figure 6.3: Interhelical angles sampled by the *Azoarcus* ribozyme
(A) The fraction of total (α_h, β_h, γ_h) space sampled between pairs of *Azoarcus* ribozyme helices computed on a 10° grid. **(B)** The fraction of total 2×(α_h, β_h, γ_h) and 3×(α_h, β_h, γ_h) global interhelical conformations sampled by the three-way and four-way pseudoknotted junctions, computed using a 60° grid. The fraction of 3×(α_h, β_h, γ_h) angles sampled by tRNA from **Chapter 4** is shown as reference. **(C)** The mutual information between the (α_h, β_h, γ_h) distributions of different helices. The (α_h, β_h, γ_h) of each helix was computed relative to P4.

To further determine the degree to which the three-way and four-way pseudoknotted junctions are globally constrained, we computed the number of unique 2×(α_h, β_h, γ_h) and 3×(α_h, β_h, γ_h) conformations sampled by each, respectively. For these calculations, P8 and P4 were used as the respective internal reference helices for the three-way and four-way junctions. This analysis reveals that the P4-P6-P7-P3 four-way junction samples only ~5% of possible global interhelical conformations and the P8-P7-P3 three-way junction only ~0.2% of possible conformations. Notably, this means that both pseudoknotted junctions are more constrained than the canonical tRNA four-way junction, despite containing more single-stranded residues. Thus,

as might be naively presumed, pseudoknots place strong topological constraints on RNA 3D conformation. Due to the large number of junctions in the ribozyme, it is impossible to estimate the degree to which the molecule as a whole is constrained by topological constraints. Nevertheless, these results indicate that topological constraints strongly limit the 3D conformation of the *Azoarcus* ribozyme.

As was observed for tRNA, one consequence of the *Azoarcus* ribozyme's topological constraints is that the orientations of different helices are correlated. We calculated the (α_h, β_h, γ_h) orientation of all helices with respect to P4, and then computed the Mutual Information (MI) between different pairs of (α_h, β_h, γ_h) as a measure of the interhelical correlation. As described in **Chapter 4**, MI is a general measure of correlation that is independent of linearity assumptions, with zero MI indicating zero correlation. Strikingly, this analysis reveals small but wide-ranging correlations throughout the entire molecule. Helices linked by the central three-way and four-way junction pseudoknots are correlated with $0.2<MI<0.6$, comparable or greater than what is observed between the helices of tRNA (**Chapter 4**). P6 and P8, which are separated by two intervening helices, are correlated with $MI\approx0.2$. While significantly attenuated, residual correlations of $MI\approx0.04$ are also observed between P6 and P2, and P6 and P9, indicating the presence of long-range communication between the core and these peripheral domains. By contrast, P2 and P9 are completely uncorrelated ($MI=0$), as expected given that they are both only loosely coupled to the rest of the molecule.

6.3.3 Topological constraints penalize formation of non-native tertiary contacts

In our studies of tRNA in **Chapters 4-5** we showed that topological constraints contribute to the specificity of tRNA 3D structure by preventing formation of non-native tertiary contacts. We

therefore explored whether topological constraints similarly contribute to Azoarcus ribozyme folding specificity. In particular, such a mechanism could help explain how the ribozyme can utilize identical GAAA/11-nt-receptor TL/TLR motifs in L2/J8 and L9/J5 without suffering from tertiary misfolding (**Figure 6.1**). The other primary set of tertiary interactions that stabilize the *Azoarcus* ribozyme, A-minor and base-triple motifs, also have limited specificity (4,26).

Repeating the approach described in **Chapter 4**, we computed the free energy cost that topological constraints impose on forming tertiary contacts from the probability $P(r_i,r_j)$ that residues r_i and r_j come within a S-S bead distance of 14 Å. These free energies are shown in **Figure 6.4** for a selection of single-stranded regions to all other residues of the ribozyme. Consistent with our findings in tRNA, the free energy penalty for forming different long-range contacts in the *Azoarcus* is highly variable, with ΔG^{topo} varying from as little as 1 k_BT to greater than 10 k_BT. Remarkably, once again, these penalties appear to be greater for non-native contacts than native contacts (native contacts are outlined in **Figure 6.4**), indicating that topological constraints are providing an inherent source of folding specificity. General features of the two different classes of tertiary interactions – A-minor/triple-type interactions and TL/TLR motifs – are discussed in detail below.

Of the two tertiary interaction classes, native A-minor and base-triple-type interactions form with the lowest ΔG^{topo} penalties, with $\Delta G^{topo}<5$ k_BT for all interactions and frequently <3 k_BT (e.g. interactions between J2/3, J3/4, J6, J6/7, J7/3, and J8/7 to helices; **Figure 6.4**). Moreover, the difference in ΔG^{topo} between native and non-native contacts is large, typically >3 k_BT. As A-minor and base-triple interactions consist of only several hydrogen bonds and generally contribute no more than -3 k_BT in stabilizing energy (27), the 3 k_BT difference encoded by topological constraints is sufficient to strongly disfavor formation of non-native contacts. This

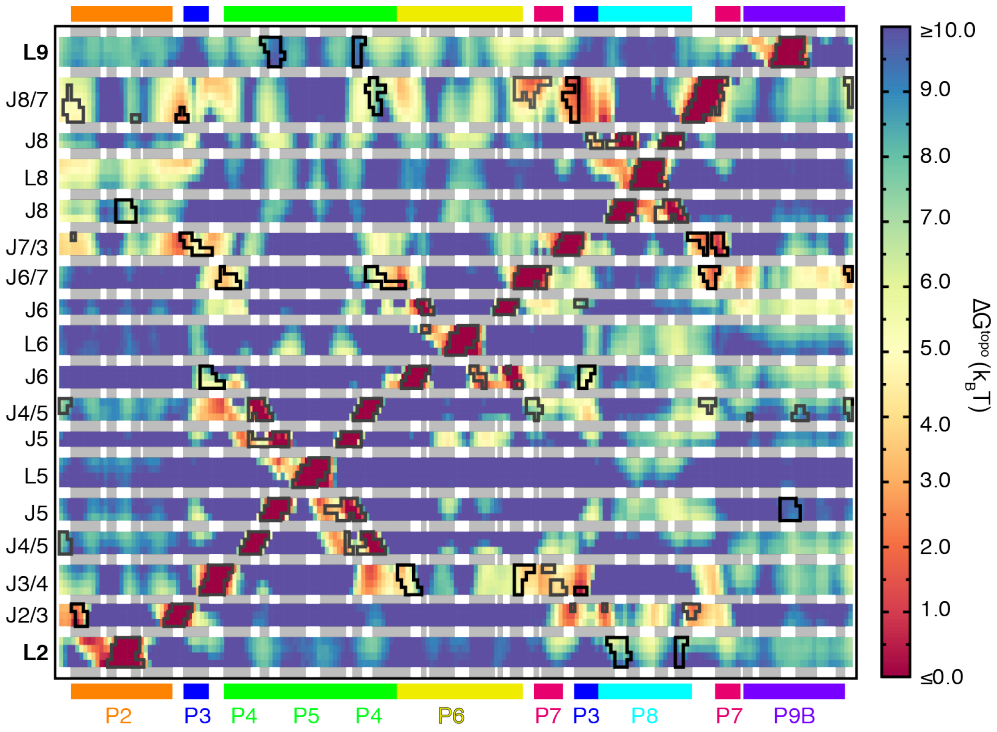

Figure 6.4 ΔG^{topo} of forming different pairwise tertiary contacts in the *Azoarcus* ribozyme
Selected loop regions are shown along the y-axis, ordered by sequence number from 5' to 3' end. Along the x-axis are shown all residues, also ordered 5' to 3'. Colored bars drawn along the bottom of the x-axis are used to indicate different secondary structure elements, with the color scheme the same as in **Figure 6.1**. Vertical gray backgrounding is used to indicate paired regions. S-S contacts observed in the 1U6B crystal structure are outlined in dark gray and black. Regions drawn in black make actual tertiary interactions, whereas regions drawn in gray correspond to residues that are in close proximity in the crystal structure but do not participate in tertiary interactions.

---

high degree of topological constraint specificity helps explain why large RNAs are able to use these sequence-non-specific interactions without suffering tertiary misfolding consequences. Interestingly, a similar pattern is observed in tRNA in **Chapter 4**, where the less-sequence-specific base-triple interactions between the V-loop and D-stem/loop form with lower ΔG^{topo} penalties and greater topological constraint specificity. Together, these observations suggest that A-minor and base-triple-type interactions are primarily 'opportunistic', occurring between loops and helices that are topologically local with few alternative available interaction partners.

Compared to the A-minor and base-triple class of interactions, the TL/TLR contacts between L2/J8 and L9/J5 possess large ΔG^{topo} penalties of >6 k_BT (**Figure 6.4**). In addition, the

TL/TLR motifs have less topological-constraint-encoded specificity. For example, L2 forms contacts with P7 with a comparable or smaller ΔG^{topo} penalty than with its native J8 receptor, and can also form contacts with P4, P5, and P9 with only 1-2 k_BT higher penalty (**Figure 6.4**). Likewise, the ΔG^{topo} penalty of forming native contacts between L9 and J5 is 7-8 k_BT, but is <6 k_BT for forming non-native contacts between L9 and P6 or P7 (**Figure 6.4**). It is worth emphasizing that this analysis only evaluates whether residues can come into close proximity; finer-grained analyses that also consider interaction geometry may reveal that L2 and L9 have greater topological-constraint-encoded specificity. Nevertheless, these results suggest that the sequence specificity of TL/TLR interactions (28) is important for preventing tertiary misfolding. Significantly, however, topological constraints strongly discriminate against formation of non-native TL/TLR interactions. Native L2/J8 TL/TLR contacts form with ~1.5 k_BT smaller ΔG^{topo} than non-native L2/J5 contacts, and with a ~2.0 k_BT smaller penalty than the competing L9/J8 interaction (**Figure 6.5**). For L9, the native L9/J8 interaction forms with ~0.5 k_BT smaller penalty ΔG^{topo} than the non-native L9/J8. Thus, for these two motifs that have identical sequences, topological constraints provide an important source of folding specificity.

An interesting difference between the two TL/TLR interactions is the ~1.5 k_BT higher ΔG^{topo} of forming L9/J5 contacts relative to L2/J8. The greater cost of L9/J5 formation correlates with the high evolutionary conservation of GAAA and 11-nt receptor motifs at L9 and J5, which form with the highest affinity of any TL/TLR interaction (1,29). By contrast, L2 and J8 are frequently substituted with weaker TL/TLR motifs in other group I introns (12,29). The higher ΔG^{topo} of L9/J5 contacts can be rationalized by the greater number of flexible linker/pivot points separating the different long-range interactions along the secondary structure. The shortest secondary structure path from L9 to J5 passes through four such flexible points – J9, J7/9, J6/7,

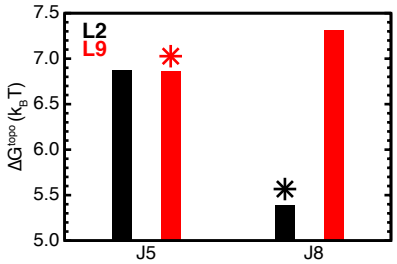

Figure 6.5: The ΔG^{topo} of forming native and non-native TL/TLR interactions
ΔG^{topo} values were computed from the probability of observing at least a single S-S contact between the respective TL and TLR. Native TL/TLR interactions are indicated by asterisks.

---

and the P4/P6 linker. By comparison, only two flexible points separate L2 and J8. Approximating these two scenarios as five-mer and three-mer ideal chains, where the monomers are uniformly long secondary structure elements, L9 and J5 would be expected to form contacts with 0.4 lower probability than L2 and J8 (30). This decreased probability translates to a ~0.8 k_BT higher ΔG^{topo}, within a reasonable range of the ~1.5 k_BT difference we observe. This comparison can also be made to tRNA, where the D- and T-loops are separated by two flexible points and form with ΔG^{topo} comparable to L2 and J8 (**Figure 4.2**). Thus, while RNA secondary structure is clearly not an ideal chain, this comparison reveals that the higher ΔG^{topo} can be understood in the context of well-established polymer physics.

6.3.4 Well-packed folds are too sparsely sampled to identify native 3D structure

In **Chapters 4-5**, we were able to harness the tertiary specificity of topological constraints to predict native tRNA 3D structure. In particular, we found that topological constraints prevented tRNA from forming non-native well-packed conformations, thereby allowing us to identify the native 3D structure as the most entropically favored well-packed conformation. Given the similar evidence of topological-constraint-encoded tertiary specificity, we thus repeated this procedure for the *Azoarcus* ribozyme. As before, we identified the 500 'best-packed' conformers from our simulation that have the maximal number of long-range contacts between single-stranded loops

and inter-helical stacking interactions. The entropy of each conformer was then computed based on how many structurally similar conformers appeared in the best-packed ensemble. Unlike for tRNA, this procedure was unable to identify the native fold (defined here as RMSD < 23 Å; (21)) (**Figure 6.6**). While several low RMSD conformations appear in the subset, the significant majority of conformations are non-native. Many non-native conformations also have high entropies, indicating that these conformations are not heterogeneous, but rather adopt similar 3D structures. Analysis of these high entropy conformations revealed that they correspond to extended conformations that form many core contacts, but do not possess any contacts involving the L2 or L9 TLs (**Figure 6.6**). Thus, while the identified 'well-packed' conformations are not biased towards the native structure, neither are they biased towards a viable non-native fold. Indeed, none of the 'best-packed' conformers are particularly well-packed, indicating that such conformations are simply not sampled.

The paucity of well-packed folds in our simulation can be rationalized by the very high entropic cost associated with forming such conformations (**Figure 6.2**). For example, given that P2 and P9 are completely uncoupled from one another (**Figure 6.3**), the free energy of forming L2/J8 and L9/J5 TL/TLR contacts simultaneously is expected to be >12 k_BT. Given this energy, the expectation is that only 6 of the 10^6 *Azoarcus* ribozyme conformations generated by our simulations should possess both contacts.

As an attempt to overcome the sparse sampling of compact conformations, we performed an additional TOPRNA simulation where the radius of gyration was restrained to 32 Å (the SAXS measured value (5)). We then used the same procedure as above to identify and compute the entropies of the 500 best-packed conformations. However, this analysis also failed to yield native-like conformations, and in fact identified a number of non-native well-packed folds as the

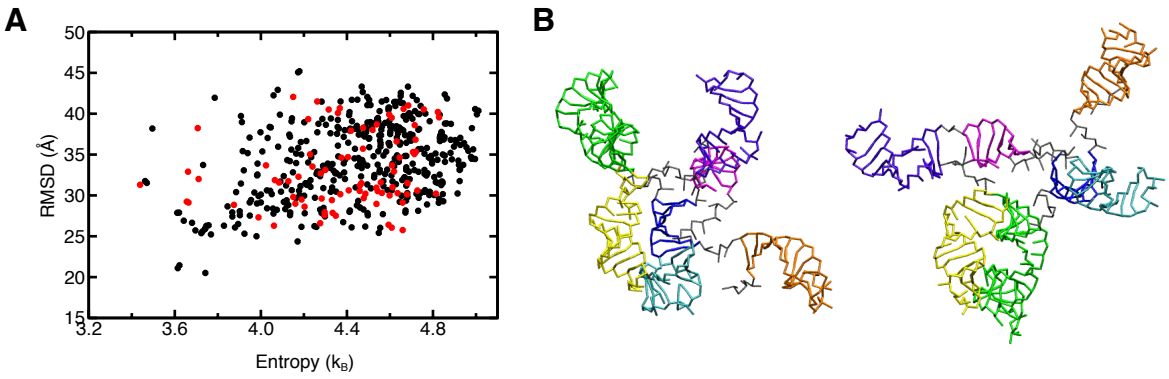

Figure 6.6: Well-packed folds sampled by the *Azoarcus* ribozyme
(A) Entropies and all P-bead RMSDs of the 500 best-packed conformers sampled by the *Azoarcus* ribozyme simulation. Conformations possessing contacts between L2 or L9 and any helical or internal loop region are drawn in red. **(B)** Two representative high entropy 'well-packed' conformations. The structural elements are colored as in **Figure 6.1**.

---

most entropically favored (not shown). Further analysis revealed that the radius of gyration restraint strongly disfavored P4/P6 stacking due to the fact that the end-to-end distance of the P4/P6 coaxial stack is ~100 Å. As discussed below, P4/P6 stacking is a central organizing feature of the ribozyme 3D structure, and thus it is not surprising that native-like conformations are underrepresented in this simulation.

6.3.5 Comparisons to an isolated TL/TLR motif indicate that topological constraints strongly destabilize Azoarcus TL/TLR motifs

Collectively, our results indicate that the secondary structure of the *Azoarcus* ribozyme possess very strong entropic penalties on tertiary folding. However, it is it difficult to fully appreciate the consequences of these penalties on folding stability without some external reference. To better contextualize the ~5.4 k_BT and ~6.9 k_BT ΔG^{topo} penalties of forming native L2/J8 and L9/J5 GAAA/11-nt-receptor interactions, we performed simulations of an isolated GAAA/11-nt-receptor motif that has been well characterized experimentally (**Figure 6.7A**) (31). These simulations reveal that the ΔG^{topo} for the isolated motif is ~4.2 k_BT. Experimentally, this motif

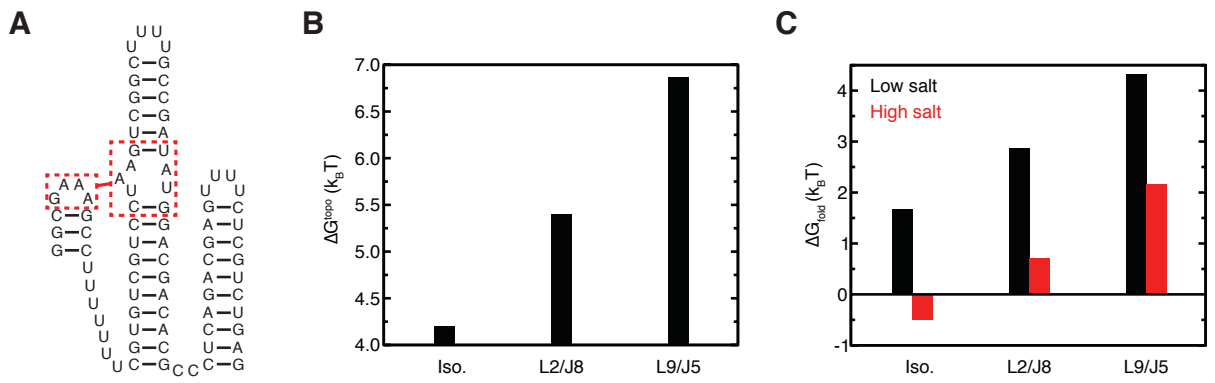

Figure 6.7: Comparison of the ΔG^{topo} between isolated and *Azoarcus* TL/TLR interactions
(A) Secondary structure of the isolated TL/TLR motif characterized experimentally by Nesbitt and colleagues (31). **(B)** ΔG^{topo} values computed from the probability of observing at least a single S-S contact between the respective TL and TLR for the different motifs. **(C)** ΔG_{fold} values measured experimentally for the isolated motif and extrapolated to the *Azoarcus* TL/TLR motif at low and high salt concentrations (31). Extrapolated values for TL/TLR motif i were obtained from the equation $\Delta G_{fold}(i) = \Delta G_{fold}(iso) + (\Delta G^{topo}(i) - \Delta G^{topo}(iso))$

---

was measured to fold with ΔG_{fold}=-0.5 k_BT at high salt concentrations and ΔG_{fold}=1.6 k_BT at low salt concentrations. Similar ΔG_{fold} values were obtained when the U_7 single-stranded linker was replaced with an A_7 linker. Therefore, assuming that all other components of the ΔG_{fold} of the GAAA/11-nt-receptor interaction are equivalent in the ribozyme and isolated motif contexts, we predict that L2/J8 should have $0.6 \leq \Delta G_{fold} \leq 2.8$ k_BT, and L9/J5 should have $2.2 \leq \Delta G_{fold} \leq 4.3$ k_BT (**Figure 6.7C**). Thus, topological constraints disfavor individual formation of these motifs, particularly L9/J5.

Notably, these predicted ΔG_{fold} values for L2/J8 and L9/J5 correlate well with experimental observations. Both equilibrium and time-resolved hydroxyl radical probing experiments indicate that L2/J8 forms first and is independently stable (15,32), with a folding midpoint (ΔG_{fold}=0) at ~0.7 mM $MgCl_2$ (32). This is in good agreement with our prediction that ΔG_{fold} should be ~0.6 k_BT by comparinng to the isolated TL/TLR motif at 1 mM $MgCl_2$ (high salt). By contrast, formation of the L9/J5 interaction occurs late in the folding pathway, coupled to formation of other tertiary interactions, and has the highest $MgCl_2$ folding midpoint of all

tertiary interactions (15,32). Similarly, our simulations predict that $\Delta G_{fold} \approx 2.2\ k_B T$ even at high $MgCl_2$ concentrations, indicating that L9/J5 is unlikely to form by itself. As discussed above, the differential stability of the two motifs also correlates well with the stronger evolutionary pressures on L9/J5 compared to L2/J8.

6.3.6 Triple-helix formation promotes cooperative folding of L9/J5

An implication of the above results is that topological constraints provide the basis of a natural hierarchical folding mechanism. Here, the comparably low ΔG^{topo} of L2/J8 allows it to form early and independently, whereas formation of L9/J5 is disfavored by a high ΔG^{topo} penalty until late in the folding pathway.

Interestingly, recent experiments have shown that formation of L9/J5 is strongly thermodynamically coupled to tertiary interactions between J6/7 and P4, and J3/4 and P6, which together comprise the central triple helix (TH) motif (5). While L9/J5 appears to form concomitantly with the TH at the time and structural resolution of experiments (15,32), there is strong reason to believe that TH folding precedes L9/J5 formation. Notably, in the related *Tetrahymena* group I ribozyme, folding of the P4/P6 domain (which requires TH folding) is a critical first folding step that nucleates assembly of the rest of ribozyme (33). Early modeling studies of Group I introns also observed that the TH motif places strong constraints on the rest of the ribozyme core, helping to organize the global architecture of the RNA (12). Finally, the TH is among the most highly conserved structural elements of all group I ribozymes (12), and although it does not directly play a role in catalysis (14), it is required for enzymatic activity due its crucial structural role (33-35). Our simulations also support that the TH can form more easily than L9/J5, with $\Delta G^{topo} \approx 3.5\ k_B T$ for both sets of base triple interactions (**Figure 6.4**). Thus,

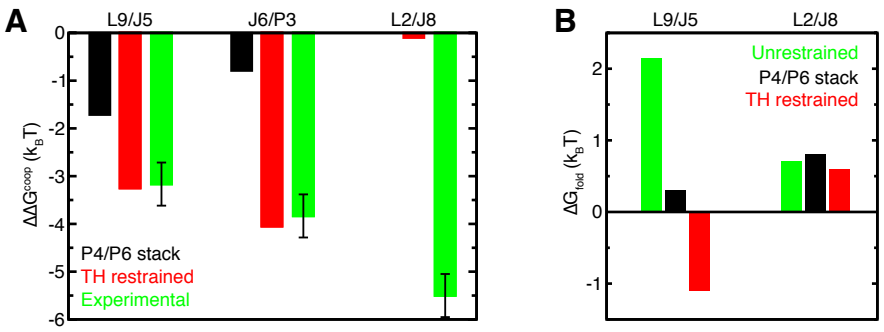

Figure 6.8: Cooperativity between the TH and distal tertiary interactions
(A) Thermodynamic cooperativity between the indicated tertiary interactions compared to values measured experimentally (5). $\Delta\Delta G^{coop}$ between interaction i and either P4/P6 stacking or restraint of TH was computed as $\Delta\Delta G^{coop} = \Delta G^{topo}(i \mid X) - \Delta G^{topo}(i)$. **(B)** ΔG_{fold} values for TL/TLR interactions conditional on either P4/P6 stacking or restraint of TH, computed according to $\Delta G_{fold}(i \mid X) = \Delta G_{fold}(iso) + (\Delta G^{topo}(i \mid X) - \Delta G^{topo}(iso))$ using the $\Delta G_{fold}(iso)$ measured at high salt concentrations (31).

---

based on this evidence, we hypothesized that formation of the TH precipitates L9/J5 formation by reducing ΔG^{topo}.

As a first of test of this hypothesis, we used our unrestrained simulation to compute the ΔG^{topo} of forming L9/J5 contacts conditional on P4 and P6 being coaxially stacked. This calculation revealed that the ΔG^{topo} of L9/J5 is lowered to ~5 k_BT, indicating cooperativity contributes ~2 k_BT to L9/J5 stability (**Figure 6.8**). This compares favorably to the ~3 k_BT cooperatvity observed experimentally between L9/J5 and the TH at moderate $MgCl_2$ concentrations (5). Moreover, returning to the comparison of the isolated GAAA/11-nt receptor above, this reduction in ΔG^{topo} reduces the predicted ΔG_{fold} of L9/J5 from 2.1 k_BT to ~0 k_BT, indicating that this cooperativity is sufficient to promote L9/J5 folding (**Figure 6.8**).

It is important to note that above calculation only measures the cooperativity between P4/P6 stacking and L9/J5 and does not presuppose formation of the TH interactions. To further explore the role of the TH interactions in folding we performed an additional simulation of the *Azoarcus* ribozyme where the central triple helix was restrained to its native conformation. Free energies of forming different contacts computed from this simulation thus reflect $\Delta G^{topo \mid TH}$.

Strikingly, $\Delta G^{topo \mid TH}$(L9/J5) ≈ 3.6 k_BT, indicating a cooperative contribution of ~3.1 k_BT that is within error of the experimental measurement (**Figure 6.8**). Referenced to the GAAA/11-nt receptor, the $\Delta G_{fold} \approx$ -1.1 k_BT, indicating L9/J5 folding is significantly favored. Thus, couplings between helical domains arising from topological constraints are sufficient to account for the cooperativity between the TH and L9/J5 tertiary interactions.

The conformational sampling of the TH restrained simulation also allowed us to compute the whole $\Delta G^{topo \mid TH}$ map for all pairs of residues (**Figure 6.9**). This reveals TH formation has a broad effect throughout the molecule, as was observed experimentally (5). Most A-minor and base-triple interactions in the core are strongly promoted, and as shown in **Figure 6.8**, the cooperativities are in reasonable agreement with experimental measurements. The exception to this agreement, the L2/J8 interaction, is discussed in detail below. More notable is the large increase in native-specificity of the free energy landscape – whereas $\Delta G^{topo \mid TH}$ is lower than ΔG^{topo} for native contacts, $\Delta G^{topo \mid TH}$ is significantly higher for non-native interactions (**Figure 6.9**).

In summary, our results indicate that topological constraints disfavor formation of L9/J5 in the absence of the folded TH and P4/P6 stacking. Upon TH folding, which is topologically favored, the energetic penalty for forming L9/J5 contacts decreases by ~3 k_BT, making L9/J5 formation exothermic. TH folding also substantially favors formation of other tertiary contacts, and increases the native-specificity of the free energy landscape. These results provide strong support that secondary structure, through topological constraints, encodes a natural hierarchical folding mechanism and provides the basis of folding cooperativity.

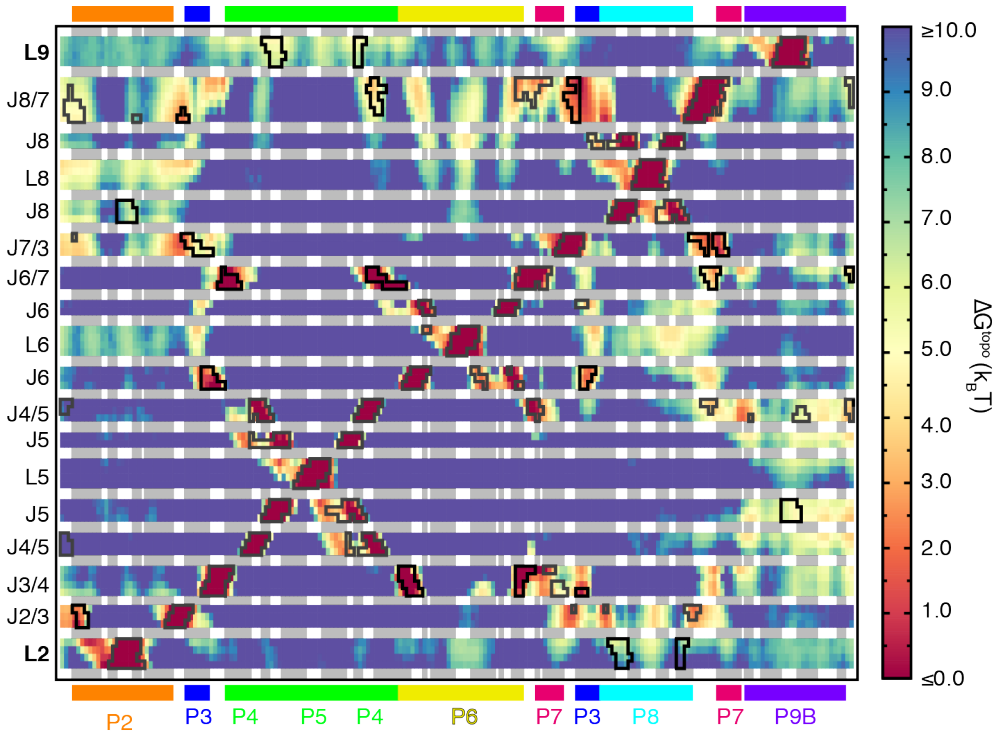

Figure 6.9: ΔG^{topo} of forming different tertiary contacts upon restraint of TH
Selected loop regions are shown along the y-axis, ordered by sequence number from 5' to 3' end. Along the x-axis are shown all residues, also ordered 5' to 3'. Colored bars drawn along the bottom of the x-axis are used to indicate different secondary structure elements, with the color scheme the same as in **Figure 6.1**. Vertical gray backgrounding is used to indicate paired regions. S-S contacts observed in the 1U6B crystal structure are outlined in dark gray and black. Regions drawn in black make actual tertiary interactions, whereas regions drawn in gray correspond to residues that are in close proximity in the crystal structure but do not participate in tertiary interactions.

6.4 Discussion and conclusions

The folding mechanisms used by large RNAs have been the subject of intense research. In general, it is known that secondary structure typically folds first, followed by counterion-mediated collapse to an ensemble of both native-like and misfolded compact intermediates, and final consolidation of tertiary structure (10). However, fundamental questions such as how large RNAs are able to fold to specific tertiary structures despite using non-specific tertiary interactions, and the source of the thermodynamic cooperativity that exists between these interactions, have remained poorly understood. Our simulations of the *Azoarcus* group I ribozyme suggest that topological constraints are an important component to the answers of these questions. In particular, we observe that topological constraints strongly confine the set of 3D

conformations accessible to the *Azoarcus* ribozyme and impose large, variable free energy penalties on forming different tertiary interactions. These free energy penalties are harnessed to selectively penalize formation of non-native tertiary interactions, providing a source of folding specificity, and long-range interhelical correlations arising from topological confinement provide a basis for encoding tertiary interaction cooperativity and hierarchical folding behavior.

As might be expected given their lack of sequence-specificity, A-minor and base-triple tertiary interactions appear to rely particularly heavily on topological-constraint-encoded specificity. We observe that native A-minor and base-triple interactions occur between structural regions that form with low ΔG^{topo} (<5 k_BT), and putative non-native A-minor or base-triple interactions are discriminated against via large ΔG^{topo} penalties. These features likely serve two important roles in folding. First, the low ΔG^{topo} of these contacts suggests that they have a low barrier to formation. This may help create a funneled folding landscape, where topologically local A-minor and base-triple interactions form as the molecule compacts to help offset the large entropic penalty associated with compaction (36). Second, topological-constraint-specificity helps discriminate against formation of non-native contacts, smoothing the folding funnel.

By contrast, TL/TLR interactions have both high ΔG^{topo} penalties and lower topological-constraint-encoded specificities. Unlike A-minor and base-triple motifs, TL/TLR interactions form with both high thermodynamic stability and sequence specificity (1), reducing the need for topological constraints to specify the folding landscape. Nonetheless, topological constraints do appear to play a key role in preventing the *Azoarcus* ribozyme from forming non-native TL/TLR interactions. Given their significant favorable enthalpies, formation of non-native TL/TLR interactions would be expected to pose large kinetic barriers to native tertiary folding (37). Thus, the ability of topological constraints to discriminate between competing TL/TLR interactions

likely explains why *Azoarcus* and other group I introns can use multiple of these motifs without suffering from tertiary misfolding.

Strikingly, in agreement with a model where formation of local A-minor and base-triple interactions help funnel the folding landscape, we observe that P4/P6 stacking and formation of the central triple-helix further increases both the specificity and stability of the L9/J5 TL/TLR interaction. The thermodynamic cooperativity we observe between TH and L9/J5, -3 k_BT, is within error of the value measured experimentally (5). Based on comparisons to a reference isolated TL/TLR motif, this cooperative contribution is sufficient to shift the free energy of L9/J5 formation from a large positive value to a negative value. Furthermore, L9 is topologically prohibited from forming contacts with the competing J8 TLR (**Figure 6.9**), indicating that native-collapse becomes progressively more specific.

Together, these data support a cohesive model where topological constraints of the native secondary structure encode a hierarchical folding funnel whose entry is initiated by formation of local A-minor and base-triple interactions. Thermodynamic cooperativity between these interactions in turn promotes formation the otherwise disfavored L9/J5 tertiary interaction. As discussed above, such a folding funnel has clear benefits by helping to progressively compensate for the large entropic penalty associated with RNA compaction. However, such a folding funnel may also play an important role in preventing formation of topologically frustrated tertiary misfolded kinetic intermediates. In particular, a recent study of the *Azoarcus* ribozyme used a non-equilibrium approach to progressively enforce different native tertiary interactions on the prefolded secondary structure (23). In the majority of cases, this non-equilibrium procedure resulted in topologically frustrated tertiary misfolds.

It is important to note that the tertiary interaction cooperativity observed in our simulations corresponds to the experimental cooperativity observed at moderate $MgCl_2$ concentrations, where the *Azoarcus* ribozyme exists as an ensemble of native-like compact intermediates (5). At higher salt concentrations, L9/J5 and the TH are no longer thermodynamically coupled, which Woodson and colleagues (5) propose is due to salt-induced reorganization of the P7 active site that antagonizes L9/J5. Given that we observe that the cooperativity between the TH and L9/J5 is inherent to the ribozyme architecture, our simulations support that antagonism must be responsible for the lack of TH and L9/J5 cooperativity at high salt.

A conspicuous exception to the agreement between our results and those of Woodson and colleagues (5) is the complete lack of cooperativity between L2/J8 and the TH in our simuatlions (**Figure 6.8**). Visual analysis of the TH-restrained simulation revealed that the TH places few constraints on either P2 or P8, as both of these helices are separated from the TH by the P3 helix. As a first explanation for this disagreement, we hypothesized that formation of the TH may promote formation of the highly conserved base triple interactions between the top of J8/7, J6/7, P4, and P7, which would then further restrict the P3-P7-P8 junction and thereby favor L2/J8 formation (**Figure 6.1**). We tested this hypothesis by performing an additional simulation where both the TH and these base triples were restrained. However, this simulation also revealed that the ΔG^{topo} of L2/J8 was minimally affected (not shown). Electrostatic or attractive interactions not modeled by our simulations could provide the source of the experimentally observed cooperativity between these tertiary interaction motifs. However, we suggest an alternative explanation that acts on the secondary structure level. Here, L2/J8 formation helps stabilize the P3 pseudoknot, which in turn stabilizes TH and the rest of the molecule. Such a mechanism fits

with the early folding of L2/J8 (15,32), and provides an interesting hypothesis to test in the future.

Clearly, our simulations represent a drastic oversimplification of the group I ribozyme folding landscape. Much of the complexity that makes the folding of these RNAs so interesting arises from secondary structure misfolding that our models cannot capture (10,25). Nevertheless, our results strongly argue that topological constraints play a critical role in defining the tertiary folding energy landscape of the *Azoarcus* ribozyme. Combined with our studies of tRNA, this suggests that topological constraints are a universally important determinant of RNA 3D structure and dynamics.

6.5 References

1. Costa, M. and Michel, F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*, **14**, 1276-1285.
2. Doherty, E.A., Batey, R.T., Masquida, B. and Doudna, J.A. (2001) A universal mode of helix packing in RNA. *Nat Struct Biol*, **8**, 339-343.
3. Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B. and Steitz, T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci USA*, **98**, 4899-4903.
4. Battle, D.J. and Doudna, J.A. (2002) Specificity of RNA-RNA helix recognition. *Proc Natl Acad Sci USA*, **99**, 11676-11681.
5. Behrouzi, R., Roh, J.H., Kilburn, D., Briber, R.M. and Woodson, S.A. (2012) Cooperative tertiary interaction network guides RNA folding. *Cell*, **149**, 348-357.
6. Sattin, B.D., Zhao, W., Travers, K., Chu, S. and Herschlag, D. (2008) Direct measurement of tertiary contact cooperativity in RNA folding. *J Am Chem Soc*, **130**, 6085-6087.
7. Kilburn, D., Roh, J.H., Behrouzi, R., Briber, R.M. and Woodson, S.A. (2013) Crowders perturb the entropy of RNA energy landscapes to favor folding. *J Am Chem Soc*, **135**, 10055-10063.
8. Kilburn, D., Roh, J.H., Guo, L., Briber, R.M. and Woodson, S.A. (2010) Molecular crowding stabilizes folded RNA structure by the excluded volume effect. *J Am Chem Soc*, **132**, 8690-8696.
9. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E. and Cech, T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147-157.

10. Woodson, S.A. (2010) Compact intermediates in RNA folding. *Annu Rev Biophys*, **39**, 61-77.
11. Cech, T.R. (1987) The chemistry of self-splicing RNA and RNA enzymes. *Science*, **236**, 1532-1539.
12. Michel, F. and Westhof, E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol*, **216**, 585-610.
13. Michel, F., Jacquier, A. and Dujon, B. (1982) Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie*, **64**, 867-881.
14. Woodson, S.A. (2005) Structure and assembly of group I introns. *Curr Opin Struct Biol*, **15**, 324-330.
15. Mitra, S., Laederach, A., Golden, B.L., Altman, R.B. and Brenowitz, M. (2011) RNA molecules with conserved catalytic cores but variable peripheries fold along unique energetically optimized pathways. *RNA*, **17**, 1589-1603.
16. Russell, R., Das, R., Suh, H., Travers, K.J., Laederach, A., Engelhardt, M.A. and Herschlag, D. (2006) The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J Mol Biol*, **363**, 531-544.
17. Brooks, B.R., III, C.L.B., Jr, A.D.M., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. *et al.* (2009) CHARMM: The biomolecular simulation program. *J Comput Chem*, **30**, 1545-1614.
18. Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J. and Strobel, S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45-50.
19. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51-55.
20. Butcher, S.E., Dieckmann, T. and Feigon, J. (1997) Solution structure of a GAAA tetraloop receptor RNA. *EMBO J*, **16**, 7490-7499.
21. Hajdin, C.E., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2010) On the significance of an RNA tertiary structure prediction. *RNA*, **16**, 1340-1349.
22. Richards, E.G. (1980) *An introduction to the physical properties of large molecules in solution*. Cambridge University Press, New York.
23. Chen, C., Mitra, S., Jonikas, M., Martin, J., Brenowitz, M. and Laederach, A. (2013) Understanding the role of three-dimensional topology in determining the folding intermediates of group I introns. *Biophys J*, **104**, 1326-1337.
24. Bai, Y., Chu, V.B., Lipfert, J., Pande, V.S., Herschlag, D. and Doniach, S. (2008) Critical assessment of nucleic acid electrostatics via experimental and computational investigation of an unfolded state ensemble. *J Am Chem Soc*, **130**, 12334-12341.
25. Thirumalai, D. and Hyeon, C. (2009) In Walter, N. G., Woodson, S. A. and Batey, R. (eds.), *Non-Protein Coding RNAs*. Springer-Verlag Berlin, Heidelberger Platz 3, D-14197 Berlin, Germany, Vol. 13, pp. 27-47.
26. Gautheret, D., Damberger, S.H. and Gutell, R.R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J Mol Biol*, **248**, 27-43.
27. Silverman, S.K. and Cech, T.R. (1999) Energetics and cooperativity of tertiary hydrogen bonds in RNA structure. *Biochemistry*, **38**, 8691-8702.
28. Fiore, J.L. and Nesbitt, D.J. (2013) An RNA folding motif: GNRA tetraloop-receptor interactions. *Q Rev Biophys*, **46**, 223-264.

29. Prathiba, J. and Malathi, R. (2008) Group I introns and GNRA tetraloops: remnants of 'The RNA world'? *Mol Biol Rep*, **35**, 239-249.
30. Dill, K.A. and Bromberg, S. (2003) *Molecular driving forces: statistical thermodynamics in chemistry and biology*. Garland Science.
31. Fiore, J.L., Holmstrom, E.D. and Nesbitt, D.J. (2012) Entropic origin of Mg2+-facilitated RNA folding. *Proc Natl Acad Sci USA*, **109**, 2902-2907.
32. Chauhan, S. and Woodson, S.A. (2008) Tertiary interactions determine the accuracy of RNA folding. *J Am Chem Soc*, **130**, 1296-1303.
33. Zarrinkar, P.P. and Williamson, J.R. (1996) The kinetic folding pathway of the Tetrahymena ribozyme reveals possible similarities between RNA and protein folding. *Nat Struct Biol*, **3**, 432-438.
34. Green, R. and Szostak, J.W. (1994) In vitro genetic analysis of the hinge region between helical elements P5-P4-P6 and P7-P3-P8 in the sunY group I self-splicing intron. *J Mol Biol*, **235**, 140-155.
35. Doudna, J.A. and Cech, T.R. (1995) Self-assembly of a group I intron active site from its component tertiary structural domains. *RNA*, **1**, 36-45.
36. Onuchic, J.N., Luthey-Schulten, Z. and Wolynes, P.G. (1997) Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem*, **48**, 545-600.
37. Fiore, J.L., Kraemer, B., Koberling, F., Edmann, R. and Nesbitt, D.J. (2009) Enthalpy-driven RNA folding: single-molecule thermodynamics of tetraloop-receptor tertiary interaction. *Biochemistry*, **48**, 2550-2558.

Chapter 7: Conclusion

7.1 General conclusions

In this dissertation we have comprehensively explored how topological constraints encoded by the connectivity and steric properties of RNA secondary structure govern RNA 3D conformation. Prior studies supplied strong evidence that topological constraints limit the 3D conformations accessible to two-way junctions (1-3), and hypothesized that these constraints may in turn provide a source of folding and dynamics specificity (1,3). Nevertheless, the full role of these constraints in determining two-way junction behavior remained unclear. More importantly, the role of topological constraints in higher-order junctions and large RNAs was unknown. Using a combination of heuristic models and the specially designed TOPRNA coarse-grained model, and studies of bulges, tRNAs, and the *Azoarcus* ribozyme, we have conclusively demonstrated that topological constraints are fundamental to RNA folding and dynamics.

Considered on an individual junction basis, the studies reported in **Chapters 2-6** characterize nearly every important class of junction found in functional RNAs. Our analyses uniformly indicate that topological constraints strongly limit junction conformation: for all but the longest bulges, two-way junctions are limited to <25% of the total interhelical space; different variants of the tRNA four-way junction are limited to <9%; a five-way junction variant of tRNA is limited to <7%; and three-way and four-way pseudoknotted junctions found in the

Azoarcus ribozyme are limited to 0.2% and 4%. The only significant junction type not discussed here, a standard three-way junction, has been simulated in separate work with similar observed topological confinement (unpublished data, Mustoe, Al-Hashimi, and Brooks). As a consequence of this confinement, the conformations of even indirectly connected helices are coupled together. Remarkably, in the *Azoarcus* ribozyme, these couplings extend across multiple junctions, indicating that in large molecules, topological constraints can restrict 3D conformation even beyond that expected based on individual junctions. Thus, taken as whole, these results strongly argue that topological constraints play a universal role in limiting RNA 3D conformation.

More importantly, our simulations demonstrate for the first time that topological constraints make significant contributions to the free energy of different allowed junction conformations. In two-way junctions, the ΔG^{topo} varied by as much as 6 kcal/mol between different interhelical conformations. In tRNA and the *Azoarcus* ribozyme, the ΔG^{topo} of forming different tertiary contacts varied by as much as 2-5 kcal/mol. These values are of similar magnitude to those obtained by Chu *et al.* (1) in their studies of PEG-linked DNA helices, and as noted in the introduction, are expected from basic polymer physics principles (4).

Experimental results in **Chapters 3** and **5** reveal that these topological constraint free energies in turn have a profound effect on RNA tertiary folding and dynamics. Experimental observables averaged over our TOPRNA-generated free energy landscapes reproduce measurements made on dynamic polypyrimidine bulges to nearly within experimental error. Bulges in the PDB also adopt conformations that closely coincide with the topological constraint free energy landscape. In **Chapter 5**, differences in topological constraints destabilized the insN and insNN mutants of mt-tRNA$^{Ser(UCN)}$ by 0.6 and 1.2 kcal/mol, approximately within experimental error of the destabilization observed by UV melting. Together, these results provide

compelling support for the hypothesis that topological constraints comprise an integral component of the RNA free energy landscape.

Strikingly, in tRNA and the *Azoarcus* ribozyme, the differential free energies encoded by topological constraints provide a robust, sequence-independent source of tertiary folding specificity by penalizing formation of non-native tertiary contacts. As noted in **Chapter 1**, the basis of tertiary folding specificity has long been poorly understood – most tertiary interactions are inherently non-specific, and motifs that are sequence-specific are often reused multiple times in the same molecule. The free energy penalties we observe between different contacts – between 1 and 2 kcal/mol – are similar in magnitude to the strength of individual tertiary interactions and thus are likely to be highly significant in determining the RNA folding outcome. For tRNA, these differences allow us to predict global 3D structure based purely on maximizing inter-loop packing. While the low probability of sampling compact conformations precluded us from predicting the *Azoarcus* ribozyme 3D structure, we suspect that a similar prediction would be possible with greater sampling. These results rationalize the observation that compact folding intermediates adopt native-like folds despite having incompletely-formed tertiary interactions (5), and why crowding agents that favor compact folds specifically stabilize the native structure (6,7). Thus, bearing out the hypothesis posed by Herschlag and coworkers (1), our data indicate that topological constraints are central to how RNAs overcome their limited nucleotide alphabet to fold to unique 3D structures.

Topological constraints are also responsible for the cooperativity and hierarchical nature of RNA tertiary folding. In both tRNA and the *Azoarcus* ribozyme, tertiary contacts formed in one part of the molecule are transduced by topological constraints into a reduction of conformational freedom of distal helices, lowering the entropic cost of forming additional

tertiary interactions by as much as -2 kcal/mol. These values are within experimental error of the tertiary interaction cooperativity measured in the *Azoarcus* ribozyme (8). Coupled with the large entropic penalty of forming individual tertiary interactions in the *Azoarcus* ribozyme, this cooperativity produces a natural folding hierarchy where less costly short-range interactions form first and promote subsequent formation of long-range interactions. Such hierarchical folding may further prevent formation of non-native tertiary interactions that could pose kinetic barriers to native folding in large RNAs.

It is important to acknowledge that our conclusions represent a greatly simplified view of RNA tertiary folding. This thesis explicitly ignored driving forces such as electrostatics and attractive interactions, which have been overwhelmingly shown to be critical to RNA folding (5,9,10). Recent studies have also made clear that our assumption that native secondary structure is always preformed during tertiary folding is an oversimplification (11,12). A full understanding of RNA tertiary folding requires consideration of these and other complexities of the RNA free energy landscape.

Nevertheless, the ability of topological constraints to encode folding specificity, cooperativity, and hierarchy implicates them as a fundamental determinant of RNA folding and dynamics. Significantly, as shown in **Chapters 2-5**, topological constraints and their resulting contributions to the folding landscape are highly tunable, requiring only small changes to the length of interhelical linkers. Thus, topological constraints can be readily optimized by evolution to achieve specific functional goals, and, *vice versa*, there is strong evolutionary pressure on RNAs to conserve secondary structures that have optimal topological constraints. This is conclusively illustrated by mt-tRNA$^{Ser(UCN)}$, where disruption of topological constraints results in

human disease. Overall, we believe that topological constraints will prove to be a crucial piece in solving the RNA free energy landscape puzzle.

7.2 Future outlook

The highly predictive framework offered by topological constraints offers many immediately promising areas for future research. The most obvious extension is to test whether topological constraints play a similarly important role in the tertiary folding of larger RNAs such as the *Tetrahymena* group I ribozyme and the group II ribozyme. These RNAs have long served as paradigm systems for understanding tertiary folding (5), and exploring the applicability of topological constraints in these RNAs would help establish the universality or reveal limitations of the mechanisms uncovered herein.

An intriguing result from our studies of the *Azoarcus* ribozyme is that topological constraints can give rise to hierarchical folding behavior. As RNAs grow in size, hierarchical folding mechanisms become universal. The most important example of this is the ribosome, where folding is tightly controlled by the hierarchical binding of proteins that induce conformational changes in the ribosomal RNA (13-16). The physical basis of how this hierarchy is enforced is poorly understood. In **Chapter 3**, we found that proteins could stabilize two-way junctions in conformations with high topological constraint free energies. Interestingly, preliminary analysis of these junction/protein complexes revealed that the majority of the complexes are from the ribosome, and when mapped to the ribosomal assembly map (14), the participating protein bound early in the hierarchical assembly pathway. Testing whether topological constraints contribute to the hierarchical assembly mechanism of the ribosome should be an exciting future topic of study.

A related deeper question is why such hierarchical assembly mechanisms are necessary. It is known that assembly proteins play an important role in neutralizing the negative charge of the RNA during hierarchical assembly of the ribosome (17). However, it may also be the case that hierarchical folding mechanisms are necessary because topological constraints are unable to effectively discriminate against non-native interactions in large RNAs.

The potential of topological constraints to encode tertiary folding specificity (at least in small RNAs) also offers new avenues for RNA 3D structure prediction. Given the enormous advances made in determining secondary structure (18), much of the field has pivoted to using secondary structure as a starting point for 3D structure prediction. However, the nonspecific nature of tertiary interactions continues to pose a significant challenge for structure prediction algorithms. Accounting for the penalties posed by topological constraints on forming different interactions may be a solution. In **Chapters 4-5**, we were able to predict tRNA 3D structure from these penalties alone, an impressive feat given that tRNA remains a difficult structure to predict for state-of-the-art methods (19). Our failure to predict the *Azoarcus* ribozyme structure indicates that future efforts to use TOPRNA as a structure-prediction method should incorporate an interhelical stacking bias and perhaps a radius of gyration restraint to increase sampling of compact conformations. Nonetheless, the *Azoarcus* ribozyme simulations did reveal the individual specificities of the TL/TLR and other tertiary interactions. Incorporating such information into preexisting methods is likely to prove fruitful (19-26). Considering conformational biases arising from topological constraints may also aid future efforts at predicting the 3D structure of long non-coding RNAs and other 'non-structural' RNAs, which, despite lacking clear stabilizing tertiary interactions, can adopt functional well-defined 3D structures (27).

While RNA dynamics were a secondary focus of this thesis, they are critical to many RNA biological processes, as discussed in **Chapter 1**. Compared to RNA folding, RNA dynamics are still much less understood. As we demonstrated in **Chapters 3-4**, topological constraints appear to play a primary role in dictating dynamics specificity, and for bulges in the low-salt limit, the overall free energy landscape. Thus, considering topological constraints is likely to be important for future efforts at understanding both the physics and function of RNA dynamics. An area of particular current interest is constructing dynamic RNA ensembles that can be used as a target for small molecule virtual screening (28). Topological constraint free energies could potentially be used to guide the generation of such ensembles to ensure that only low-energy RNA conformations are being screened against.

Beyond the above applications, several fundamental questions regarding the nature of topological constraints remain. In particular, we explicitly ignored electrostatics throughout this work. However, repulsion due to the negative charge of the RNA backbone is just as much an inherent constraint on RNA 3D conformation as topological constraints. While theoretical studies of model helical junctions suggest that electrostatic repulsions are significantly screened at the high salt concentrations where RNAs fold (29,30), it is likely that electrostatics will still disfavor particularly close packing of helices. It may be the case that electrostatics act in synergy with topological constraints to further discriminate against non-native tertiary conformations. Future studies may also explore how inclusion of atomic-level accuracy modifies the topological constraint free energy landscape; we recently developed an enhanced sampling method that uses (α_h, β_h, γ_h) angles as order parameters that could facilitate such studies (31). Lastly, as noted earlier, secondary structure is not always 'prefolded', as we assumed in our studies. Indeed, many studies have shown that tertiary interactions can form concomitantly with native secondary

structure, and that topological constraints posed by newly formed helices can bias downstream secondary structure folding steps (32-35). Studies of topological constraints in the context of these complexities are needed.

Finally, our studies of tRNA in **Chapters 4-5** suggest that topological constraints underlie secondary structure conservation. While secondary structure has long been appreciated as the most highly conserved aspect of RNA, to our knowledge we are the first to provide a rationalization for this conservation. Exploring the link between topological constraints and secondary structure conservation in other RNAs may uncover new principles of RNA folding and function. In particular, many non-coding RNAs appear to lack traditional signatures of secondary structure conservation (36); it may be that these RNAs conserve topological constraints through variable secondary structure architectures. Better understanding the link between secondary structure and 3D folding is also likely to aid efforts at designing novel RNA folds. There has been speculation that the ability of RNAs to remain functional as long as secondary structure is conserved – and thus tolerate a high degree of primary sequence drift – was central to the viability of the RNA-world (37). In this respect, topological constraints may be at the very center of RNA's ancient role as the progenitor of life.

7.3 References

1. Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S. and Herschlag, D. (2009) Do conformational biases of simple helical junctions influence RNA folding stability and specificity? *RNA*, **15**, 2195-2205.
2. Bailor, M.H., Sun, X.Y. and Al-Hashimi, H.M. (2010) Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation. *Science*, **327**, 202-206.
3. Sim, A.Y.L. and Levitt, M. (2011) Clustering to identify RNA conformations constrained by secondary structure. *Proc Natl Acad Sci USA*, **108**, 3590-3595.
4. Richards, E.G. (1980) *An introduction to the physical properties of large molecules in solution*. Cambridge University Press, New York.

5. Woodson, S.A. (2010) Compact intermediates in RNA folding. *Annu Rev Biophys*, **39**, 61-77.
6. Kilburn, D., Roh, J.H., Behrouzi, R., Briber, R.M. and Woodson, S.A. (2013) Crowders perturb the entropy of RNA energy landscapes to favor folding. *J Am Chem Soc*, **135**, 10055-10063.
7. Denesyuk, N.A. and Thirumalai, D. (2011) Crowding promotes the switch from hairpin to pseudoknot conformation in human telomerase RNA. *J Am Chem Soc*, **133**, 11858-11861.
8. Behrouzi, R., Roh, J.H., Kilburn, D., Briber, R.M. and Woodson, S.A. (2012) Cooperative tertiary interaction network guides RNA folding. *Cell*, **149**, 348-357.
9. Tinoco, I., Jr. and Bustamante, C. (1999) How RNA folds. *J Mol Biol*, **293**, 271-281.
10. Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct*, **26**, 113-137.
11. Koculi, E., Cho, S.S., Desai, R., Thirumalai, D. and Woodson, S.A. (2012) Folding path of P5abc RNA involves direct coupling of secondary and tertiary structures. *Nucleic Acids Res*, **40**, 8011-8020.
12. Wu, M. and Tinoco, I., Jr. (1998) RNA folding causes secondary structure rearrangement. *Proc Natl Acad Sci USA*, **95**, 11555-11560.
13. Kim, H., Abeysirigunawarden, S.C., Chen, K., Mayerle, M., Ragunathan, K., Luthey-Schulten, Z., Ha, T. and Woodson, S.A. (2014) Protein-guided RNA dynamics during early ribosome assembly. *Nature*, **506**, 334-338.
14. Mizushima, S. and Nomura, M. (1970) Assembly mapping of 30S ribosomal proteins from E. coli. *Nature*, **226**, 1214.
15. Mulder, A.M., Yoshioka, C., Beck, A.H., Bunner, A.E., Milligan, R.A., Potter, C.S., Carragher, B. and Williamson, J.R. (2010) Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. *Science*, **330**, 673-677.
16. Stagg, S.M., Mears, J.A. and Harvey, S.C. (2003) A structural model for the assembly of the 30S subunit of the ribosome. *J Mol Biol*, **328**, 49-61.
17. Trylska, J., McCammon, J.A. and Brooks, C.L. III (2005) Exploring assembly energetics of the 30S ribosomal subunit using an implicit solvent approach. *J Am Chem Soc*, **127**, 11125-11133.
18. Hajdin, C.E., Bellaousov, S., Huggins, W., Leonard, C.W., Mathews, D.H. and Weeks, K.M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci USA*, **110**, 5498-5503.
19. Ding, F., Lavender, C.A., Weeks, K.M. and Dokholyan, N.V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods*, **9**, 603-608.
20. Lavender, C.A., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2010) Robust and Generic RNA Modeling Using Inferred Constraints: A Structure for the Hepatitis C Virus IRES Pseudoknot Domain. *Biochemistry*, **49**, 4931-4933.
21. Seetin, M.G. and Mathews, D.H. (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J Comput Chem*, **32**, 2232-2244.
22. Flores, S.C. and Altman, R.B. (2010) Turning limited experimental information into 3D models of RNA. *RNA*, **16**, 1769-1778.
23. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189-199.

24. Das, R., Kudaravalli, M., Jonikas, M., Laederach, A., Fong, R., Schwans, J.P., Baker, D., Piccirilli, J.A., Altman, R.B. and Herschlag, D. (2008) Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci USA*, **105**, 4144-4149.
25. Gherghe, C.M., Leonard, C.W., Ding, F., Dokholyan, N.V. and Weeks, K.M. (2009) Native-like RNA Tertiary Structures Using a Sequence-Encoded Cleavage Agent and Refinement by Discrete Molecular Dynamics. *J Am Chem Soc*, **131**, 2541-2546.
26. Yang, S.C., Parisien, M., Major, F. and Roux, B. (2010) RNA Structure Determination Using SAXS Data. *J Phys Chem B*, **114**, 10039-10048.
27. Fang, X., Wang, J., O'Carroll, I.P., Mitchell, M., Zuo, X., Wang, Y., Yu, P., Liu, Y., Rausch, J.W., Dyba, M.A. *et al.* (2013) An unusual topological structure of the HIV-1 Rev response element. *Cell*, **155**, 594-605.
28. Stelzer, A.C., Frank, A.T., Kratz, J.D., Swanson, M.D., Gonzalez-Hernandez, M.J., Lee, J., Andricioaei, I., Markovitz, D.M. and Al-Hashimi, H.M. (2011) Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nat Chem Biol*, **7**, 553-559.
29. Tan, Z.J. and Chen, S.J. (2012) Ion-mediated RNA structural collapse: effect of spatial confinement. *Biophys J*, **103**, 827-836.
30. Bai, Y., Chu, V.B., Lipfert, J., Pande, V.S., Herschlag, D. and Doniach, S. (2008) Critical assessment of nucleic acid electrostatics via experimental and computational investigation of an unfolded state ensemble. *J Am Chem Soc*, **130**, 12334-12341.
31. Dickson, A., Mustoe, A.M., Salmon, L. and Brooks, C.L., III (2014) Fast molecular dynamics exploration of RNA interhelical conformation using Euler angles and WExplore. *Nucleic Acids Res,* in press.
32. Cho, S.S., Pincus, D.L. and Thirumalai, D. (2009) Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *Proc Natl Acad Sci USA*, **106**, 17349-17354.
33. Lin, J.C., Hyeon, C. and Thirumalai, D. (2012) RNA under tension: Folding Landscapes, Kinetic partitioning Mechanism, and Molecular Tensegrity. *J Phys Chem Lett*, **3**, 3616-3625.
34. Feng, J., Walter, N.G. and Brooks, C.L., III. (2011) Cooperative and directional folding of the preQ1 riboswitch aptamer domain. *J Am Chem Soc*, **133**, 4196-4199.
35. Lin, J.C. and Thirumalai, D. (2008) Relative stability of helices determines the folding landscape of adenine riboswitch aptamers. *J Am Chem Soc*, **130**, 14080-14081.
36. Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26-46.
37. Schultes, E.A. and Bartel, D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448-452.

Appendix A: Supporting Information for Chapter 2

A.1 Derivation of linker length as a function of (α_h, β_h, γ_h)

We can express the linker length for a bulge-like connected system by reducing each helix to two points, a pivot point located at the origin (representing the *Y* strand) and a linker terminus (representing the *X* strand) on the opposite side of the helix (**Figure A.4**). The linker termini are located a distance *D* from the origin, where *D* is the span of the helix diameter. In the initial state, both helices are coaxially stacked, so $Y_{H1} = Y_{H2} = (0,0,0)$ and $X_{H1} = X_{H2} = (D,0,0)$. However, in some ($\alpha_h$, β_h, γ_h) conformation $Y_{H1} = Y_{H2} = (0,0,0)$, $X_{H1} = (D,0,0)$, and $X_{H2} = R(-\gamma_h, -\beta_h, -\alpha_h) \times (D,0,0)^T$. For the given ($\alpha_h$, β_h, γ_h):

$$R(-\gamma_h,-\beta_h,-\alpha_h) = \begin{pmatrix} \cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) - \sin(\alpha_h)\sin(\gamma_h) & \cos(\alpha_h)\sin(\gamma_h) + \sin(\alpha_h)\cos(\beta_h)\cos(\gamma_h) & -\sin(\beta_h)\cos(\gamma_h) \\ -\cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h) - \sin(\alpha_h)\cos(\gamma_h) & \cos(\alpha_h)\cos(\gamma_h) - \sin(\alpha_h)\cos(\beta_h)\sin(\gamma_h) & \sin(\beta_h)\sin(\gamma_h) \\ \cos(\alpha_h)\sin(\beta_h) & \sin(\alpha_h)\sin(\beta_h) & \cos(\beta_h) \end{pmatrix} \quad \textbf{[A.1]}$$

And therefore

$$X_{H2} = R(-\gamma_h - \beta_h - \alpha_h) \times (D,0,0)^T \quad \textbf{[A.2]}$$

$$= D\begin{pmatrix} \cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) - \sin(\alpha_h)\sin(\gamma_h) \\ -\cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h) - \sin(\alpha_h)\cos(\gamma_h) \\ \cos(\alpha_h)\sin(\beta_h) \end{pmatrix} \quad \textbf{[A.3]}$$

From the coordinates of X_{H1} and X_{H2} we can calculate the distance, *L*, between the two linker termini:

$$L = \sqrt{\begin{aligned} &D^2\left(\cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) - \sin(\alpha_h)\sin(\gamma_h) - 1\right)^2 \\ &+D^2\left(-\cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h) - \sin(\alpha_h)\cos(\gamma_h)\right)^2 + D^2\left(\cos(\alpha_h)\sin(\beta_h)\right)^2 \end{aligned}} \quad \textbf{[A.4]}$$

Through an expansion of the squared terms and simplification this reduces to:

$$\frac{L^2}{D^2} = 2 - 2\cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) + 2\sin(\alpha_h)\sin(\gamma_h) \quad \textbf{[A.5]}$$

This can be written as a function *F*:

$$F = 1 - \cos(\alpha_h)\cos(\beta_h)\cos(\gamma_h) + \sin(\alpha_h)\sin(\gamma_h) - \frac{L^2}{2D^2} \quad \textbf{[A.6]}$$

with the total derivative of *F* with respect to an arbitrary time, *t*:

$$\frac{dF}{dt} = \frac{\partial F}{\partial \alpha_h}\frac{\partial \alpha_h}{\partial t} + \frac{\partial F}{\partial \beta_h}\frac{\partial \beta_h}{\partial t} + \frac{\partial F}{\partial \gamma_h}\frac{\partial \gamma_h}{\partial t} + \frac{\partial F}{\partial L}\frac{\partial L}{\partial t} \quad \textbf{[A.7]}$$

In the case where *L, D*, and β_h are held constant, then

$$\begin{aligned} 0 = &\left(\sin(\alpha_h)\cos(\beta_h)\cos(\gamma_h) + \cos(\alpha_h)\sin(\gamma_h)\right)\frac{\partial \alpha_h}{\partial t} \\ &+ \left(\cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h) + \sin(\alpha_h)\cos(\gamma_h)\right)\frac{\partial \gamma_h}{\partial t} \end{aligned} \quad \textbf{[A.8]}$$

and

$$\frac{d\alpha_h}{dt} = \frac{\sin(\alpha_h)\cos(\gamma_h) + \cos(\alpha_h)\cos(\beta_h)\sin(\gamma_h)}{-\cos(\alpha_h)\sin(\gamma_h) - \sin(\alpha_h)\cos(\beta_h)\cos(\gamma_h)}\frac{d\gamma_h}{dt} \quad \textbf{[A.9]}$$

The singularities that occur in **[A.9]** correspond to *L*=0 or *L*=2*D,* where the linker is maximally or minimally extended and the total derivative of the function is 0. In the coaxial limit of $\beta_h \rightarrow 0°$, $\frac{d\alpha_h}{dt} = \frac{d\gamma_h}{dt}$ and changes in α_h, which denotes a clockwise twist of the 3' helix, must be exactly offset by negative changes in γ_h, or clockwise rotations of the 5' helix. Similarly, as $\beta_h \rightarrow 180°$, $\frac{d\alpha_h}{dt} = -\frac{d\gamma_h}{dt}$. The maximum allowable α_h and γ_h values for a given linker distance

are achieved when $\beta_h \rightarrow 0°$ or $\beta_h \rightarrow 180°$, causing the topologically allowed α_h-γ_h projection to appear as a superposition of these limits with the expected positive and negative α_h-γ_h correlations (Fig 2A)**.** While the specific cases of $\beta_h=0°$ and $\beta_h=180°$ are locations of degeneracies in the Euler space (1), the approximately linear relationships between $\frac{d\alpha_h}{dt}$ and $\frac{d\gamma_h}{dt}$ hold for β_h near these values. As β_h diverges from 0°, α_h and γ_h remain correlated for fixed β_h and linker length; however the proportionality constant between the differentials is no longer unity. Eventually, as $\beta_h \rightarrow 90°$, the α_h-γ_h relationship breaks down. When $\beta_h=90°$ and α_h or $\gamma_h=0°$, ±90°, or ±180°, changes in α_h are completely uncorrelated to changes in γ_h.

A.2 Supporting tables and figures

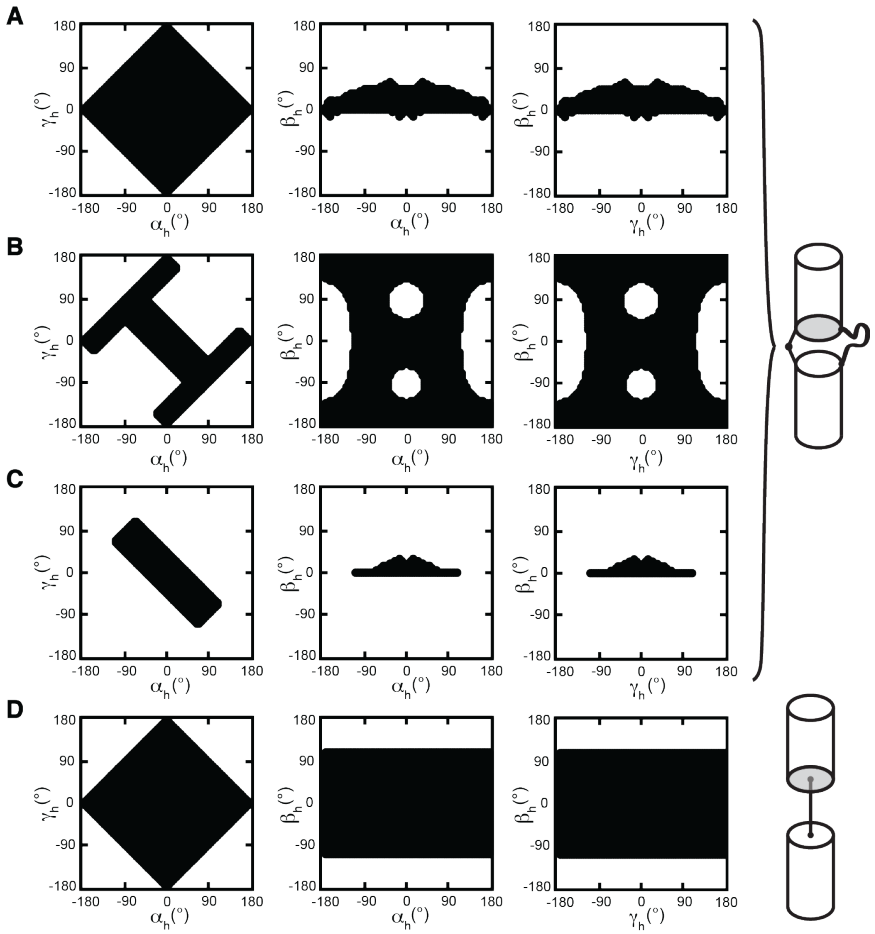

Figure A.1: (α_h, β_h, γ_h) projections of model cylinder systems
(A) Sterically allowed, **(B)** connectivity allowed, and **(C)** the union of sterically and connectivity allowed conformations for a bulge-like connected cylinder system with a 14.5 Å linker (equivalent to 2-nt bulge). **(D)** The set of allowed conformations satisfying the steric constraints for a center connected cylinder system with a 14.52 Å connector length. There are no connectivity constraints as the linker-length is invariant across the rotations.

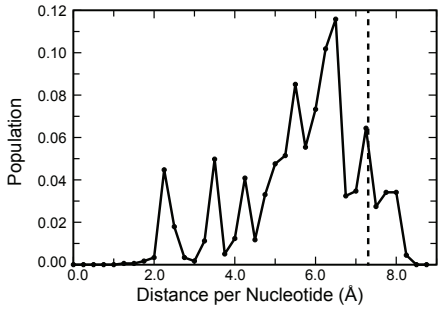

Figure A.2: The per-nucleotide O3'-P atomic distance in PDB junctions
The distribution was computed from 1787 $H_3S_{4\geq X\geq 1}H_3S_0$ crystal structure bulges indentified with the RNA FRABASE (2). The dashed line denotes the chosen linker cutoff of 7.26 Å.

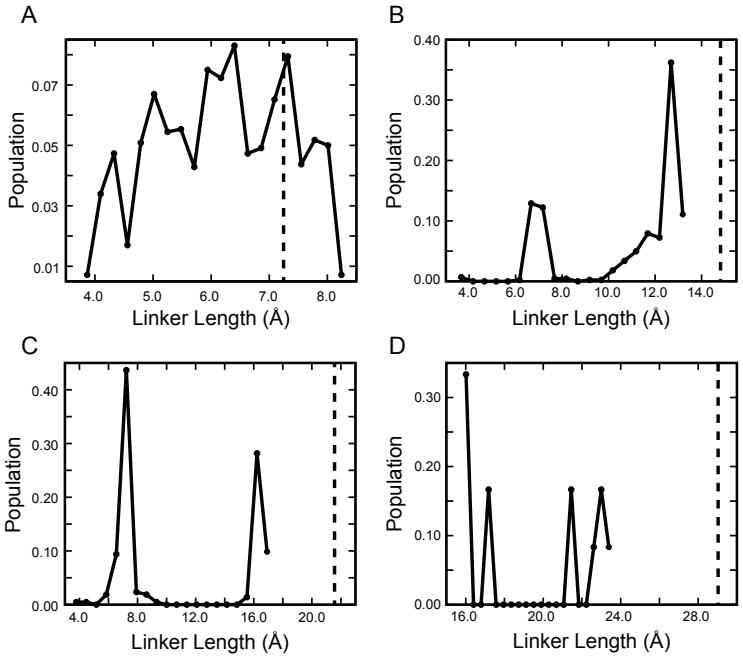

Figure A.3: Bulge linker distance distributions for junctions in the PDB
(A)1120 S_1S_0, **(B)** 442 S_2S_0, **(C)** 213 S_3S_0, and **(D)** 12 S_4S_0 crystallographic junctions identified with the RNA FRABASE (2). The linker cutoff of 7.26 Å per bulged nucleotide used in our calculations is marked with a dashed vertical line at 7.26 Å, 14.52 Å, 21.78 Å, and 29.04 Å for S_1S_0, S_2S_0, S_3S_0, and S_4S_0 bulges respectively.

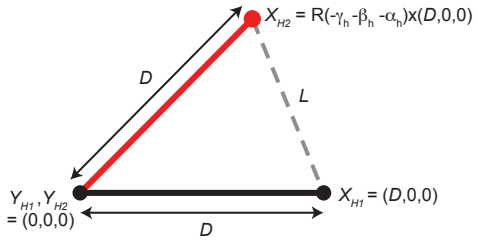

Figure A.4: Diagram of linker length dependence on (α_h, β_h, γ_h)
L is the linker length and *D* is the helix diameter.

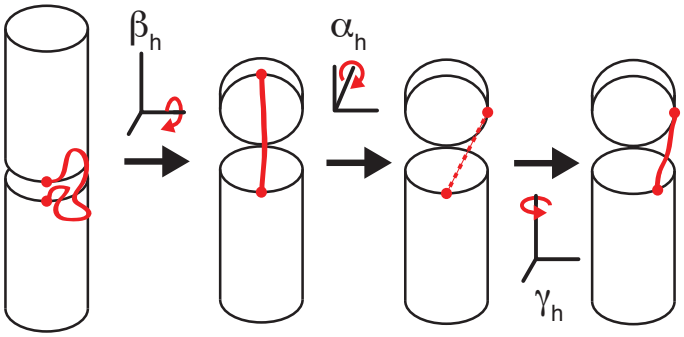

Figure A.5: Illustration of α_h-γ_h correlations due to connectivity constraints
Shown is a scenario where a rotation of the upper helix about its helical axis by α_h over-stretches the linker. This must be offset by a corresponding rotation of the lower helix around its axis by γ_h.

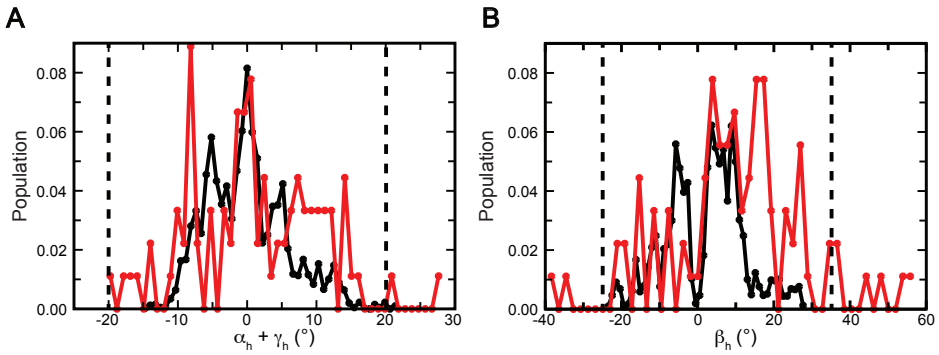

Figure A.6: Distribution of interhelical twist and bend in contiguous PDB helices
(A) Interhelical twist (α_h+γ_h) and **(B)** bend β_h observed for 4062 x-ray (black) and 90 NMR (red) $H_6S_0H_6S_0$ structures identified with the RNA FRABSE (2). Inter-helical angles were measured by splitting the contiguous, 6 base pair helices. Vertical dashed lines denote the cutoff values used in the construction of the topologically allowed space.

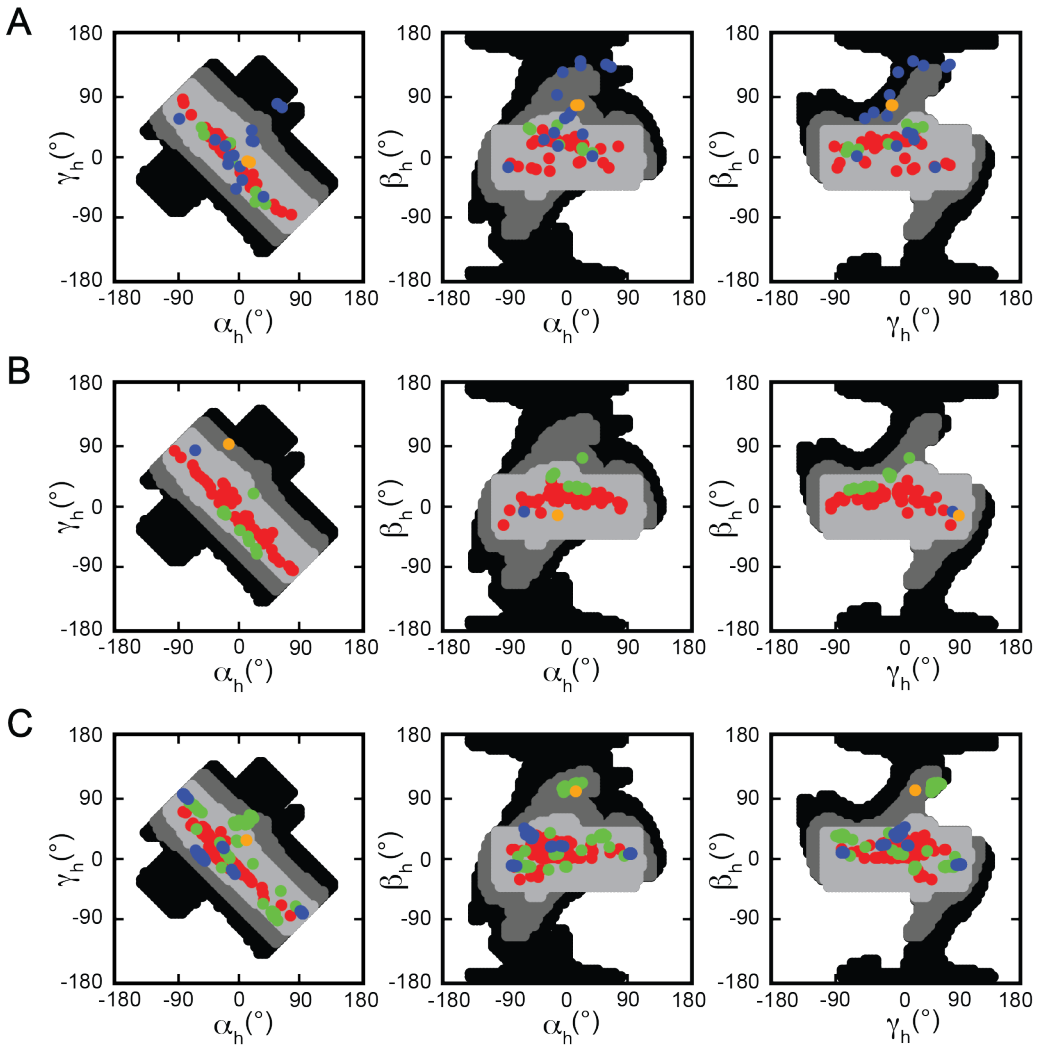

Figure A.7: (α_h, β_h, γ_h) angles of GU-containing, NMR, or cryo-EM PDB junctions
Interhelical angles measured for **(A)** NMR, **(B)** cryo-EM, or **(C)** GU closing base pair junctions. The total topologically allowed space is plotted for the connectivity constraints of 1 (light grey), 2 (dark grey), or 3 (black) bulged nucleotides, where allowed conformations for *i* number of bulged nucleotides are inclusive of those for *i-1*, *i-2*, *...1* bulged nucleotides. (α_h, β_h, γ_h) angles measured for S_1S_0, S_2S_0, S_3S_0, and S_4S_0 junctions are plotted in red, green, blue, and orange respectively. Shown are 34 $H_3S_1H_3S_0$, 7 $H_3S_2H_3S_0$, 15 $H_3S_3H_3S_0$, and 2 $H_3S_4H_3S_0$ NMR junctions; 59 $H_3S_1H_3S_0$, 12 $H_3S_2H_3S_0$, 1 $H_3S_3H_3S_0$, and 1 $H_3S_4H_3S_0$ EM junctions; and 432 (392 X-ray, 31 EM, 9 NMR) $H_3S_1H_3S_0$, 308 (282 X-ray, 25 EM, 1 NMR) $H_3S_2H_3S_0$, 157 (148 X-ray, 9 EM, 0 NMR) $H_3S_3H_3S_0$, and 2 (2 X-ray, 0 EM, 0 NMR) $H_3S_4H_3S_0$ GU closing base pair containing junctions.

PDB	Residues	Sequence	5' Helix RMSD	3' Helix RMSD
1QC0	C.102:109	CAC**CG**UUG	0.84	0.70
	D.130:137	GUG**GC**AAC		
1QC0	C.103:110	ACC**GU**UGG	0.90	0.74
	D.129:136	UGG**CA**ACC		
1QC0	C.104:111	CCG**UU**GGU	0.77	0.87
	D.128:135	GGC**AA**CCA		
1QC0	C.105:112	CGU**UG**GUA	0.73	0.90
	D.127:134	GCA**AC**CAU		
1QC0	C.108:115	UGG**UA**GCG	0.87	0.80
	D.124:131	ACC**AU**CGC		
1QC0	C.109:116	GGU**AG**CGG	0.90	0.73
	D.123:130	CCA**UC**GCC		
1QC0	C.110:117	GUA**GC**GGU	0.78	0.68
	D.122:129	CAU**CG**CCA		
1QCU	A.2:9	GGG**GG**GGG	0.56	0.64
	B.14:21	CCC**CC**CCC		
1QCU	C.101:108	CCC**CC**CCC	0.56	0.78
	D.115:122	GGG**GG**GGG		
3DIL	A.2:9	GCC**GA**CGG	0.86	1.21
	A.166:173	UGG**CU**GCC		
3DIL	A.85:92	GUG**CA**GAG	0.65	0.71
	A.102:109	UAC**GU**CUC		
3DIL	A.117:124	GGU**AU**GGG	0.72	0.93
	A.130:137	CCA**UA**CCC		
1Z43	A.155:162	CAU**CU**CCU	1.14	1.14
	A.167:174	GUA**GA**GGG		
1Y26	X.14:21	GCU**UC**AUA	0.70	0.80
	X.75:82	UGA**AG**UAU		
3G78	A.356:363	ACC**GC**CGU	1.42	0.85
	A.380:387	UGG**UG**GCA		
433D	A.5:12	UUG**CG**GUA	0.94	1.03
	B.17:24	GGC**GU**UAU		
1HR2	A.228:235	UCA**AC**AGA	0.49	0.94
	A.239:246	AGU**UG**UCU		
2GCS	A.10:17	CUU**AA**AGC	0.79	0.73
	B.24:31	GAA**UU**UCG		

Table A.1: Helices used to computed the non-idealized topologically allowed space
Base pairs at the site of the 'junction' are bolded.

A.3 References

1. Bailor, M.H., Mustoe, A.M., Brooks, C.L., III and Al-Hashimi, H.M. (2011) 3D maps of RNA inter-helical junctions. *Nat Protoc*, **6**, 1536-1545.
2. Popenda, M., Blazewicz, M., Szachniuk, M. and Adamiak, R.W. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res*, **36**, D386-391.

Appendix B: Supporting Information for Chapter 3

B.1 Supporting tables and figures

ID	Sequence
1	5'GGU**B**CCG CCA GGC5'
2	5'CAU**B**GCG GUA CGC5'
3	5'AGC**B**UUU UCG GAA5'
4	5'AGA**B**AUC UCU UAG5'
5	5'UUG**B**UCA AAU AGU5'
6	5'ACU**B**GUG UGA CAC5'
7	5'GGC**B**CGU CCG GCA5'
8	5'UGU**B**CCA ACA GGU5'
9	5'AGC**B**CCG UCG GGC5'
10	5'CGA**B**GCC GCU UGG5'

Table B.1: Sequences of TOPRNA simulated two-helix bulges
Bolded 'B's are used to indicate where the 'bulged' single stranded nucleotides were inserted into the sequences, with the inserted bulges consisting of randomized sequences of lengths varying from 1 to 7 residues.

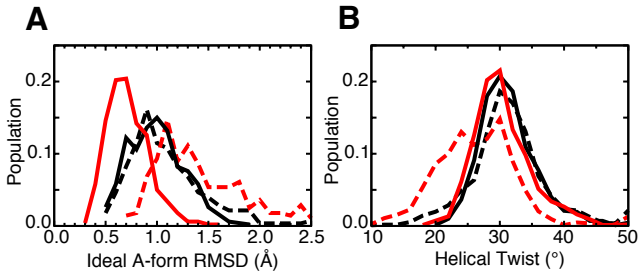

Figure B.1: Comparison of TOPRNA and NMR-MD HIV-2 TAR helices
The distribution of helical parameters measured from TOPRNA simulations (solid lines) and NMR-MD dynamic ensemble of HIV-2 TAR (dashed lines). Parameters were measured for the first five base pairs of the lower helix (black) and the four base pairs of the upper helix (red), and the populations of the helical twist parameters represent collections over both the different conformations and constituent base-pair steps. The divergence between the NMR-MD and TOPRNA helical parameters for the upper helix is due to a transient secondary structure change in NMR-MD ensemble.

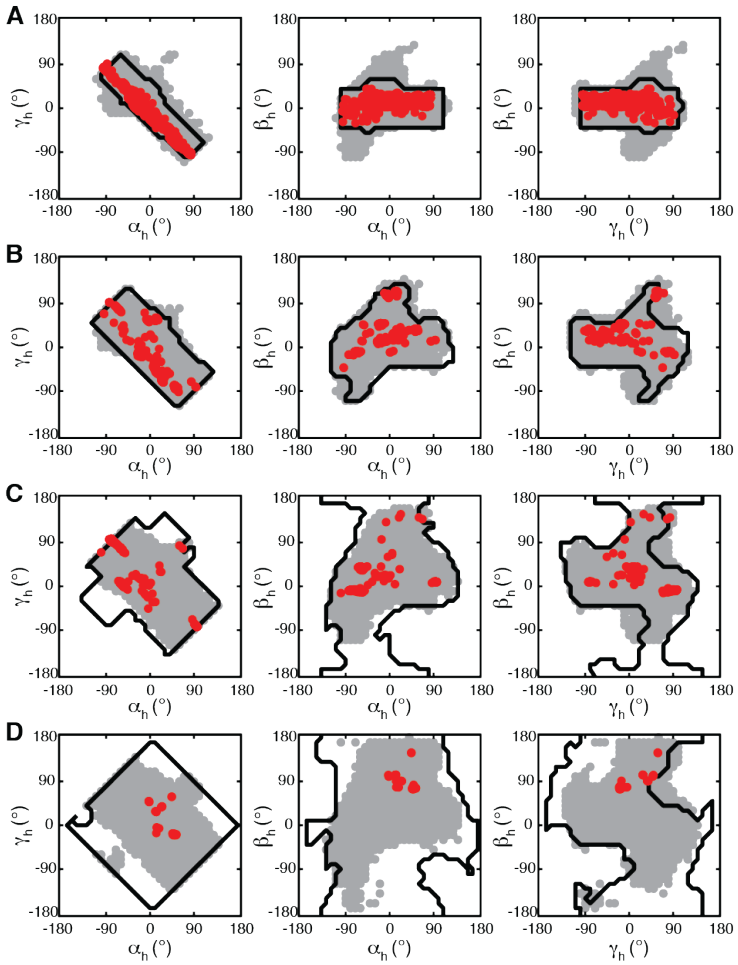

Figure B.2: Interhelical angles sampled TOPRNA bulge simulations
Angles sampled by TOPRNA are shown in gray, measured in the PDB shown in red, and predicted to be allowed by the heuristic model shown in black for **(A)**1-nt, **(B)** 2-nt, **(C)** 3-nt, and **(D)** 4-nt bulges.

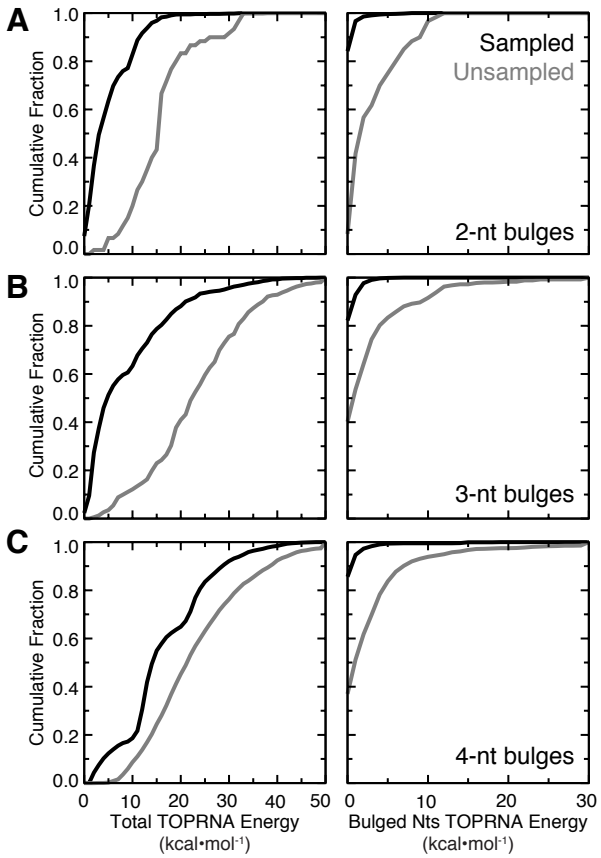

Figure B.3: Energy distributions of 'TOPRNA-unsampled' bulge conformations
TOPRNA force-field energies of interhelical conformations predicted to be allowed by the heuristic model that are sampled (black) and unsampled (gray) by TOPRNA simulations for **(A)** 2-nt, **(B)** 3-nt, **(C)** and 4-nt bulge systems. On the left are distributions of the total system energy and on the right distributions of the bulge-comprising nucleotides' energy. Energies were obtained after minimization of a TOPRNA bulge to a scaffold possessing the given (α_h, β_h, γ_h) conformation (see Chapter 3 methods).

Appendix C: Supporting Information for Chapter 4

C.1 Supporting tables and figures

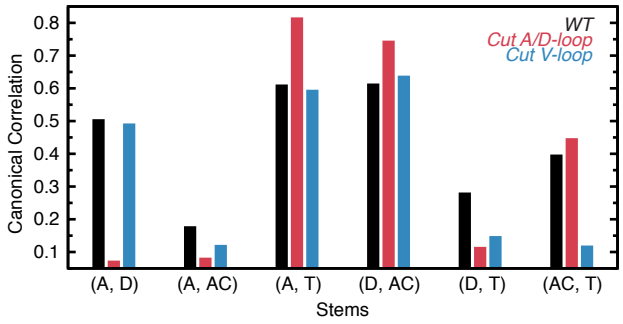

Figure C.1: Canonical correlations between the centers of mass of tRNA helices
Correlations were measured as described in Chapter 4. The increased correlations between the (A, T), (D, AC), and (AC, T) stems of the cut tRNAs arise because global translations and rotations of these molecules were less effectively removed by reference helix superpositions. These residual global motions inflate the canonical correlation measured between helices. Note that these increased correlations are not observed by mutual information measures (**Figure 4.2**).

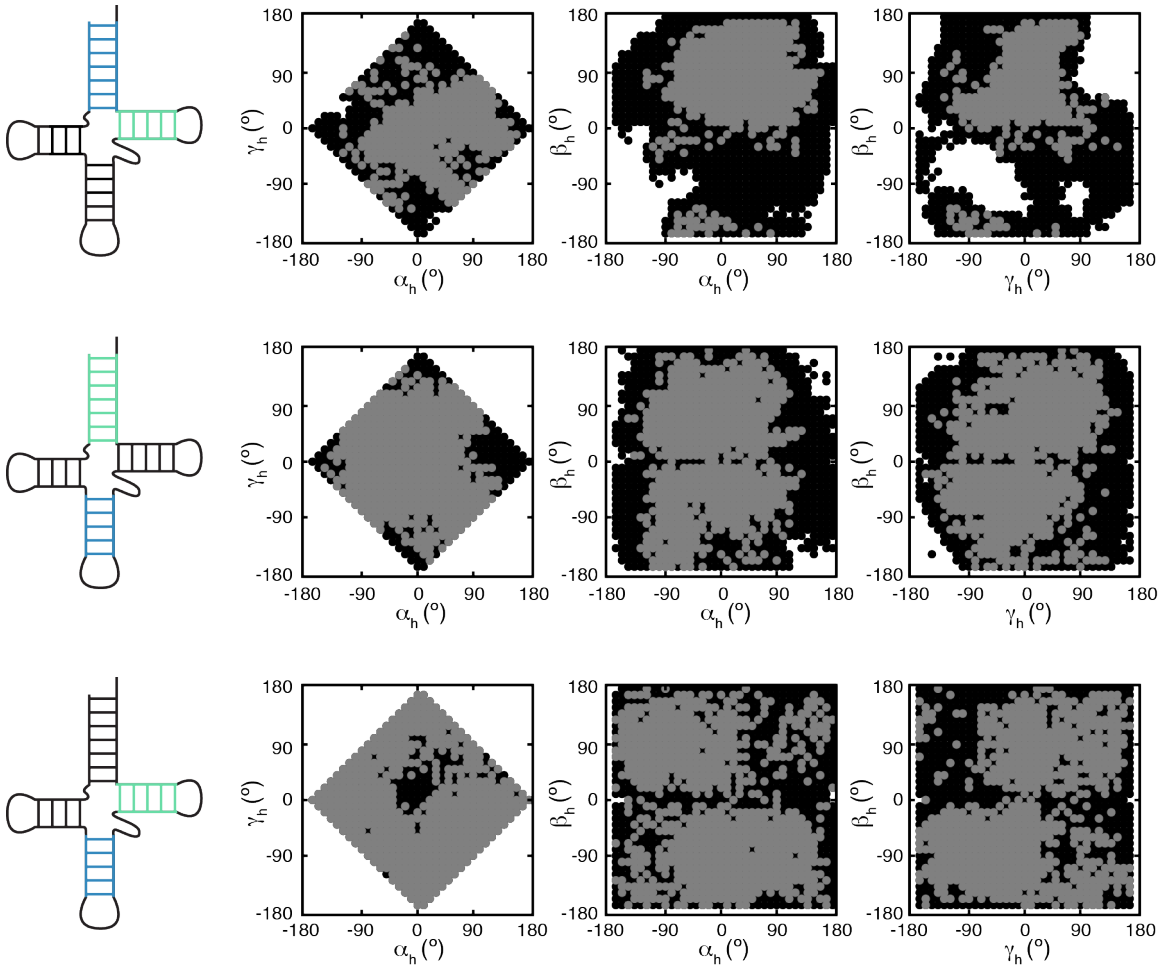

Figure C.2: D-to-AC-stem stacking constrains other tRNA helices
2D projections of the (α_h, β_h, γ_h) angles sampled between stems of WT tRNA when the D- and AC-stems are coaxially stacked (gray points), superimposed on top of all (α_h, β_h, γ_h) angles sampled by unrestrained WT tRNA (black points). Thus, black points denote otherwise accessible regions of (α_h, β_h, γ_h) space that are inaccessible when the D- and AC-stems are stacked. Cartoons on the left indicate the plotted inter-helical angle, with stems highlighted in blue denoting the H1 reference helix and stems highlighted in green denoting the H2 helix. Cutoffs of $|\beta_h|<20$ and $|\alpha_h + \gamma_h|<50$ between the D- and AC-stems were used to identify coaxially stacked conformations.

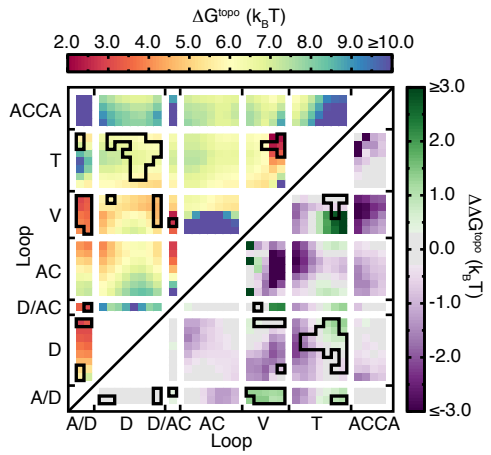

Figure C.3: The energetic cost of forming tertiary contacts in cut V-loop tRNA
The ΔG^{topo} is shown in upper triangle. The $\Delta\Delta G^{topo}$ between the cut V-loop relative to WT tRNA is shown in the lower triangle.

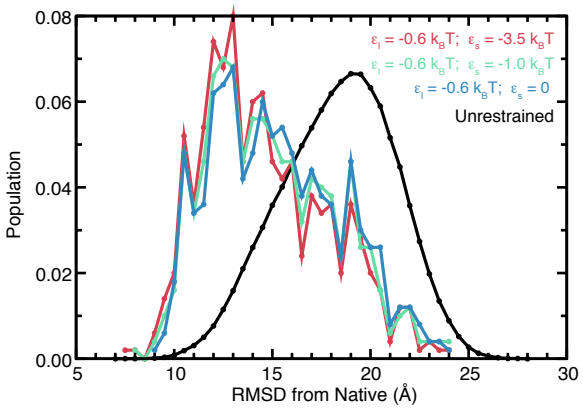

Figure C.4: RMSD histogram of 500-best packed conformers of tRNA
RMSDs were computed with respect to the 6TNA crystal structure using all-P atoms for all conformers sampled by the unrestrained simulation of WT tRNA (black), and of the 500 best-packed conformers identified using different values of ε_l and ε_s (colored; see key).

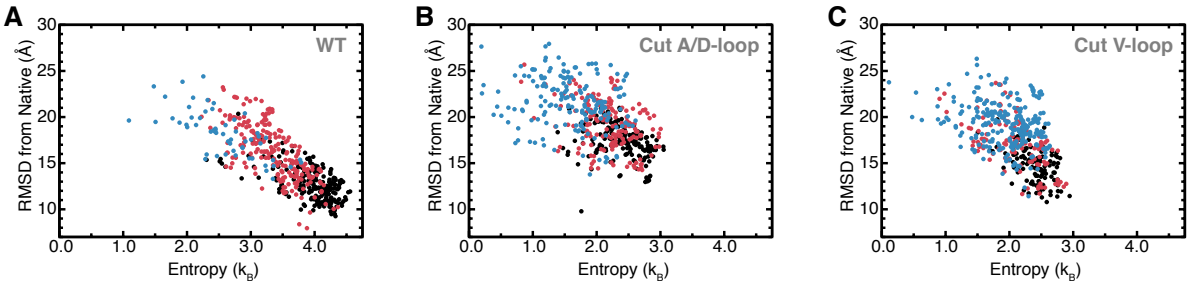

Figure C.5: Entropies of the 500 best-packed conformers of the cut V-Loop tRNA
Entropies were computed using ε_l=-0.6 k_BT and ε_s=-3.5 k_BT for **(A)** WT, **(B)** cut A/D-loop, and (**C**) cut V-loop tRNAs. Conformations that possess only native-consistent contacts and have contacts between the D- and T-loops are shown in black, conformations that possess only native-consistent contacts but do not have D-T loop contacts are shown in red, and conformations possessing native-inconsistent interactions are shown in blue (see main text methods). Note that **A** and **B** are reproduced from **Figure 4.3** for comparison.

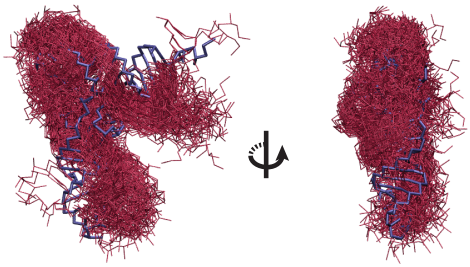

Figure C.6: High RMSD conformations sampled by the $tRNA_{9R}$ simulation
Shown in red are 50 randomly selected conformations from a total of 1249 conformations sampled by the $tRNA_{9R}$ simulation with >15 Å RMSD from the crystal structure, computed using all P atoms. These structures are superimposed on the 6TNA crystal structure, shown in blue.

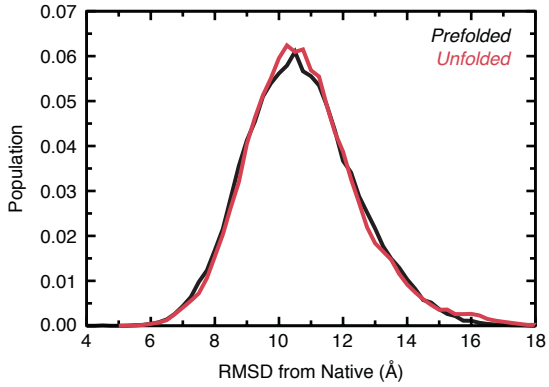

Figure C.7: RMSD of $tRNA_{4R}$ simulations initiated from unfolded coordinates
'Unfolded' initial coordinates were folded using a hierarchical folding scheme described in Chapter 4. The 'unfolded' curve shown in red represents the aggregate RMSD distribution computed from the production runs of the members of the largest cluster. The prefolded curve shown in black is computed from $tRNA_{4R}$ simulation begun from the crystal structure and is identical to that shown in **Figure 4.4**. G26•A44, U8•A14, G15•C48, and G19•C56 were restrained in
both simulations.

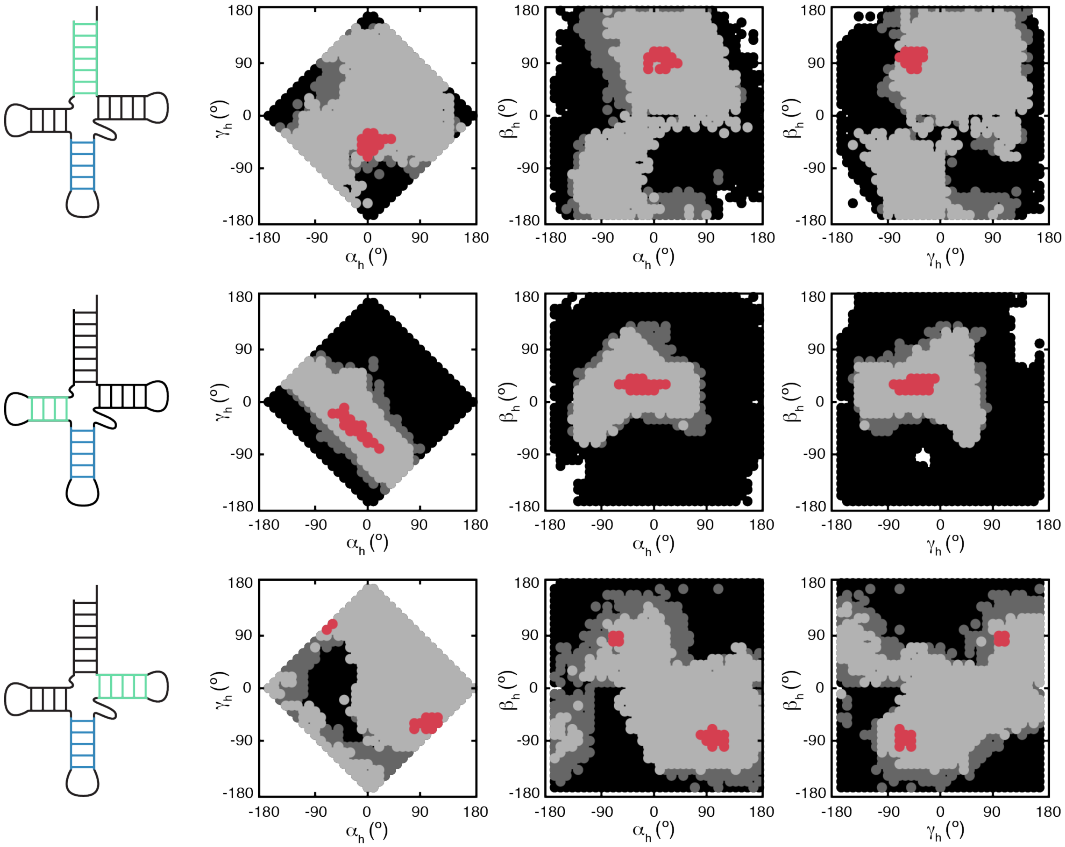

Figure C.8: Interhelical angles sampled by WT tRNA
2D projections of the (α_h, β_h, γ_h) angles sampled between different pairs of helices are shown for simulations of WT tRNA that are unrestrained, restrained by the non-redundant interactions G26•A44, U8•A14, G15•C48, and G19•C56, and fully restrained by all tertiary interactions shown in main text Figure 3A. Angles only sampled by unrestrained tRNA are shown in black, angles sampled by both unrestrained and non-redundantly restrained tRNA are shown in dark gray, and angles sampled by all three simulations are shown in light gray. Red points correspond to angles measured from 109 different tRNA crystal structures. Cartoons on the left indicate the plotted inter-helical angle, with stems highlighted in blue denoting the H1 reference helix and stems highlighted in green the H2 helix.

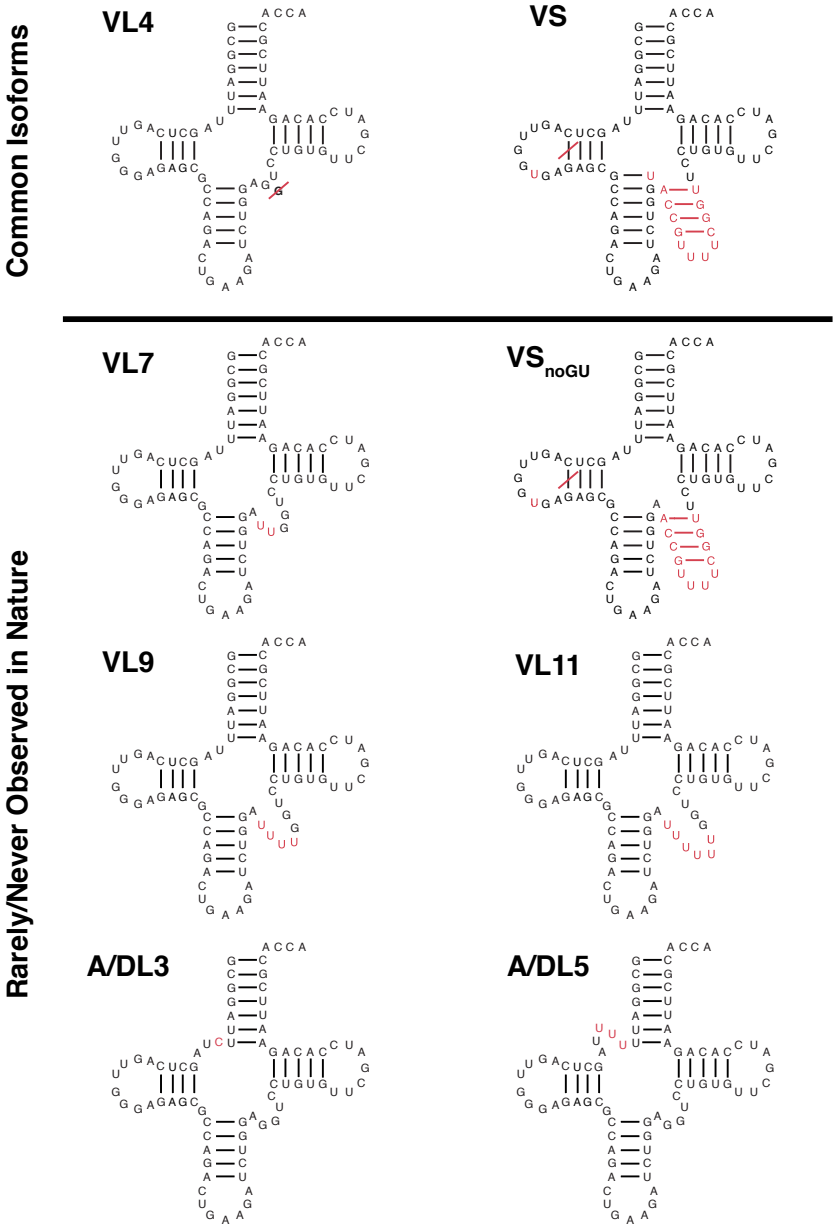

Figure C.9: Sequences and secondary structures of simulated tRNA variants
Inserted features compared to WT tRNA are shown in red, and deleted features are marked by a red strike.

Appendix D: Supporting Information for Chapter 5

D.1 Update of the TOPRNA force field

Simulations unrelated to this dissertation revealed that the original TOPRNA force field (TOPRNA1) occasionally allowed helices to undergo significant unnatural unwinding transitions. We determined that this was due to systematic errors in the dihedral potential placed along the backbone of consecutively paired residues. These errors are artifacts of the fitting procedure used to parameterize TOPRNA1, which used a four-term cosine series to fit the entire backbone statistical potential derived from the RNA05 structural database (1, 2). In fitting to the whole statistical potential, global agreement is prioritized over faithfully representing the shape and proper minima of the A-form energy well (Figure S1). Given that paired residues in TOPRNA are meant to permanently assume A-form conformations, and in fact can never adopt non-A-form conformations due to base-pair bonds, it thus makes sense to refit the potential to only the A-form energy well.

We reparameterized the backbone dihedral potentials using a single-term cosine of the form

$$U_{\phi,i} = K_i(1 + \cos(\phi - \delta_i)) \qquad \textbf{[D.1]}$$

and fitting it to the same RNA05-derived statistical potential described previously (1). The minimum of dihedral i, δ_i, was derived as $\delta_i = \pi + \langle\phi\rangle_{E<0.25}$, the average of all dihedral values

where the statistical potential had E<0.25 kcal/mol. K_i was derived by taking the average over all dihedral values within ±0.5 radians of δ_i

$$K_i = \left\langle \frac{E(\phi)}{\left(1+\cos(\phi-\delta_i)\right)} \right\rangle_{\delta-0.5 \le \phi \le \delta+0.5} \quad \textbf{[D.2]}$$

where E(ϕ) is the value of the statistical potential at dihedral value ϕ. All other aspects of the model were left unchanged. We name this new parameter set TOPRNA2, which we recommend as the new default all future TOPRNA simulations.

Benchmark simulations of a series of random sequence 8-bp hairpins confirmed that TOPRNA2 better represents long helices (**Figure D.1**). The simulated hairpins were constructed using the **–init** option of *toprnaCreate.pl* (brooks.chem.lsa.umich.edu), and contained a 4-nt capping loop with each stem base pair having a 1/9 probability of being a GU base pair and 8/9 probability being an AU or GC pair. Simulations were performed for 10^7 steps of Langevin dynamics at 300 K, using parameters described in Chapter 3. As a further test, we also repeated a number of our previous simulations of bulge motifs and cc-tRNA using TOPRNA2. Analysis showed that TOPRNA2 produced highly similar results to TOPRNA1 (not shown). For consistency, all values reported in Chapter 5 and Appendix D for cc-tRNA are calculated from a simulation performed using TOPRNA2.

D.2 Estimating changes in mutant melting temperature

At their respective melting temperatures the mutant and WT can be related through:

$$\Delta H^{WT}_{fold} - T^{wt}_m \Delta S^{WT}_{fold} = \Delta H^{mut}_{fold} - T^{mut}_m \Delta S^{mut}_{fold}. \quad \textbf{[D.3]}$$

We assume that inserting one or two nucleotides into the V-loop will negligibly affect the enthalpy of folding, allowing us to directly relate the melting temperature T_m to the total entropy change upon folding, ΔS_{fold}:

$$T_m^{wt}\Delta S_{fold}^{WT} = T_m^{mut}\Delta S_{fold}^{mut} \quad \textbf{[D.4]}$$

Assuming that the global rotational and translational degrees of freedom are the same in the folded and unfolded state, ΔS_{fold} can be split into two primary terms: $\Delta S_{fold,\ other}$, the change in entropy contributed by ions, water, and increased local flexibility of residues previously participating tertiary interactions; and $\Delta S_{fold,\ 3Dconf}$, the change in entropy due to the increased number of global 3D conformations populated in the unfolded state. We assume that $\Delta S_{fold,\ other}$ will not be affected by the insertion mutations, allowing us to express **[D.4]** as

$$T_m^{wt}\Delta S_{fold,\ 3Dconf}^{WT} = T_m^{mut}\Delta S_{fold,\ 3Dconf}^{mut} + \left(T_m^{mut} - T_m^{wt}\right)\Delta S_{fold,\ other}. \quad \textbf{[D.5]}$$

Finally, noting that $\Delta S_{fold,\ other}$ is expected be of a similar magnitude to $\Delta S_{fold,\ 3Dconf}$ under high salt solution conditions (5), and (T_m^{mut}-T_m^{WT}) to be ~10 K compared to T_m~300 K, the final term of [S5] should be negligible, allowing [S5] to be simplified to

$$T_m^{wt}\Delta S_{fold,\ 3Dconf}^{WT} \approx T_m^{mut}\Delta S_{fold,\ 3Dconf}^{mut}. \quad \textbf{[D.6]}$$

The primary contributor to $\Delta S_{fold,\ 3Dconf}$ will be the greatly increased number of interhelical conformations that the tRNAs populate in the unfolded state. The interhelical entropy of both the folded and unfolded states can be computed exactly as

$$S_{3Dconf} = -k_B \int p\left((\boldsymbol{\alpha}_\mathbf{h},\boldsymbol{\beta}_\mathbf{h},\boldsymbol{\gamma}_\mathbf{h})\right)\ln\left[p\left((\boldsymbol{\alpha}_\mathbf{h},\boldsymbol{\beta}_\mathbf{h},\boldsymbol{\gamma}_\mathbf{h})\right)\right]\mathbf{dV}, \quad \textbf{[D.7]}$$

where the integral is performed over all allowed intehelical angles, and $(\boldsymbol{\alpha}_\mathbf{h}, \boldsymbol{\beta}_\mathbf{h}, \boldsymbol{\gamma}_\mathbf{h})$ represents the set of 9D Euler angles describing a given tRNA conformation. **dV** is the volume element

$\sin\beta_{h,1}d\alpha_{h,1}d\beta_{h,1}d\gamma_{h,1}\ldots\sin\beta_{h,3}d\alpha_{h,3}d\beta_{h,3}d\gamma_{h,3}$. We estimated this integral from our simulations by constructing a 9D histogram and summing over all populated bins

$$S_{3Dconf} = -k_B \sum_i \frac{p_i}{V_i} \ln\left(\frac{p_i}{V_i}\right) V_i \qquad \textbf{[D.8]}$$

where V_i is the volume of bin i , and p_i is the population of bin i. In the folded state, only a single bin is populated and thus $S_{3Dconf}=k_B\ln(V_{nat})$, where V_{nat} is the volume of the native bin, yielding the final expression

$$\Delta S_{fold,3Dconf} = k_B \ln\left(V_{nat}\right) + k_B \sum_{i \neq nat} \frac{p_i}{V_i} \ln\left(\frac{p_i}{V_i}\right) V_i \ . \qquad \textbf{[D.9]}$$

As shown in **Figure D.6**, the ΔT_m estimates we obtain are relatively insensitive to choice of histogram bin size. We repeated our calculations using MBAR (6) to combine data from all eight temperature windows of our replica exchange simulations and found negligible differences between ΔT_m estimates obtained using only data at 300 K (not shown). Since helices also possess translational degrees of freedom, we also repeated the entropy calculation using the distance of the T-stem closing base pair from its native coordinate (see above) as a tenth histogram dimension. Here, the calculation was performed the same as in **[D.9]** with V_i multiplied by the distance dimension bin width.

Finally, it is worth noting that we also estimated $\Delta S_{fold,\ 3Dconf}$ using less descriptive order parameters such as structural RMSD or single axis rotation distance of the (α_h, β_h, γ_h) angles to the native conformation. These calculations produced much smaller estimates of difference between the mutant and WT $\Delta S_{fold,\ 3Dconf}$ due to their inability to fully capture the increased structural heterogeneity that is caused by the mutation. In other words, there are many tRNA conformations that have 15 Å RMSD from the native state, and RMSD alone cannot capture

changes in the heterogeneity of these non-native conformations. This highlights the importance of choosing good order parameters for entropy calculations.

D.3 Supporting tables and figures

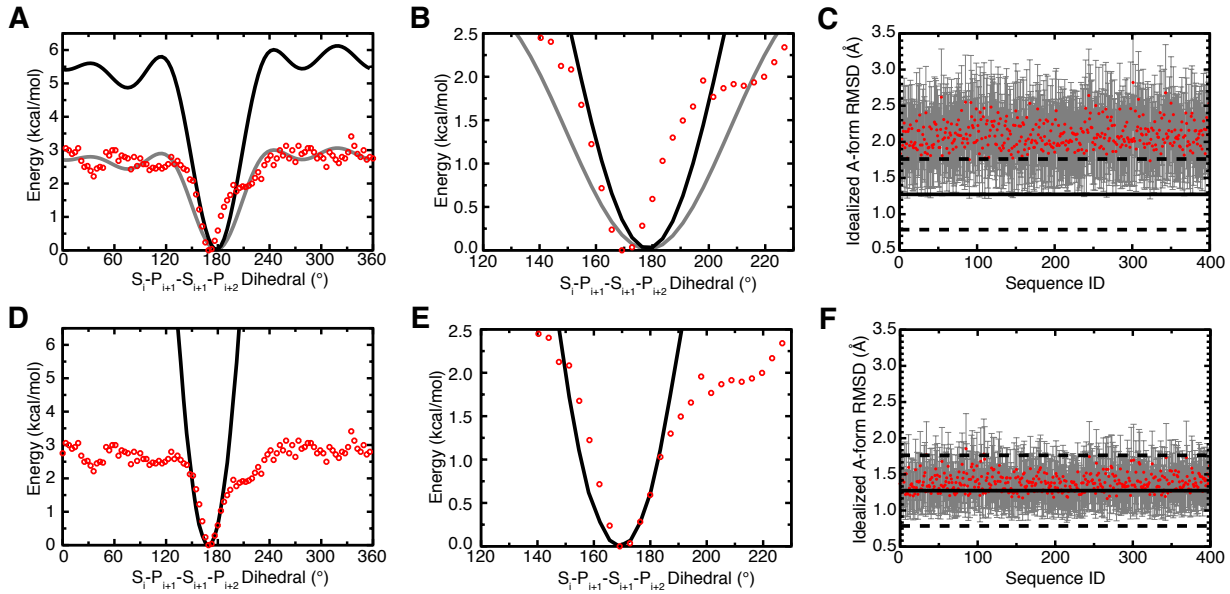

Figure D.1: Correction to the TOPRNA force field
(**A, B**) The sequence independent S_i-P_{i+1}-S_{i+1}+P_{i+2} dihedral as parameterized in TOPRNA1. The statistical potential derived from the RNA05 database is shown using open red circles, the original four-term cosine series fit shown in gray, and final TOPRNA1 potential where the fitted K_ϕ were doubled shown in black (1). A detailed view of the energy minima is shown in (B). (**C**) The RMSD from idealized A-form structure of 400 8-bp hairpins of random sequence simulated using the TOPRNA1 force field. RMSDs were computed the S and P-beads of the stem residues, with the mean RMSD over 10^7 of dynamics steps shown as red circles, and standard deviation shown with gray error bars. The mean and standard deviation of the RSMD from idealized A-form structure observed among 9284 8-bp helices in the PDB identified from a search of RNA FRABASE (7) are shown as horizontal solid and dashed black lines. These RMSDs were also computed using S and P-beads by converting the all-atom PDB structures to TOPRNA representations. (**D, E**) The S_i-P_{i+1}-S_{i+1}+P_{i+2} dihedral in TOPRNA2 with the same color scheme as in (A,B). (**F**) Same as (C), except simulated with TOPRNA2. All sequences with mean RMSD greater than ½ standard deviation above the PDB mean have at least one GU base pair, or for four special cases, contain ≤2 GC base pairs.

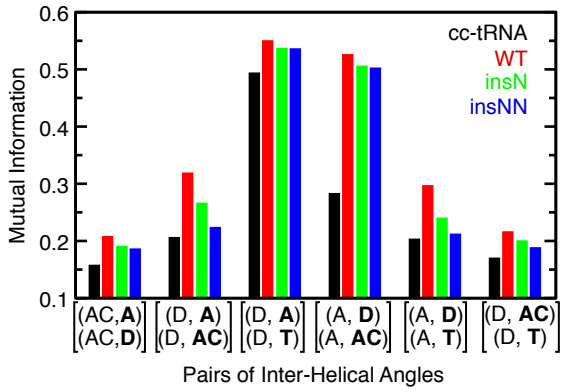

Figure D.2: Mutual information between mt-tRNA^{Ser} interhelical angles
The two helices whose orientations are being correlated are bolded. MI of 0 indicates no correlation.

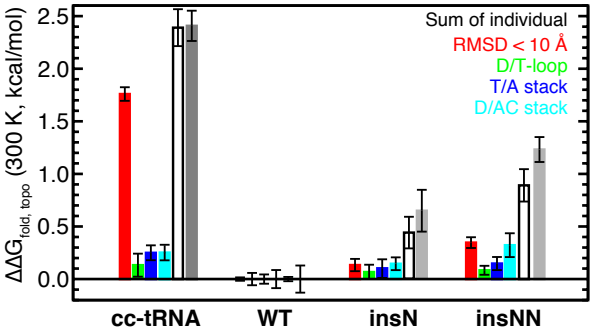

Figure D.3: Contributions to $\Delta\Delta G_{fold,\ topo}$ between mutant and WT mt-tRNA^{Ser} and cc-tRNA
$\Delta\Delta G_{fold,\ topo} = \Delta G_{fold,\ topo}(i) - \Delta G_{fold,\ topo}(WT)$. Shown in gray is the $\Delta\Delta G_{fold,\ topo}$ computed when native structure is defined as RMSD<10 Å with the existence both D-to-T-loop contacts and both interhelical stacks. The $\Delta\Delta G_{fold,\ topo}$ of forming conformations with RMSD < 10 Å is shown in red. The $\Delta\Delta G_{fold,\ topo}$ of additionally forming individual tertiary interactions given RMSD < 10 Å is shown by the green and blue colored bars. Open bars represent the sum of the individual contributions. Differences between the sum and the total $\Delta\Delta G_{fold,\ topo}$ (gray) are due to losses in cooperativity among the individual interactions.

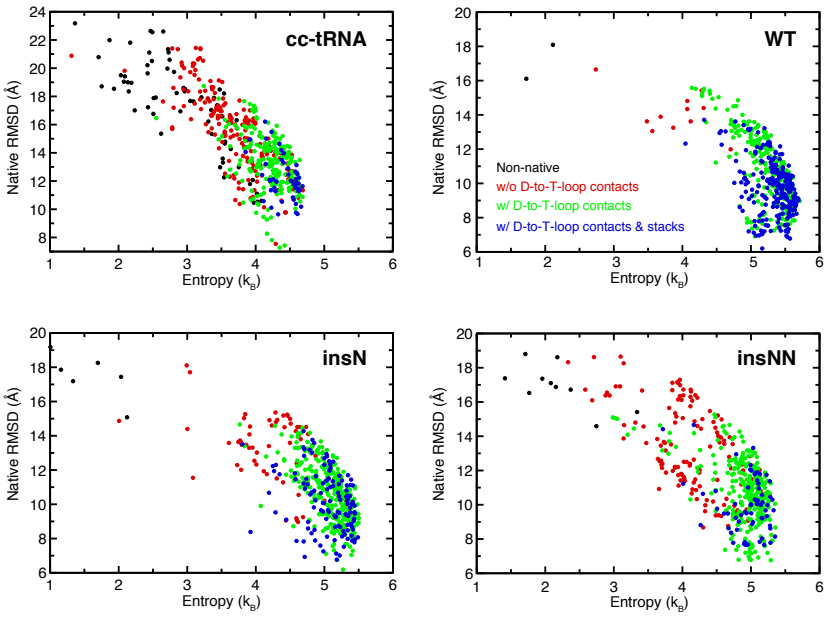

Figure D.4: Entropies of the best-packed mt-tRNA^{Ser} conformers
Entropies and native RMSDs are shown for the 500 best-packed conformers sampled by the different tRNA simulations (G26-U44 pairing enforced). Entropies were computed based on how many other structurally similar conformations are in the best-packed ensemble (see methods). Conformations that possess native-inconsistent contacts are shown in black. All other conformations are colored based on the subset of native contacts that they possess, which are labeled in the figure. Note that high entropies indicate conformers that are thermodynamically favored.

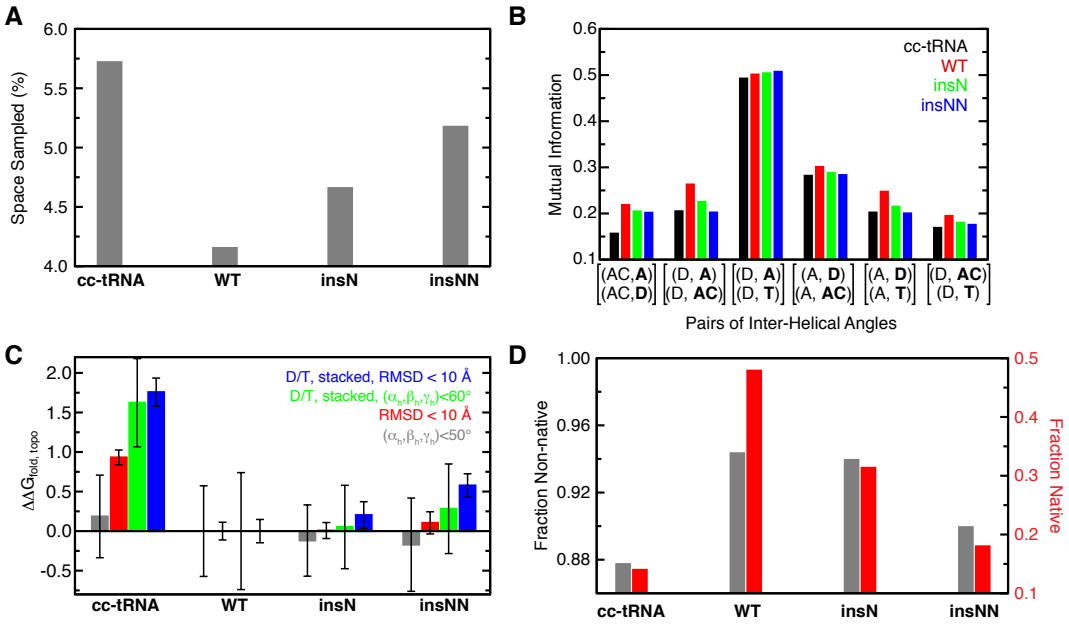

Figure D.5: Constraints of mt-tRNA^{Ser} in the absence of G26-U44 pairing
(A) The fraction of theoretically possible (α_h, β_h, γ_h) states sampled by TOPRNA simulations. **(B)** Mutual information measured between different pairs of tRNA interhelical angles, measured with respect to a common reference helices. The two helices whose orientations are being correlated are bolded. **(C)** Differences in the topological constraint contribution to the folding free energy $\Delta\Delta G_{fold,\ topo} = \Delta G_{fold,\ topo}(i) - \Delta G_{fold,\ topo}(WT)$. Values and error bars represent the mean and standard deviation of $\Delta\Delta G$ computed by block averaging over thirds of the simulations. **(D)** Fraction of the 500 best-packed conformations from each tRNA simulation that have non-native contacts (gray) or that have D-to-T-loop contacts and both interhelical stacks (red). The fraction of conformations with both native interactions is weighted by conformational entropy based on how many structural similar conformations are in the best-packed ensemble.

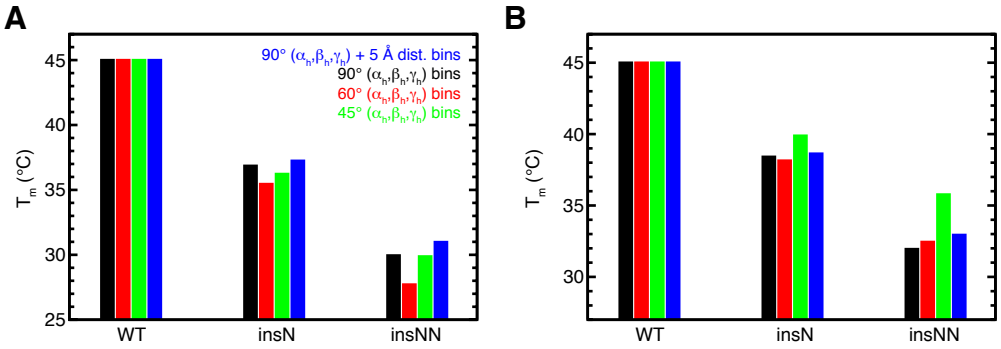

Figure D.6: TOPRNA estimates of insN and insNN T_m
Assumed WT T_m=45° C for mt-tRNAs with **(A)** and without **(B)** G26-U44 pairing. Calculations were done for several different bin sizes with the obtained results shown in different colors (SI methods). As the bin width becomes finer, the accuracy of the bin populations decreases. Note that although the bin widths are large, because we are binning based on (α_h, β_h, γ_h) value rather than single axis rotation distance to the native state, these bins are still able measure changes in heterogeneity of accessible states near the native conformation.

	$T_{m,1}$	ΔH_1	$T_{m,2}$	ΔH_2	$T_{m,3}$	ΔH_3	$T_{m,4}$	ΔH_4
WT	40	47	53	70	68	42	81	73
insG	34	51	50	33	71	56	83	120
insA	31	43	52	27	71	63	81	73
insU	33	46	53	36	66	80	73	52
insUU	29	54	54	35	68	94	74	77

Table D.1: van't Hoff fit parameters to UV melting data at 2mM $MgCl_2$
All temperatures are in units of °C, and ΔH in units of kcal/mol. Estimated errors are denoted by superscripts, which were estimated manually based on repeating the fits with different initial parameters, bootstrapping analysis, and comparison between experimental replicates. $T_{m,1}$ errors are estimated to be ±1 °C, and $T_{m[2,3,4]}$ errors are estimated to be ±5 °C. ΔH_1 errors are estimated to be ±10% and $\Delta H_{[2,3,4]}$ errors are estimated to be ±50%.

	$T_{m,1}$	ΔH_1	$T_{m,2}$	ΔH_2	$T_{m,3}$	ΔH_3	$T_{m,4}$	ΔH_4
WT	45	51	57	91	71	40	92	123
insG	40	42	53	65	61	65	74	52
insA	37	40	50	70	57	58	73	46
insU	39	40	54	71	60	58	73	48
insUU	34	47	59	43	70	83	76	50

Table D.2: van't Hoff fit parameters to UV melting data at 5 mM $MgCl_2$
All temperatures are in units of °C, and ΔH in units of kcal/mol. Estimated errors are denoted by superscripts, which were estimated manually based on repeating the fits with different initial parameters, bootstrapping analysis, and comparison between experimental replicates. $T_{m,1}$ errors are estimated to be ±1 °C, and $T_{m[2,3,4]}$ errors are estimated to be ±5 °C. ΔH_1 errors are estimated to be ±10% and $\Delta H_{[2,3,4]}$ errors are estimated to be ±50%.
