Modeling RNA Junctions with Applications to Structure Predictions of Regulatory Regions of Viral RNAs

by

Segun Jung

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Program in Computational Biology

New York University

September 2013

Professor Tamar Schlick

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

"*Success does not consist in never making mistakes but in never making the same one a second time.*"

George Bernard Shaw

DEDICATION

For my parents, brothers, Yoonyoung and Minjun, whose love, support, and patience sustain me during the studies

ACKNOWLEDGEMENTS

I express my sincere gratitude for all the support and encouragement from my mentors, colleagues, friends and family during the academic years. I cannot imagine completion of this work without them.

First and foremost, I would like to thank my thesis advisor Prof. Tamar Schlick, whose guidance, inspiration, and advice have been indispensable over my graduate study. I am especially grateful for her perspective in science as well as in mathematical and computational biology, which will no doubt shape my own philosophy I am today. She always encourages being an independent thinker that takes crucial part of the research training in Schlick lab. I also enjoyed her broad interest in sports and theater: as she is an excellent athletics in various sports including swimming, yoga, and running. Some of her excellent suggestions for group outing are Shakespeare in the Park and Anna Zieglers play Photograph 51.

I am deeply grateful to have my thesis committee members, Drs. Tamar Schlick, Yingkai Zhang, Bud Mishra, Michael Shelley, and Daniel Tranchina, for their interest and support in my thesis work.

I am also thankful for the friendship and help offered by current and previous Schlick lab members: Abdul Iqbal, Dr. Antoni Luque, Dr. Christian Laing, Dr. Durba Roy, Dr. Giulio Quarta, Dr. Hin Hark Gan, Dr. Jeremy Curuksu, Joseph Izzo, Leif Halvorsen, Dr. Mai Zahran, Dr. Meredith Foley, Dr. Namhee Kim, Dr. Ognjen Perisic, Dr. Rosana Collepardo, Shereef Elmetwaly, and Yunlang Li. I value especially my close friendship with Dr. Meredith Foley; she has always been available to listen my personal concerns, scientific ideas during my stay at NYU. I also acknowledge Dr. Christian Laing for his help when I first joined the

lab and his continuing friendship during the ensuing years. Although I cannot name them all in this limited space, I would also like to extend my gratitude to all of my friends and colleagues in the NYU Department of Chemistry.

I thank my fellow students (past and present) in Computational Biology Program at NYU, whom I have shared this incredible journey throughout the graduate study. Special thanks are to Drs. Kenneth Ho, Gulseher Sarah Sirin, and David Rooklin. Most importantly, I am deeply thankful for Dr. Michael Shelley as being an excellent adacemic mentor and friend throughout the graduate study.

I am deeply indebted as well to Drs. Louise C. Showe and Michael K. Showe at the Wistar Institute whose guidance and encouragement made me begin this odyssey journey here at NYU. I am also fortunate to work with Dr. Jeff Bell who guided me an exciting summer research project at Schrodinger Inc., and to have Dr. Marc Parisien at the University of Chicago for friendship and his help for a computational program.

I would like to acknowledge several fellowships for their financial support since my arrival in the United States in 2004; the two year full scholarship from the Korean Government when I first arrived at the Drexel University in Philadelphia for masters program in Biomedical Engineering; biomedical sciences training fellowship from Sackler Institute at the NYU Langone Medical Center and McCracken fellowship from the Graduate School of Arts and Science at NYU throughout the Computational Biology IGERT Doctoral Program. This work was supported by the NSF (EMT award CCF-0727001), NIH (GM081410 and GM100469), the Human Frontier Science Program, by a joint NSF/NIGMS initiative in Mathematical Biology (DMS-0201160) and partially by NIH (R01-GM055164 and R01 ES 012692).

Last but certainly not least, I am eternally thankful to my family in Korea—my parents Sunsang Jung and Youngim Lim, and my two older brothers Yugeun and Youngeun and their beloved own families, whose trust helped me sustain difficult times—and to my own family Yoon Young and Minjun Jung, who have always inspired me with unforgettable joy, overflowing love, profound wisdom, and sound encouragement on a daily basis.

ABSTRACT

RNA junctions are key motifs in the organization of RNAs since they define the global topology of the biomolecule. RNA junction analysis and prediction with applications to viral RNAs are the main subjects of this thesis.

Motivated by previous studies of 3 and 4-way RNA junctions, we analyze higher-order RNA junction structures using a non-redundant dataset of RNA crystal structures to show that sub-junctions contain helical arrangements of lower-order RNA junctions and that recurrent tertiary motifs such as A-minor interactions stabilize junction architecture.

Based on the knowledge obtained from the RNA junction analysis, we develop a novel bioinformatics/data mining approach to predict helical topologies of RNA junctions as tree graphs, called RNAJAG (RNA Junction-As-Graph). Using a large set of 200 junctions, we show that RNAJAG predicts reasonably the helical junction configurations in 3 and 4-way junctions of the native RNAs.

The combined advances in RNA junction analysis, prediction, and modeling lead us to propose candidate RNA junction structures of regulatory regions, called internal ribosome entry site (IRES), of the foot-and-mouth-disease virus (FMDV). Based on all available experimental data, we model junction topologies, build atomic 3D models, and investigate candidate structures by MD simulations to determine the most energetically favorable configurations and analyze tertiary interactions. Our collective findings suggest a plausible theoretical tertiary structure of the apical region in FMDV IRES domain 3. Our work provides insights into the potential role of the long-range interactions for structural stability and organization of domain 3.

There is also much interest in the dynamic nature of RNA junctions as

they are capable of undergoing conformational changes that are often linked to important biological functions. Thus, we study the dynamic properties of 4-way RNA junctions, as found in FMDV IRES domain 3, employing molecular dynamics (MD) simulations. Our results suggest a mechanism of interconversion between different conformations of the junction via a rotation between helical axes of coaxial stacking conformers. Together with the theoretical candidate structure investigation, this mechanism is crucial to understand the possible conformational change of the junction that will help elucidate required tertiary contacts for RNA structure stability and their roles in important biological functions.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Background and Introduction

1.1 Biological roles of RNA

RNA has long been known to play a central role in information transfer by delivering to DNA instructions on protein synthesis. This classical paradigm had been expanded since the recent discovery of non-coding RNAs [30] to numerous functional roles that encompass gene regulation at all stages of the cell life cycle. Deciphering the functions of these gene-regulating RNAs presents an exciting challenge for the next decade.

To understand their biological functions, determination of the structural features of RNAs is essential because sequence alone does not provide enough information. Indeed, RNA structural biology has been providing insights into detailed descriptions of structure-function relationship. One of the prominent examples was the determination of high resolution crystal structures of large non-coding RNA, called ribosome [5, 149, 140], providing overall architecture of RNA folding and its interaction with proteins; contribution of these works was rewarded by the 2009 Nobel Prize in Chemistry.

Characterizing structural aspects of RNA has been a great challenge Since the very first key discovery of RNA about 60 years ago; only in 1956, RNA structural biology began with a report that two single-stranded RNA molecules (polyribo U and polyribo A) could spontaneously hybridize to form a double-stranded RNA helix [150]; about 20 years later, the first crystal structure of complex transfer RNA with full atomic details was solved by Klug [151] followed by the determination of large ribosomal RNAs in 2000 [5, 149, 140]. Several decades of effort on the work of RNA structural biology has led to remarkable progress on an understanding the details of RNA structures (e.g., RNA 3D motifs, RNA-protein interactions) to their biological functions. However, we face more challenges today with the findings of new non-coding RNAs [152, 153]. The functional roles of these RNAs remain elusive, and exciting new discoveries of regulatory roles have yet to come.

1.2 Fundamental structural elements of RNA

RNA is a single-stranded polymeric molecule composed of four nucleotides—adenine (A), guanine (G), cytosine (C), and uracil (U). Each nucleotide is composed of three different entities—base, sugar, and phosphate; the bases come from two groups: purine (adenine and guanine) and pyrimidine (cytosine and uracil) (Figure 1.1A). These nucleotides can interact with each other to pair bases via hydrogen bond formations: three hydrogen bonds for G-C and two for A-U base pairs (Figure 1.1B). The base pairs stack to form a double-stranded helix (Figure 1.1C) where the base pairing interactions can be classified in three different types: canonical Watson-Crick base pairs (A-U, G-C), wooble base pair (G-U), and non-canonical base pairs (A-A, A-G, A-C, C-C, C-U, G-G,

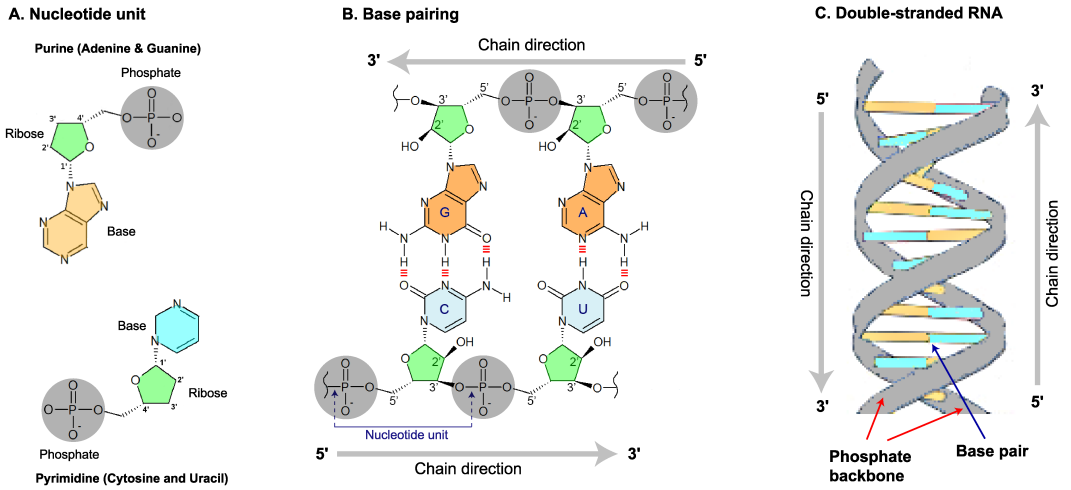


Figure 1.1: Three chemical entities of nucleotide unit and base pairing in RNA. **(A)** Each nucleotide unit is composed of base (purine or pyrimidine, each colored orange and cyan), sugar (colored green), and phosphate (colored gray). **(B)** Nucleobases can form hydrogen bonds (colored dotted red) to pair bases: GC and AU base pairs are shown, for example. **(C)** Double-stranded RNA formed by stacked base pairs

U-U) [154, 86].

These different base pair types result in the formation of fundamental RNA 2D structural elements—loops, bulges, helices, and junctions (Figure 1.2). A hairpin is formed when two regions of the same strand complements each other and build a double-stranded helix that ends in an unpaired loop. A bulge and internal loop are defined when introduced unpaired residue(s) on one side and both sides of strands between two stems, respectively. An RNA junction serves as a hub for different double-stranded helical arms [90]; for instance, 3-way junction is composed of three different helical stems. RNA junctions play crucial

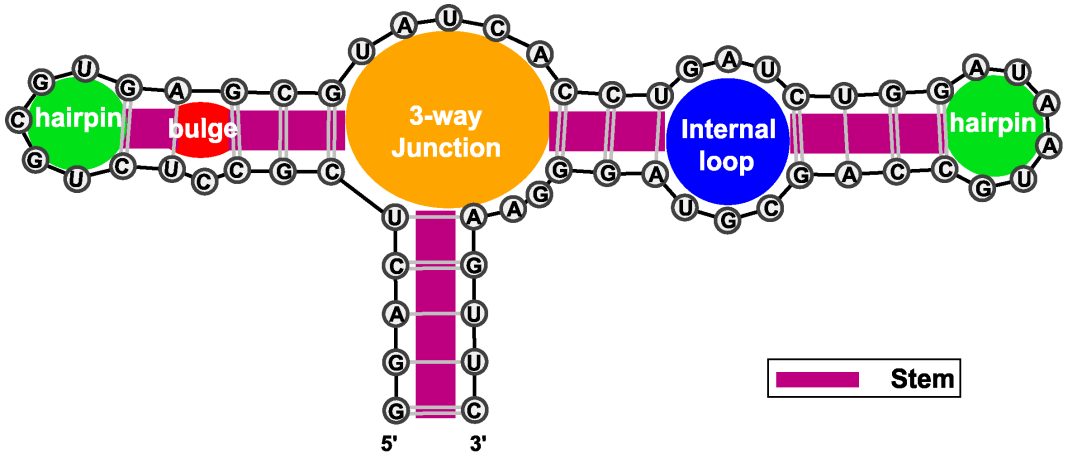


Figure 1.2: Secondary structural elements of RNA: bulges, hairpin loops, internal loops, and junctions. Hairpin loop (colored green) when forming a double helix with that ends in an unpaired loop; bulge and internal loop (colored red and blue, respectively) when introduced unpaired residue(s) on one side and both sides of strands between stems, respectively; 3-way junction (colored orange) connecting three different helical arms.

roles in RNA folding, serving as guides to the overall RNA architecture [182], and will be studied extensively in this thesis.

RNA junction motifs and structural analyses. RNA junctions are ubiquitous, found in a wide range of species from small RNAs [6, 65, 183] to large ribosomal subunits [15, 106, 58]. Thus, structural, energetic, and dynamic aspects of the junction motifs are essential to advance our current understanding of a functional role in RNAs.

A growing amount of RNA structural data obtained mainly from X-ray crystallography and NMR (nuclear magnetic resonance) have provided an exceptional opportunity to study structural properties of RNA junctions. For

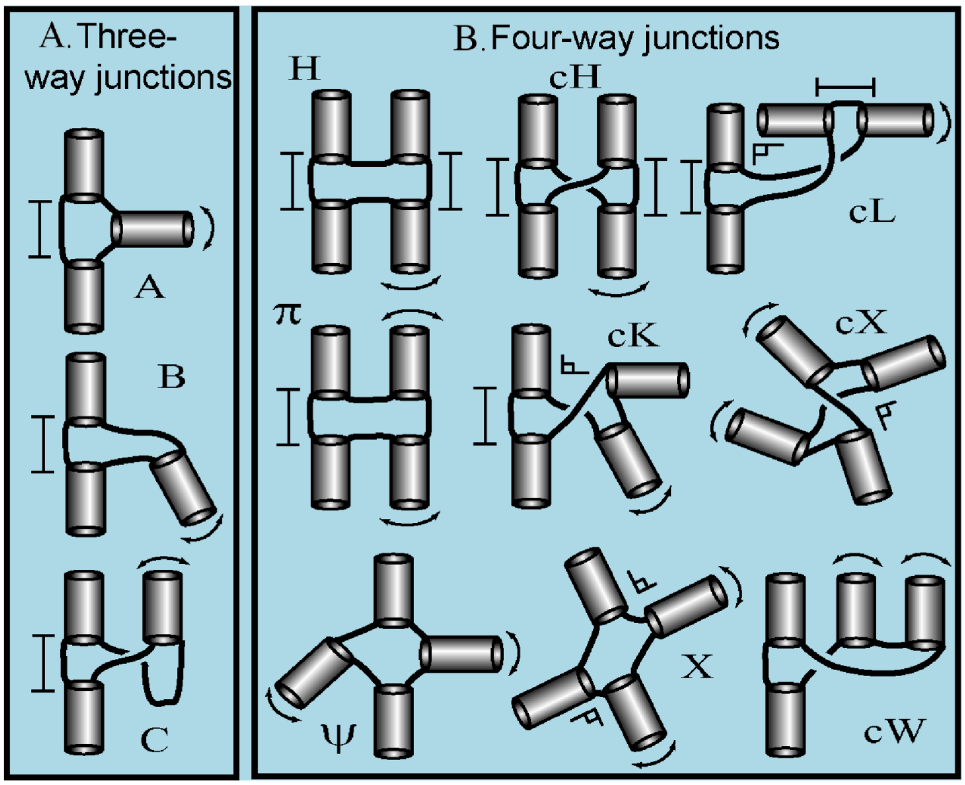


Figure 1.3: Schematic representation of 3 and 4-way junction families. **(A)** Three major family types—A, B, and C—are found in 3-way junctions where the helical arm, not involved in coaxial stacking, has different helical arrangements with respect to the coaxially stacked helices. **(B)** Nine major families—H, cH, cL, cK, ϕ, cW, ψ, cX, and X—are determined in 4-way junction based on coaxial stacking and overall helical arrangements.

example, Lescoute and Westhof [87] compiled and analyzed the 3-way junctions, classifying the topologies into three different families and formulating the rules of coaxial helical stacking formation (Figure 1.3A); two adjacent helices most likely stack each other when connected with $\leq$2 nucleotides (*nt*). Laing *et al.* [72, 73] extended the topology classification and analysis to study higher-order junctions, grouping 4-way junctions into nine family types based on coaxial stacking formation and overall helical arrangements (Figure 1.3B). Bindewald *et al.* [11] developed the RNAJunction database, which provides tens of thousands of solved RNA junctions with detailed structural annotations. All these studies, however, provide limited insights into dynamic properties.

Conformational transitions in RNA junctions. RNA junctions are dynamic structural entities capable of undergoing conformational transitions. A prominent example is the 4-way junction of hepatitis C virus (HCV) IRES where two different conformations—parallel and antiparallel configurations—were reported by crystallography and single-particle cryo-EM techniques [58, 185]; later, Lilley *et al.* [192] studied the IRES RNA junction using fluorescence resonance energy transfer (FRET) and showed that these two different stacking conformations are related via continuous interconversion. Such dynamic characteristics of 4-way junctions often have functional significance. For example, U1 snRNA [184, 186] plays a crucial role in organizing a whole RNA structure; hairpin ribozyme [187] involves in forming a catalytic site for RNA self-cleavage reaction; and viral mRNAs [188] are involved in the maturation gene translation.

An example of RNA junctions in viral RNA. The viral replication of foot-and-moth-disease virus (fmdv) begins with a translation initiation by form-

ing a specific RNA structure called internal ribosome entry site (IRES) on which ribosomal complexes can bind for a gene expression. Structural elements of IRES such as 4-way junctions in domain 3 can form a compact three-dimensional (3D) structure and thus the 3D structural determination of IRES is crucial to exploring and deciphering the initiation mechanism of translation. However, little is known about the structure. We will study the candidate junction structures extensively in this thesis.

Base stacking interactions. In addition to the base pairing interactions, base stacking interactions are also an important factor contributing significantly to maintain 2D structural elements via the London dispersion forces [155], known as dispersion forces between atoms and molecules, and electrostatic interactions [156, 157], attractive or repulsive forces due to the presence of electronically charged particles. While maintaining the 2D structure, RNA folds into a compact 3D shape via various tertiary interactions, also called motif.

1.3 RNA folding principles

The folding of RNA is hierarchical [154]: starting from a single sequence, a 2D structure is formed composed of various helical elements followed by a 3D structure formation via pairwise tertiary interactions. For example, Figure 1.4 shows the hierarchical folding of TPP riboswitch (PDB entry 2GDI) [180] where two tertiary interactions, A-minor (colored red) [105] and ribose zipper (colored cyan) [131], are involved to bring cooperatively the helices in distance to fold; the A-minor motif is defined when adenines are inserted into the minor groove of neighboring helices while the ribose zipper motif is a tertiary interaction

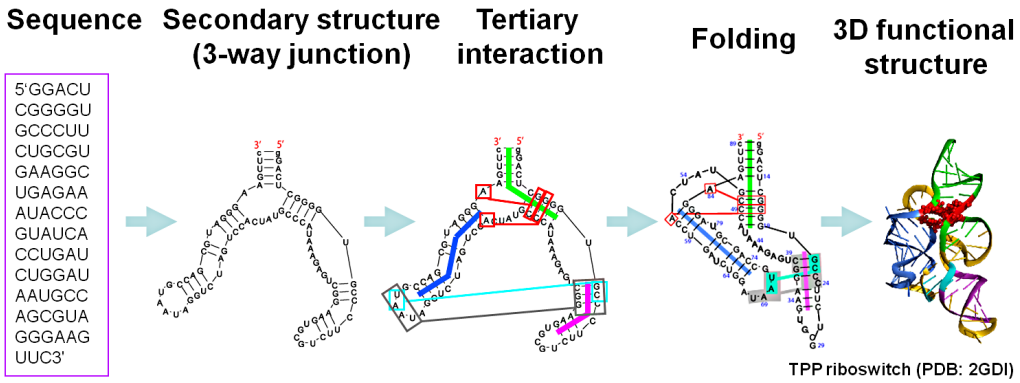


Figure 1.4: RNA structure and folding. RNA folding is hierarchical: starting from a single sequence to form 2D structure followed by a 3D structure via tertiary interactions.

between the backbone ribose 2-hydroxyls of two different regions in an RNA chain.

Indeed, tertiary interactions help stabilize RNA fold and are largely classified in three different groups: loop-loop interactions (e.g., kissing hairpin and pseudoknots), loop-helix interactions (e.g., A-minor, ribose zipper), and helix-helix interactions (e.g., coaxial stacking) [158]. Figure 1.5 shows some of the RNA-RNA tertiary interactions in fmdv ires domain 3 that are the essential key to structural stability and organization [148]. Besides these interactions, ion, solvent, or other molecules such as ligand and protein can also affect the fold of RNA.

1.4 RNA 3D structure prediction

RNA 3D structure is essential to understand its role in biological processes. Though often compared with protein folding prediction problems, RNA fold-

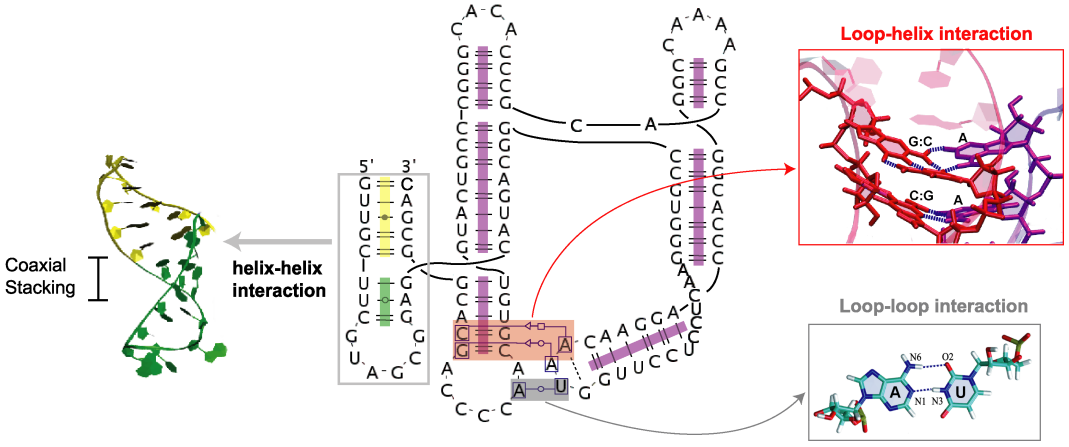


Figure 1.5: Annotated diagram of domain 3 in foot-and-mouth-disease virus internal ribosome entry site shows 3D motifs (loop-loop, loop-helix, and helix-helix interaction) working cooperatively to help stabilize the overall conformation.

ing prediction is still relatively young and immature, encountering difficulties dealing with large and complex structures. In order to help overcome and accelerate the investigation of RNA molecules, mathematical and computational approaches have contributed to the RNA structure prediction field.

RNA2D3D [159] and ASSEMBLE [160] are semi-automated programs that build first-order approximations of RNA 3D models using secondary or tertiary structure information from homologous RNAs. Other automated 3D structure prediction programs have been developed; FARNA [23], iFoldRNA [161], and NAST [55] rely on coarse-grained modeling with simulations to fold RNAs with the guidance of physics or knowledge-based energy functions; MC-Sym [109] predicts all-atom models of RNA by inserting small cyclic motif fragments, collected from solved RNA structures. BARNACLE [162] uses a coarse-grained probabilistic model of RNA to predict atomic models by efficient sampling of RNA conformations. MOSAIC [163] is another approach to efficiently and ac-

Table 1.1: List of programs for RNA 3D structure prediction

Name	Description	Reference
RNA2D3D	Semi-automated program that builds a first-order approximation of RNA 3D models using sequence and secondary structure information	[159]
ASSEMBLE	Semi-automated program that can analyze, manipulate, and build RNA 3D models	[160]
FARNA	Automated de novo prediction program that builds native-like RNA tertiary structures guided by a knowledge-based energy function	[23]
iFoldRNA	Automated RNA structure prediction and folding program using discrete molecular dynamics simulations	[161]
NAST	Automated RNA folding program that uses a knowledge-based potential coupled with a coarse-grained molecular dynamics simulation	[55]
MC-Sym	Automated RNA prediction program that builds RNA using small cyclic motif fragments collected from solved RNA structures	[109]
BARNACLE	A coarse-grained probabilistic RNA structure prediction program by efficient sampling of RNA conformations	[162]
MOSAIC	A Monte Carlo sampling approach that uses local and global hierarchical moves of RNA to efficiently and accurately model RNAs	[163]

curately model RNAs by including the local and global hierarchical folding principles.

While these advances are significant, current limitations of all such programs, however, lie in predicting large or complex RNA structures, mainly due to the large size of the conformational space. In particular, predicting the 3D structures of RNA junctions, formed by multiple helical arms, is challenging because the spatial organization is often determined by non-canonical base pairs and base stacking interactions. Furthermore, even if these programs can successfully generate models that locally resemble native RNA structures, the spatial organization of helical elements in junctions tend to be inaccurate, thus requiring manual intervention, as recently reviewed by Laing and Schlick [74].

1.5 Graph modeling approaches

Graph theory is a field of mathematics and computer science that study graphs to model pairwise relations between objects. A graph is made up of vertices (nodes) and edges (lines) that connect the vertices (Figure 1.6). A graph may be directed or undirected depending on whether two vertices associated with each edge are distinctive. Graphs or networks can be used to model various types of relationships and dynamic processes in physical, social, medical, and biological contexts [59].

As an application of graph theory, graphical representations are used to catalog and organize structural features of RNA [37, 63, 60, 62]. The main advantage of graph theoretical representation is a much reduced conformational sampling space. Indeed, using tree graphs to describe the discrete repertoire of RNA molecules has led to prediction of new RNA folds and design of novel

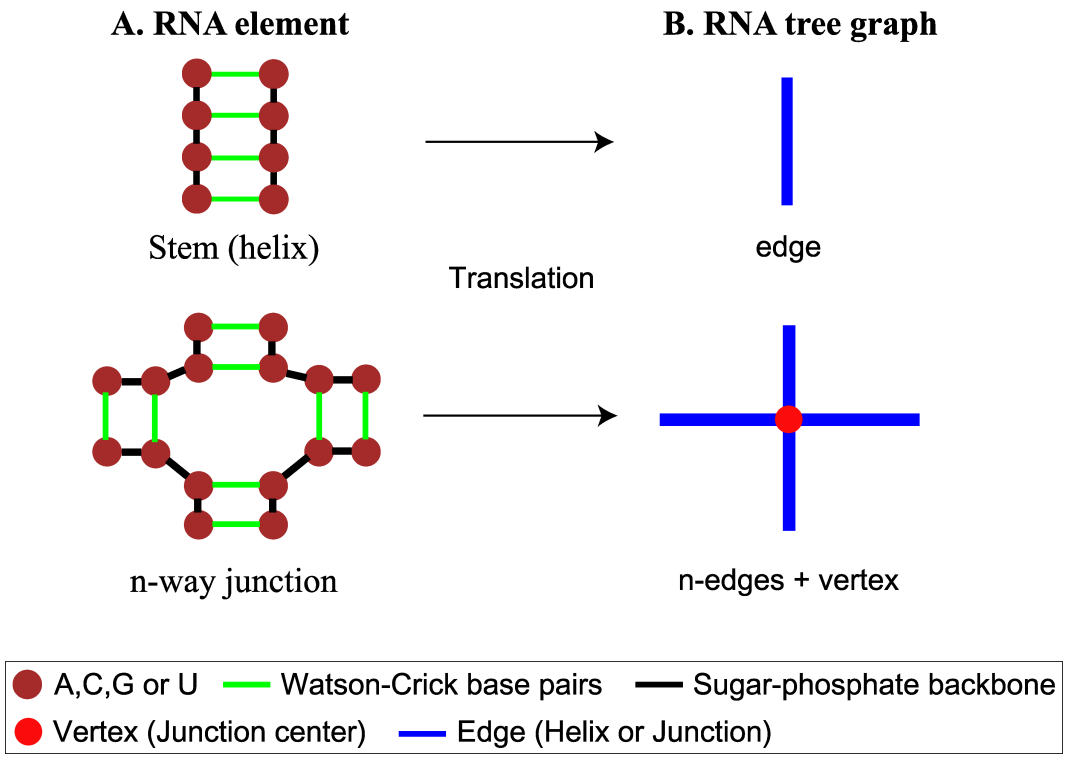


Figure 1.6: RNA graph representations. **(A)** RNA junction elements in a secondary structure. **(B)** RNA tree graph representation, which describes a helix as an edge and a loop as a vertex.

motifs [75, 164]. These graphical approaches began with pioneering works of Waterman [165], Shapiro [166], and others [168, 167, 37]. Recently, Schlick and coworkers introduced the RNA-As-Graph (RAG) tree and dual graphs to represent RNA 2D topologies, catalogue all possible topologies [37, 63], and predict novel RNA motifs [63, 60, 62, 61]. Knisley and coworkers applied the RAG tree graphs to analyze secondary structures of RNAs and predict larger RNA-like structures by merging two RNA graphs and applying neural network analysis [68]. Gopal *et al.* applied RAG to model large viral RNAs [169]. Many other applications of RAG have been reported (see review in [164]). The

much reduced RNA conformational space using these graphs opens new ways to describe and predict large RNA topologies, as described in Chapter 3 of the present thesis.

1.6 Molecular dynamics simulations of RNA

Understanding how this central molecule of life is able to accomplish such a variety of functions necessarily involves RNA structure and dynamics at the atomistic level, a challenge that can be addressed using a combination of experimental studies and Molecular dynamics (MD) simulations.

Molecular dynamics is a computer-based simulation that deals with physical motions of atoms in a given system. The atoms are allowed to interact for a period of time, giving in turn a trajectory of the simulated system. The trajectories are basically by-products of numerical solution of the Newton′s equations of motion for interacting atoms, where forces between the atoms and potential energy are defined by molecular mechanics force fields. Although MD in general has limitations due to a number of approximations (e.g., force field) [98, 170, 171] and sampling [124], it is still useful to capture dynamical features of simulated RNA systems that provides invaluable insights to interpret experimental data such as molecule binding sites or interactions involving ribosomal RNAs [174, 173], binding interaction between small molecule and riboswitch [172], and important structural motions of RNA junctions in the ribosome [175].

Only in 1983, it became possible to achieve pioneering MD simulations of 12 ps time-scale on tRNA with many approximations (e.g., in vacuum condition). Although the high expectations that computer simulations might soon replace

experiments in the laboratory [124] have not come true, continuous improvements, such as finer treatments of solvent and electrostatics, have occurred. Only in 1995, the particle mesh Ewald summation techniques with explicit solvent made possible to simulate molecular dynamics for RNA [176].

Another major problem in the field of RNA MD simulations was the lack of high-quality RNA crystals that serve as a starting template. For example, a catalytic core of the hammerhead ribozyme demonstrates the importance of solved RNA structures [177] that could lead to irrelevant trajectory of MD. An accurate starting structure is crucial for a realistic assessment of dynamics in the simulated system. Fortunately, more than 6,000 structures as of May 2013 are available in the NDB (Nucleic Acid Database).

The increasing computing power available for MD calculations has dramatically boosted the field in recent years [124]. Yet, current challenges are to improve the quality of the force fields, specifically for backbone parameters [119]. Among the two popular force fields for nucleic acids are—AMBER [17, 111] and CHARMM [181]—AMBER has been extensively used for RNA/DNA simulations while CHARMM for describing protein or DNA/RNA-protein complexes. The combination of high-quality starting structures, improved force fields, and increased computational power make MD a very promising technique to study the structure and dynamics of RNA, especially in conjunction with experiments.

This historical overview led us to propose the "field expectation [191]" curve of RNA modeling and simulation shown in (Figure 1.7) that particularly emphasizes the transition from initially unrealistic excitements with high hopes followed by disappointments to a more practical viewpoint with many productive progress in theory, technology, and experiment.

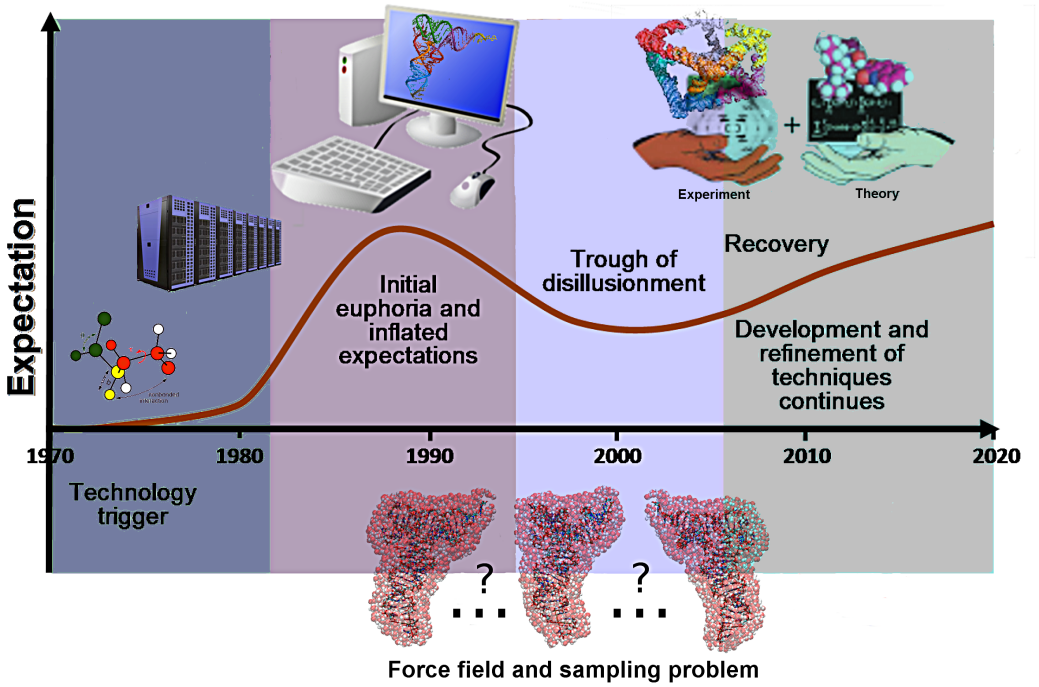


Figure 1.7: Proposed expectation curve for the field of RNA modeling and simulation, with approximate timeline [124]. The field launched when comprehensive molecular mechanics efforts started, and received a momentum with the availability of fast workstations as well as supercomputers. The unrealistically high expectations raised from the first simulation of tRNA gave in turn disappointments for a decade mainly due to force field problems. Since then, the field becomes more mature with realistic progress and balance between theory and experiment.

1.7 Challenges in the field of RNA 3D structure prediction

Although significant advances made by increasing computational resources and technologies have been achieved over the past decades in the field of RNA modeling, many challenges still remain.

Perhaps the most challenging problem in RNA modeling is the prediction of long-range tertiary interactions. Although not highly reliable, one possible way to predict candidates of long-range interactions is to perform analysis using multiple sequence alignment where patterns of conservation can be observed by covariation (mutation and counter-mutation) of sequences at different positions in the RNA molecules. Currently, programs such as ISFOLD [178] and SHEVEK [179] are capable of predicting the tertiary contacts using the MSA technique.

Several RNA 3D structure prediction methods have shown that small to medium-sized system (e.g., ≤ 50 nt) could be achieved well or at least reasonably with existing prediction programs, whereas predicting larger systems remains not feasible due to the versatile nature of RNA molecules. Specific examples ($\geq$ 100 nt) include a regulatory region of viral RNAs (e.g., hrv (human rhinovirus), fmdv (foot-and-mouth-disease virus)).

Predicting the 3D structures of RNA junctions, formed by multiple helical arms, is challenging because the overall configuration is often determined by the non-Watson-Crick base pairs and base stacking interactions. Furthermore, when computational prediction programs can successfully generate models that locally resemble native RNA junction structures, the accurate positioning of helical

elements in junctions is still challenging, often requiring manual manipulation.

RNA folding pathway often has functional significance as exemplified by riboswitch, regulating gene expression by ligand-induced conformational changes [189]. Although a growing amount of RNA structural data using experimental techniques have provided an excellent opportunity to study structural properties of RNAs, these studies failed to provide insights into dynamic properties; particularly, RNA junctions are dynamic structural entities undergoing conformational transitions. Computational approaches such as MD simulations to study the RNA folding pathway are rather limited only to small RNAs (e.g., hairpin) [190].

1.8 Overview of this thesis

While RNA structural elements are all important with a greater scope of larger RNA folding problems, this work focuses on various aspects of RNA junction structures and its application to predict 3D structure of viral RNAs. In Chapter 2, we begin with RNA junction structure analysis, the central motif of a larger architecture in RNA folding, using non-redundant dataset of RNA crystal structures from the PDB database; this is upon previous analyses on 3 and 4-way junctions that we extend to higher-order junctions.

Grounded in the knowledge/information obtained from the RNA junction structure analyses, we describe in Chapter 3 our novel program for predicting helical topologies of RNA junctions as tree graphs, called RNAJAG (RNA Junction-As-Graph). RNAJAG consists of two components—junction topology prediction and graph modeling—and yields fairly good representations against the helical configurations in native RNAs for a large set of 200 junctions.

With the advances in analysis, prediction, and modeling of RNA Junctions

altogether, in Chapter 4 we propose candidate RNA junction structures of regulatory regions, called internal ribosome entry site (IRES), in foot-and-mouth-disease virus (FMDV). Together with all available experimental data, we model junction topologies, build atomic 3D models, and investigate each of the candidate structures by molecular dynamics simulations to determine most energetically favorable configurations and analyze specific tertiary interactions. These collective findings, together with available experimental data, suggest a plausible theoretical tertiary structure of the apical region in FMDV IRES domain 3.

RNA junctions are dynamic, capable of undergoing conformational changes. Therefore, there is much interest in their conformational pathways. In Chapter 5, we study dynamic properties of fully base-paired 4-way RNA junctions (namely 4H junction) that are found in FMDV IRES domain 3. We suggest a potential folding pathway for interconversion between different conformations of the 4H junction.

Chapter 2

Classification of Higher Order RNA Junction Topologies

2.1 Introduction

RNA junctions are present in a wide range of RNA molecules.[1] Spatial arrangements of these secondary structures (Figure 2.1A) are involved in various biological functions that include the self-cleaving catalytic properties of the hammerhead ribozyme [147], promotion of functional folded states of the hairpin ribozyme [139], recognition of the binding pocket domain by purine riboswitches [6, 125], and translation initiation of the HCV virus at the internal ribosome entry site (IRES) [58]. Several junctions also occur within ribosomal RNA subunits [15, 106, 144] where they play important roles and often bind to ribosomal proteins [66]. Because junctions serve as major architectural features and building blocks in RNA, it is essential to better understand structural,

[1]This chapter is based on one of our published articles [72].

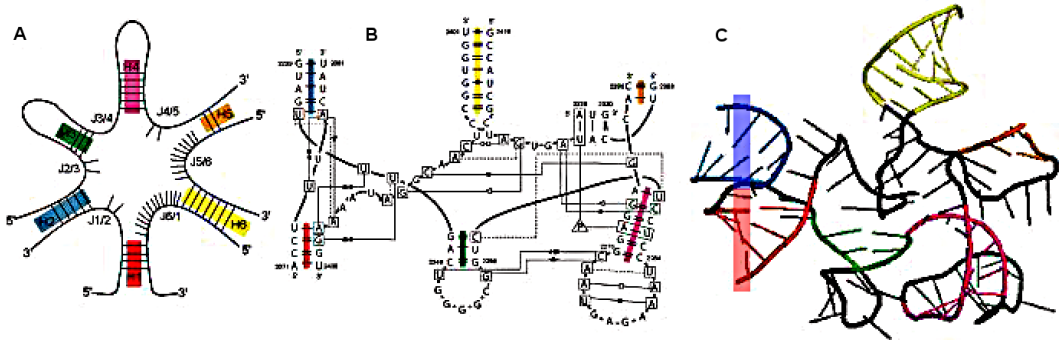


Figure 2.1: Junction architecture for *E. Coli* 23S rRNA (2AW4_2073 from Table 2.1). **(A)** Secondary structure diagram of the 6-way junction element composed of six helices labeled H_1 to H_6 (color-coded) and six loop regions labeled $J_{1/2}$ to $J_{6/1}$ with nucleotide positions marked in black. Helices and loop regions are labeled uniquely according to the $5'$ to $3'$ orientation of the entire RNA structure, by labeling H_1 as the first helix encountered while entering the junction in the $5'$ to $3'$ direction; subsequent helices within the junction are labeled as one moves along the nucleotide chain in the same direction. Lines inside the helices represent the canonical Watson-Crick base pairs G-C, A-U, and the G-U wobble base pair. **(B)** Network interaction diagram representing base pairs of the same 6-way junction according to the Leontis-Westhof nomenclature [86]. **(C)** 3D representation diagram containing coaxial stacking between helices H_1 (red) and H_2 (blue).

energetic, and dynamic aspects of these elements.

2.1.1 Earlier works of RNA junction classifications

RNA crystallography, NMR, and other experimental techniques such as fluorescence resonance energy transfer (FRET) and small-angle X-ray scattering (SAXS) have offered unprecedented opportunities to analyze RNA tertiary structure [144, 5, 14, 91, 133, 137, 140]. Such aspects have revealed structural properties of junctions; specifically, coaxial stacking of helices and long-range tertiary interactions [49, 50, 65, 141] (see Figure 2.1B and C). For instance, Lilley *et al.* [48, 88, 89] analyzed the conformations of 3 and 4-way junction examples in nucleic acids using FRET and observed transitional changes and flexibility in their helical configuration under various ionic strength of Mg^{2+} and Na^{+}. Lescoute and Westhof [87] compiled and analyzed a topological characteristics of 3-way junctions in folded RNAs, categorizing these junctions into three major families and specifying rules to predict coaxial stacking; which occurs when two separate helical regions stack to form coaxial helices as a pseudo-continuous helix (see Figure 2.1A and C). The loop connecting the stacked helices constrains the conformational space that these helical axes can explore. Laing and Schlick [73] analyzed a topology of 4-way RNA junctions and grouped them into nine major families based on coaxial stacking interactions and helical conformation signatures. Tyagi and Mathews [135] performed coaxial stacking prediction of RNAs based on free energy minimization.

2.1.2 RNA junction dataset preparation and classification

The dataset of our 3D RNA junctions as collected from the RCSB Protein Data Bank [9]. Based on available structures as of April 2009, 554 high-resolution structures were selected, with repetitions omitted by choosing the more recent structures. Junction elements were searched within these and analyzed for base pair interactions.

To perform our comprehensive search of n-way junctions ($3 \leq n \geq 10$) in the set of RNA structures above, we first considered the secondary structure associated with every 3D structure defined in terms of its canonical Watson-Crick (WC) base pairs and the single stranded regions. The search for canonical WC base pairs was performed using the program FR3D [122]. Second, we searched for sets of n distinct strands connecting in a cyclical way by at least two consecutive canonical WC base pairs (Figure 2.1A). For simplicity, pseudoknots were automatically removed during the search, but later re-inserted for statistical analysis. Visual inspection was also used to verify the correctness of our procedure. In addition, we compared our search outcome to data available from the RNAJunction database [11], to ensure the verity of all junctions.

Crystal structures containing at least one junction were identified, 43 in total. The structures include the two high resolution crystal structures of the 16S (PDB entry 2AVY and 2J00) and four 23S rRNAs (PDB entry 1NKW, 1S72, 2AW4, and 2J01). Although the 3D shape of equivalent rRNA molecules is highly conserved among species, differences are informative because they help understand evolutionary changes that *Nature* allows while keeping their molecular function intact. In total, our dataset thus contains 207 RNA junctions as

listed in Table 2.1 and Tables A.1in Appendix A and A.2.

Non-canonical base pairing with alternate hydrogen bonding patterns often occur in RNA. A consensus between FR3D and RNAVIEW [142] was considered to classify base pairs. When discrepancies occur, we employed visualization programs such as Pymol (Schrodinger, LLC) and Swiss PDB viewer [43] to check the structures further. Additionally, the junction data were analyzed from different perspectives: sequence signatures, length of loop regions, 3D motifs, and the 3D organization of their helices. Orientation aspects such as in coaxial stacking, helices that form perpendicular interhelical angles, and helices aligning their axes in parallel without the use of stacking forces were analyzed on by inspection. Pairwise structure alignment between junction domains was performed using the ARTS web server [27].

Network interaction diagrams describing base pair interactions were represented symbolically according to the Leontis and Westhof base pairing classification [83, 86]. The diagrams were created using VMD [51] and S2S [56], a visual aid program grounded in RNAVIEW.

2.1.3 Overview of Results

Our combined analyses of RNA structures have unraveled recurrent structural motifs across a variety of RNA molecules. Previous work on annotation and analysis of RNA tertiary motifs [141] based on a representative set of high-resolution RNA structures showed that coaxially stacked helices are abundant tertiary motifs that often cooperate with other tertiary motifs such as A-minor [105] for long-range interactions to stabilize RNA structure. Building upon existing work on 3 and 4-way junctions [87, 73], we extend the analysis to

Table 2.1: List of RNA 3D structures containing 23 5-way junctions, nine 6-way junctions, four 7-way junctions, one 9-way junction, and two 10-way junctions. The name describes the PDB entry and the number of the first residue of helix H_1 in the junction. The nomenclature is based on [90] and the helices are numbered according to the scheme in Leffers *et al.* [80]. Single line between rows separates junctions with the same number of coaxial helices. Double line between rows separates the junction′s degree of branching.

Name	Degree	RNA type	Coaxial stacks	Helical alignments	Nomenclature	Domain	Helix numbers
1U6B_8	5WJ	GI intron *Azoarcus*	H1H2, H3H4		$HS_1HS_2 2HS_{11}HS_2$		P2-3-8-7.1-10
1Y0Q_43	5WJ	GI intron *Twort*	H2H3, H4H5		$HS_4 2HS_{10}HS_3HS_5$		P3-4-6-7.1-7.2
1S72_238	5WJ	23S rRNA *H. Marismortui*	H1H2, H3H5		$HS_4HS_5HS_6HS_6HS_2$	I	H14-15-16-21-22
2J01_267	5WJ	23S rRNA *T. Thermophilus*	H1H2, H3H5		$HS_4HS_5HS_5HS_6HS_2$	I	H14-15-16-21-22
2BTE_6	5WJ	Leuyl-tRNA *T. Thermophilus*	H1H5, H2H3		$HS_2 3HS_2H$		H1-2-3-4-5
2NR0_6	5WJ	Leucyl-tRNA *E. Coli*	H1H5, H2H3		$HS_2HS_4HS_1HS_2H$		H1-2-3-4-5
2J01_45	5WJ	23S rRNA *T. Thermophilus*	H2H5, H3H4		$HS_6HS_4 2HS_1HS_1$	I	H4A-5-8-9-10
2AW4_46	5WJ	23S rRNA *E. Coli*	H2H5, H3H4		$HS_6HS_4 2HS_1HS_1$	I	H4A-5-8-9-10
1NKW_45	5WJ	23S rRNA *D. Radiodurans*	H2H5, H3H4		$HS_6HS_4 2HS_1HS_1$	I	H4A-5-8-9-10
3BWP_23	5WJ	GII intron *O. Iheyensis*	H2H3		$HS_1HS_2HS_2HS_3HS_3$		IA-IB-IC-ID1-I(ii)
2J00_35	5WJ	16S rRNA *T. Thermophilus*	H3H4		$HS_2HS_2HS_4HS_6HS_2$	5′	H3-4-16-17-18
2AVY_35	5WJ	16S rRNA *E. Coli*	H3H4		$HS_2HS_2HS_5HS_7HS_2$	5′	H3-4-16-17-18
1NKW_592	5WJ	23S rRNA *D. Radiodurans*	H4H5		$HS_4HS_2HS_2 2HS_5$	II	H26-27-32-36-46
1S72_640	5WJ	23S rRNA *H. Marismortui*	H4H5		$HS_4HS_2HS_2HS_4HS_{10}$	II	H26-27-32-36-46
2AW4_583	5WJ	23S rRNA *E. Coli*	H4H5		$HS_4HS_2HS_2HS_2HS_8$	II	H26-27-32-36-46
2J01_583	5WJ	23S rRNA *T. Thermophilus*	H4H5		$HS_4HS_2HS_2HS_2HS_8$	II	H26-27-32-36-46

Name	Degree	RNA type	Coaxial stacks	Helical alignments	Nomenclature	Domain	Helix numbers
1S72_657	5WJ	23S rRNA *H. Marismortui*	H4H5		$HS_2HS_3HS_5HS_1HS_3$	II	H27-28-29-30-31
2AVY_57	5WJ	16S rRNA *E. Coli*	H4H5	H2H3	$HS_2HS_6HS_1HS_1HS_3$	5'	H5-6-7-13-14
2J00_56	5WJ	16S rRNA *T. Thermophilus*	H4H5	H2H3	$HS_3HS_6HS_1HS_1HS_4$	5'	H5-6-7-13-14
1NKW_2036	5WJ	23S rRNA *D. Radiodurans*	H2H3	H4H5	$HS_9HS_8HS_{11}HS_7HS_8$	V	H73-74-89-90-93
1S72_2097	5WJ	23S rRNA *H. Marismortui*	H2H3	H4H5	$HS_6HS_8HS_9HS_4HS_4$	V	H73-74-89-90-93
2AW4_2056	5WJ	23S rRNA *E. Coli*	H2H3	H4H5	$HS_6HS_8HS_9HS_4HS_4$	V	H73-74-89-90-93
2J01_2053	5WJ	23S rRNA *T. Thermophilus*	H2H3	H4H5	$HS_9HS_8HS_{11}HS_7HS_8$	V	H73-34-89-90- H93
2A64_20	6WJ	RNase P type B *B. Stearothermophilus*	H1H2, H5H6		$HS_1HS_{13}HS_3HS_2HS_8$		P2-3-5-15-15.1-15.2
1NKW_2056	6WJ	23S rRNA *D. Radiodurans*	H1H2	H3H6	$HS_2HS_2HS_2HS_{10}HS_{13}$	V	H74-75-80-81-82-88
2J01_2073	6WJ	23S rRNA *T. Thermophilus*	H1H2	H3H6	$HS_2HS_2HS_2HS_{10}HS_{13}$	V	H74-75-80-81-82-88
1S72_2114	6WJ	23S rRNA *H. Marismortui*	H1H2	H3H6	$HS_2HS_3HS_2HS_2HS_{10}HS_{10}$	V	H74-75-80-81-82-88
2AW4_2073	6WJ	23S rRNA *E. Coli*	H1H2	H3H6	$HS_2HS_2HS_2HS_{10}HS_{13}$	V	H74-75-80-81-82-88
1S72_38	6WJ	23S rRNA *H. Marismortui*	H2H3		$HS_2HS_4HS_{11}HS_3HS_3HS_9$	I	H4-4A-11-12-13-14
2J01_43	6WJ	23S rRNA *T. Thermophilus*	H2H3		$2HS_4HS_{11}HS_3HS_3HS_9$	I	H4-4A-11-12-13-14
2AW4_44	6WJ	23S rRNA *E. Coli*	H2H3		$2HS_4HS_{11}HS_3HS_3HS_9$	I	H4-4A-11-12-13-14
1NKW_43	6WJ	23S rRNA *D. Radiodurans*	H2H3		$2HS_4HS_{11}HS_3HS_3HS_9$	I	H4-4A-11-12-13-14
1NKW_829	7WJ	23S rRNA *D. Radiodurans*	H2H3, H6H7	H1H4	$HS_42HS_2HS_4HS_3HS_3HS_7$	II	H36-37-38-39-40-41-45
2AW4_816	7WJ	23S rRNA *E. Coli*	H2H3, H6H7	H1H4	$HS_4HS_2HS_5HS_4HS_32HS_4$	II	H36-37-38-39-40-41-45
2J01_816	7WJ	23S rRNA *T. Thermophilus*	H2H3, H6H7	H1H4	$HS_42HS_2HS_5HS_32HS_4$	II	H36-37-38-39-40-41-45
1S72_909	7WJ	23S rRNA *H. Marismortui*	H2H3, H6H7	H1H4	$HS_4HS_2HS_5HS_4HS_32HS_4$	II	H36-37-38-39-40-41-45
2J01_6	9WJ	23S rRNA *T. Thermophilus*	H4H5		$HS_7HS_3HS_{18}HS_7HS_{11}HS_2HS_4HS_{19}HS_5$	I	H1-2-25-26-47-72-73-94-99
1NKW_7	10WJ	23S rRNA *D. Radiodurans*	H4H5, H9H10		$HS_8HS_9HS_{18}HS_7HS_{11}HS_2HS_4HS_2HS_4HS_4$	I	H1-2-25-26-47-72-73-94-98-99
2AW4_7	10WJ	23S rRNA *E. Coli*	H4H5, H9H10		$HS_6HS_7HS_{18}HS_7HS_{11}HS_2HS_4HS_2HS_5HS_5$	I	H1-2-25-26-47-72-73-94-98-99

higher order junctions (5 to 10-way junctions) and combine our findings to describe common tertiary motifs, including recurrent helical configurations; they occur across all junctions found in solved structures, regardless of their degree of branching. Our analysis reveals novel tertiary interaction motifs formed between perpendicular alignments of helices as well as common internal base pairs that help form long-range interactions. We also discuss how RNA junctions arrange their helical arms in similar configurations, regardless of their degree of branching. Statistical data showing preferred base pair and base stacking interactions are also reported.

2.2 Results

Network interaction diagrams (see Figure 2.1B) indicating base pair interactions have proven useful in understanding RNA tertiary motifs [81, 82, 87] and in investigating the topology of 3 and 4-way junctions [87, 72]. Here, we extend such analyses to higher order junctions ranging the degree of branching helices from 5 to 10. We begin with a description of the higher order junctions using network interaction diagrams. For clarification, we label and color code helices sequentially according to the $5'$ to $3'$ orientation of the entire RNA as shown in Figure 2.1A. We define a helix when at least two consecutive Watson-Crick base pairs—G-C, A-U and G-U—are present. The single stranded region between each pair of consecutive helices H_i and H_{i+1} is labeled by $J_{i/i+1}$. Each junction element is labeled by its Protein Data Bank (PDB) entry [9] followed by the first residue number of the first helix H_1 in the junction. The point where strands cross over is called the point of strand exchange. We use the Leontis and Westhof notation [83, 86] to study base pair interactions occurring within

junctions and to describe common motifs. Our list of 207 junctions contains junctions of degree 3 to 10 (see Table 2.1 and Tables A.1 and A.2, Appendix A) and are assembled by taking all high-resolution RNA structures from the PDB database [9] as of April 2009.

In our previous analysis of 4-way RNA junctions [73], we identified nine broad 4-way junction families according to coaxial stacking patterns and helical configurations (Figure 1.3B in Chapter 1). Helices within these junctions stabilize their conformations using common tertiary motifs like coaxial stacking, loop-helix interaction, and loop-loop interactions. Novel interactions involving A-minor motifs and coaxial stacking of helices were observed at the point of strand exchange in many 4-way junctions within the three families *cH*, *cL* and *cK* (Figure 1.3B in Chapter 1). In our analysis of higher order junctions, we find more disorder in the organization of their components. Still, similar to 3 and 4-way junctions, helices tend to arrange locally in parallel and perpendicular patterns. In addition, similar motifs such as the A-minor interactions and the sarcin/ricin like motifs [105, 85] are also commonly encountered.

2.2.1 Higher order junction classification

Due to the small number of examples available for higher order junctions (Figure 2.2), it is not possible to design a classification scheme similar to the families assigned in junctions with relatively small degrees [87, 73]. However, a number of recurrent interaction patterns and motifs can be observed, and their helical elements can be organized using coaxial stacking patterns and other helical arrangements as described below.

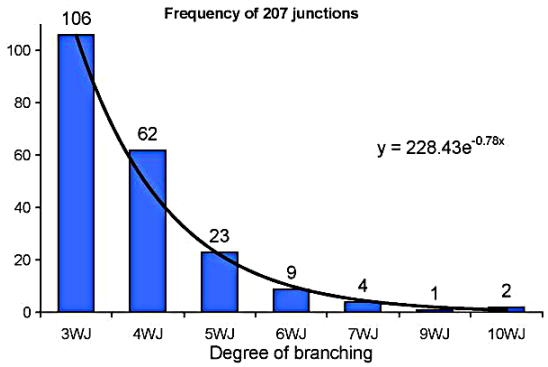


Figure 2.2: Histogram from a total of 207 RNA junctions sorted by degree of branching helices ranging from 3 to 10.

5-way junctions

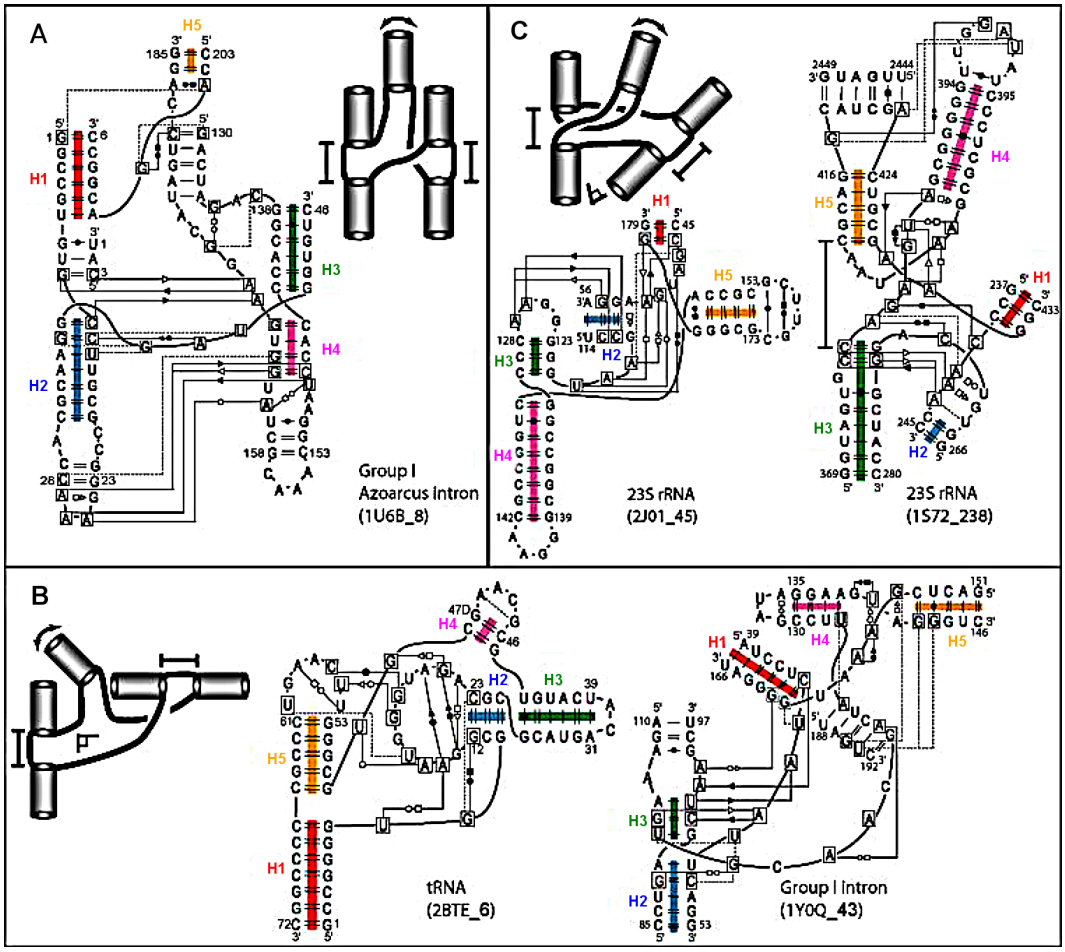
5-way junctions resemble lower-order junctions in terms of their helical arrangements. For instance, Figure 2.3A-C shows junction diagrams with two coaxial stacking interactions (seen as aligned colored helices) analogous to families in 4-way junctions [73]. Specifically in Figure 2.3A, a junction found in the *Azoarcus* intron [1] contains all its helical axes aligned roughly in a coplanar and parallel arrangement and stabilized by long-range interactions, forming a crossing at the point of strand exchange similar to elements in the 4-way junction family *cH*. A-minor interactions [105] (denoted by empty and solid triangles known as Sugar-Sugar interactions) are the most conserved interactions responsible for such crossings. Similarly, the junction 2BTE_6 in Figure 2.3B corresponds to the transfer RNA, where four helices form the well known "L" shape while an extra helix bulges out of the "L" shape. Also of interest, both Figure 2.3B and C contain junction examples with a pair of perpendicular coaxial stacking interactions. While the pattern in Figure 2.3B is a coaxial stacking produced

between consecutive helices, that in Figure 2.3C is a coaxial stacking between pairs of non-consecutive helices (H_2H_5 and H_3H_5 for each case). Thus, coaxial stacking interactions are not exclusively formed between neighboring helices.

Figure 2.3D-F shows junction diagrams with one coaxial stacking perpendicular to at least one helix. Specifically, Figure 2.3D illustrates a junction with one coaxial stack and one helical alignment (helices aligned without stacking interactions) arranged in a perpendicular configuration. As observed in 3 and 4-way junctions, such perpendicular arrangements among helices are stabilized by loop-loop interactions (2BTE_6 in Figure 2.3B and 2AVY_57 in Figure 2.3D), loop-helix interactions (2J01_45 in Figure 2.3C) or helix-helix interactions (1S72_657 and 2AVY_35 in Figure 2.3F). Loop-loop interactions typically involve Hoogsteen or Sugar edge interactions, but can also involve WC base pairs. Loop-helix interactions primarily involve Sugar-Sugar interactions forming A-minor motifs. Helix-helix interactions involve minor groove interactions and will be discussed below in more detail. Junction diagrams in Figure 2.3F resemble family *cK* of 4-way junctions, which are composed of one coaxial stacking between two helices, while a third helix aligns perpendicular to the coaxial stack. The remaining two helices are arranged based upon the length of their flanking loop elements.

6 to 10-way junctions

In contrast to the compact globular shapes that many protein structures have, RNA molecules prefer rather compact prolate ellipsoidal shapes [5, 132]. This property reflects the way junctions form by keeping most of their helical axes roughly coplanar. Compared to junctions with a low degree of branching, higher order junctions are more disordered in the organization of their components;


A
H5
H1
H2
H3
H4
Group I
Azoarcus Intron
(1U6B_8)
C
23S rRNA
(2J01_45)
23S rRNA
(1S72_238)
B
tRNA
(2BTE_6)
Group I Intron
(1Y0Q_43)

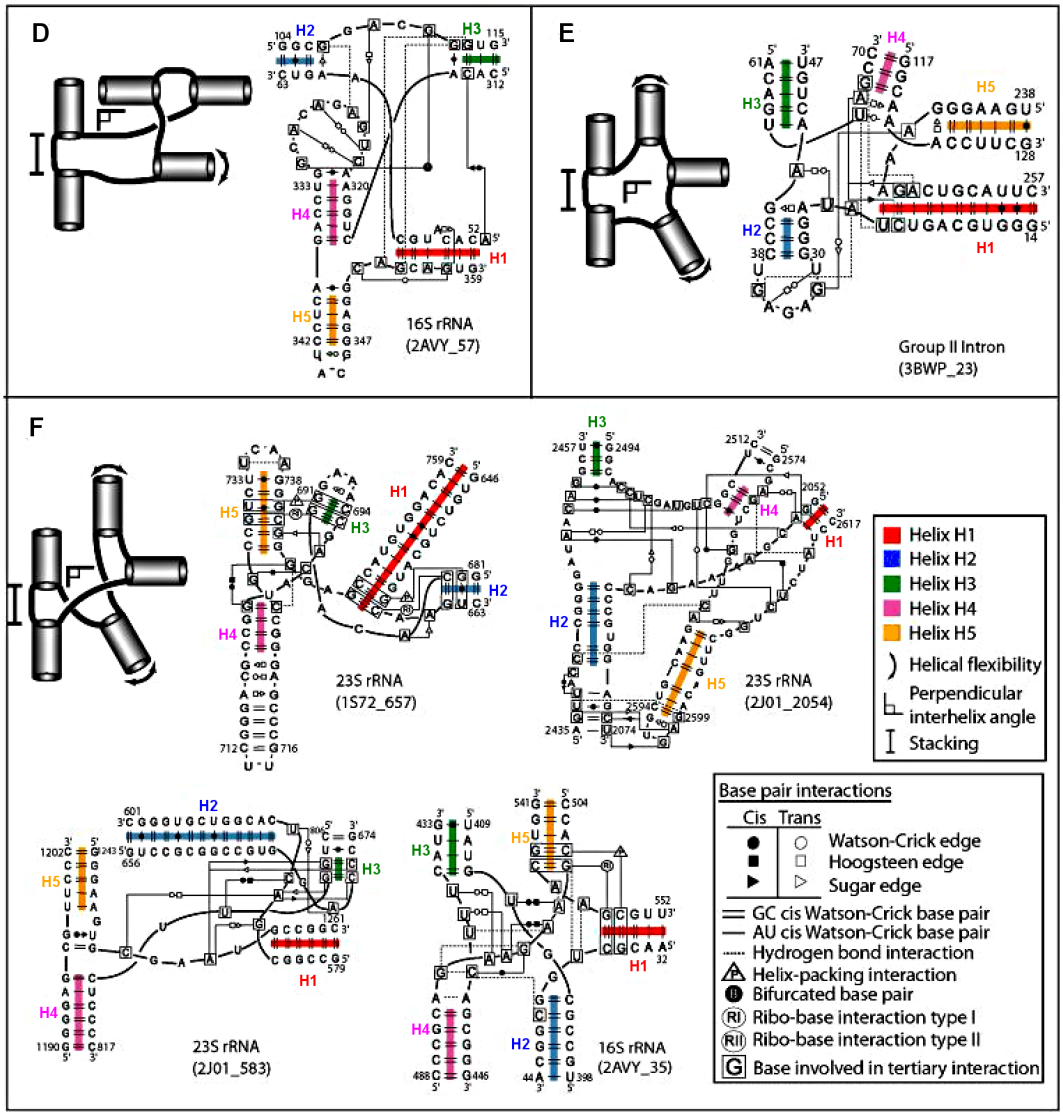


Figure 2.3: Network interaction diagrams of 5-way junctions sorted by coaxial stacking and helical configurations. The network symbology follows the Leontis-Westhof notation [86] (see inset boxes). Figures **A-C** show junction diagrams with two coaxial stacks aligned either in parallel **(A)** or perpendicular to each other **(B-C)**, while figures **D-F** portraits junction diagrams with one coaxial stack perpendicular to at least one helical arm.

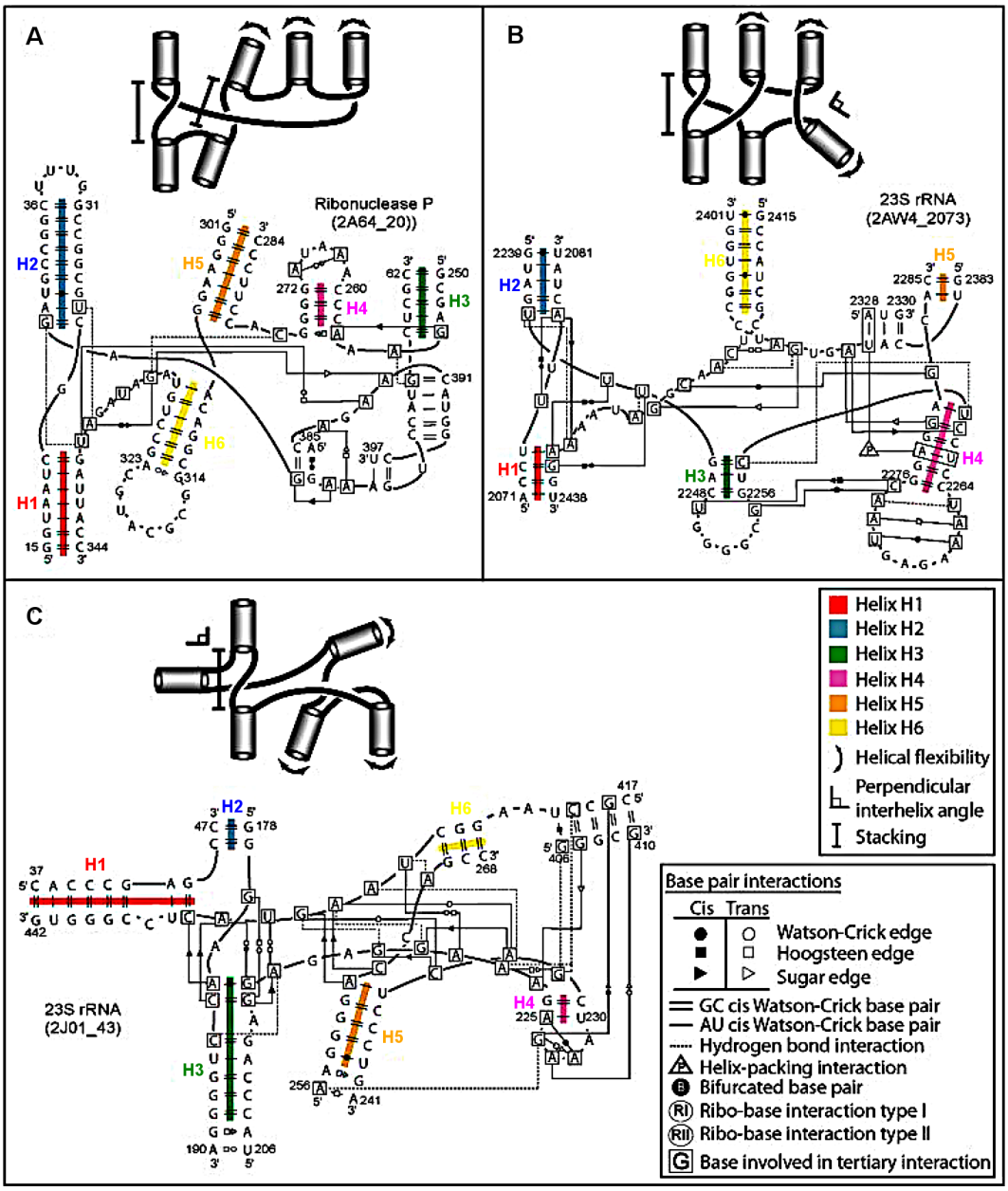


Figure 2.4: Network interaction diagrams of 6-way junctions sorted by coaxial stacking and helical configurations. The network symbology follows the Leontis-Westhof notation [86] (see inset boxes). **(A)** Junction diagram with two coaxial stacks, and **(B-C)** portraits junction diagrams with one coaxial stack and one perpendicular helical alignment.

still, the basic helical arrangements such as coaxial stacking (present in every high order junction), parallel, and perpendicular helical axes are retained, as described next.

Figure 2.4A shows a 6-way junction from the ribonuclease P, forming a coaxial helix H_1H_2 and helices H_3H_4 in a plane, while the coaxial helix H_5H_6 leaving this plane. The conformation produced by coaxial helices H_1H_2 and H_5H_6 is similar to the antiparallel conformation found in the 4-way junction in the hairpin ribozyme [115]. The diagram in Figure 2.4B shows a 6-way junction with the helical axes in a plane. The single strand $J_{5/6}$ contains nucleotides 2385-2387 base pairing with a hairpin loop, forming a (pseudoknot) helix perpendicular to H_4. The homologous 6-way junctions found in the *H. Marismortui* and *T. Thermophilus* (1S72_2114 and 2J01_2073 in Table 2.1) shows helices H_3 and H_6 aligned.

In Figure 2.4C, the 6-way junction 2J01_43 contains helices H_1-H_3 arranged by forming a coaxial helix H_2H_3 which aligns perpendicular to H_1, in a similar conformation to members of family *A* in 3-way junctions such as the M-box riboswitch (2QBZ_53 in Table A.1, Appendix A), and 3-way junctions found in the large ribosomal subunit (1S72_51, 1S72_1403 and 1S72_2130 in Table A.1, Appendix A).

The 7-way junction in Figure 2.5A is formed by three coaxial helices aligning their axes more or less in a plane. The coaxial stacking between non-neighboring helices H_1 and H_4 is due to a sarcin/ricin motif [85] formed between strands $J_{1/2}$ and $J_{7/1}$. The pair of coaxial helices H_2H_3 and H_1H_4 aligns similar to family *cH* in 4-way junctions [73], where a crossing occurs at the point of strand exchange caused by A-minor interactions. At the same time, the pair of coaxial helices H_1H_4 and H_6H_7 aligns similar to 4-way junction family *H* with its extra helix

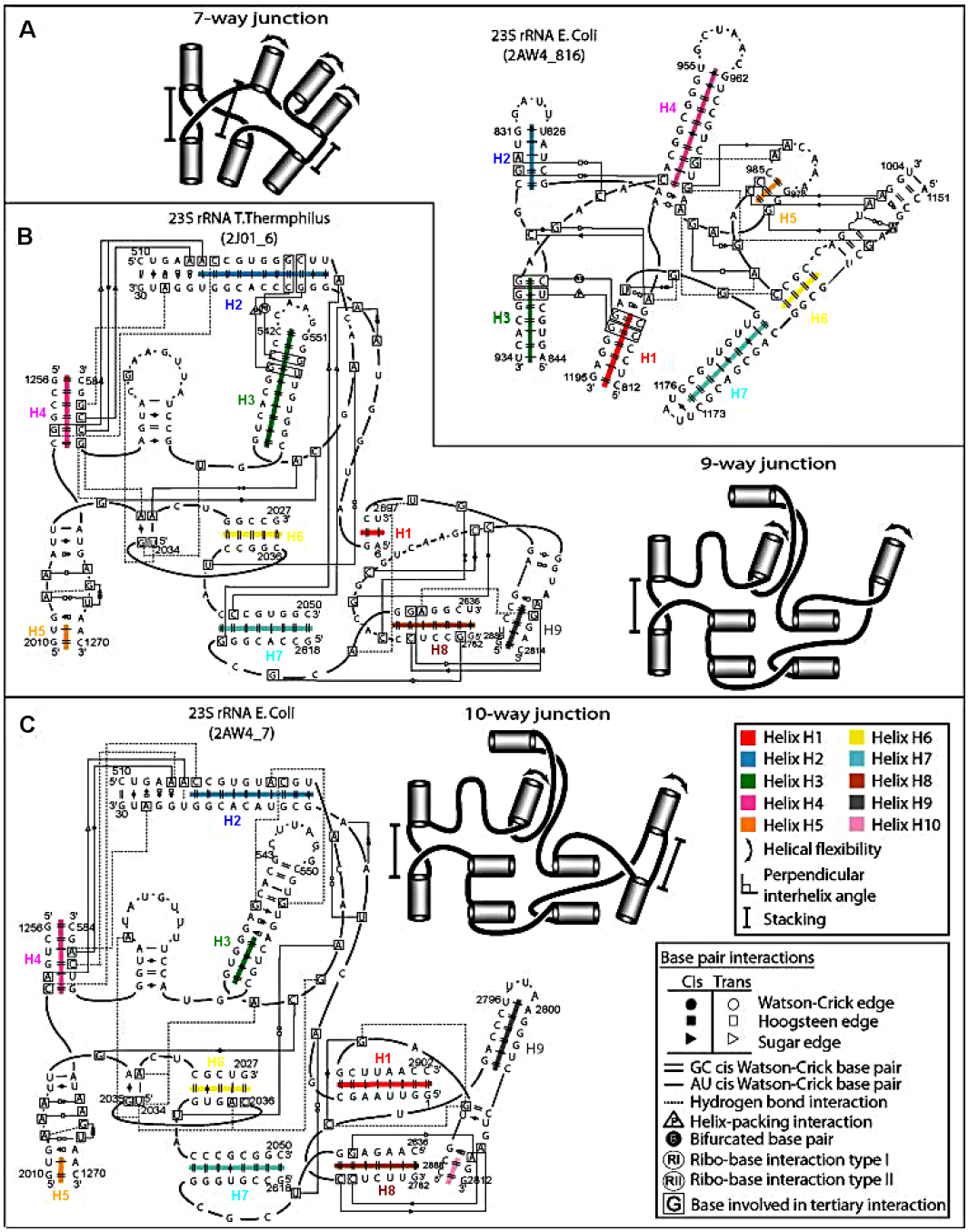


Figure 2.5: Network interaction diagrams of 7 to 10-way junctions. **(A)** 7-way junction, **(B)** 9-way junction and **(C)** 10-way junction. The network symbology follows the Leontis-Westhof notation [86] (see inset boxes).

H_5 in between. Helices H_1 and H_3 arrange perpendicular to each other.

The 9 and 10-way junctions shown in Figure 2.5BC correspond to the central junction connecting all domains in the 23S rRNA. The 10-way junction contains an extra helix presumably formed through evolutionary variation. Note that in both cases the strand $J_{3/4}$ forms a "helical region" composed of alternating canonical and non-canonical WC base pairs. Our definition of a helix requires at least two consecutive WC base pairs to be formed; therefore, this region is considered as a strand. Both junctions are non-planar due to the high degree of branching and form three small globular helical regions. The first region is composed of helices H_1 and H_8-H_9 (and H_{10} for the 10-way junction) arranging similarly to family *cK* in 4-way junctions [73]. Helices H_2, H_3, H_6, and H_7 align similar to family *X* in 4-way junctions. The third region is the coaxial helix between H_4 and H_5.

Another common characteristic of higher order junctions is that long single stranded elements occur to reduce steric clashes caused by junctions with many helical arms, while preserving the preferred prolate and ellipsoid shapes of RNA 3D structures. The single strands connecting two helices often traverse or "jump over" a third helix in between as it occurs in the strand $J_{3/4}$ shown in Figure 2.4C. Moreover, these single strands interact with several junction components while traversing as in the example 2AVY_35 in Figure 2.3F. Here the strand $J_{4/5}$ connecting helices H_4 (magenta) and H_5 (orange), interacts with $J_{3/4}$ and with itself, then interacts with $J_{2/3}$ and finally with $J_{5/1}$. These longer strands between helices allow frequent formation of pseudoknots (Figure 2.4AB and Figure 2.5BC). Other properties of higher order junctions that are shared by junctions with lower degrees are described in the following sections.

2.2.2 Common statistical features in RNA junctions

From our dataset of 207 RNA junctions listed in Table 2.1 and Tables A.1 and A.2 in Appendix A, more than half are 3-way junctions, and the number decreases as the degree of branching increases. Figure 2.2 shows that the frequency of junctions arranged by degree of branching can be estimated by the exponential function y= $228.4e^{0.78x}$ ($R^2 = 0.94$), but it is not clear how this estimate will change with increased RNA structures. Junctions of higher degree of branching are observed in large RNAs such as the ribonuclease, group II intron, and ribosomal RNA. In contrast, junctions with a small degree of branching occur in a wide range of RNAs, from riboswitches to ribosomal RNAs.

The single stranded loop regions connecting helical elements in junctions are composed by uneven proportions of nucleotide composition as shown in Figure 2.6A. While a low percentage of Cs (14%) can be noted, loop regions are strikingly A-rich (40%) for two reasons: A-minor interactions are important in stabilizing helical arms, and adenines offer flexibility to the loop regions. Conversely, the lower concentration of Cs in loop regions corresponds to the smaller number of non-canonical WC base pairs known involving cytosine; however, a reasonable number of these Cs (14%) participate in pseudoknot formation or WC G-C base pairs between loops within the junction. In addition, the concentration of WC base pairs near the end of helices (first and second position) produce a high concentration of G-C (73%) base pairs, compared to lower A-U (20%) and G-U wobble (7%) base pairs (data not shown); this might be explained by the high stability (3 hydrogen bonds) of G-C base pairs.

Figure 2.6B describes the distribution of the loop size for all loops within helical junctions (blue), loops between coaxial helices (stacked loop, shown in

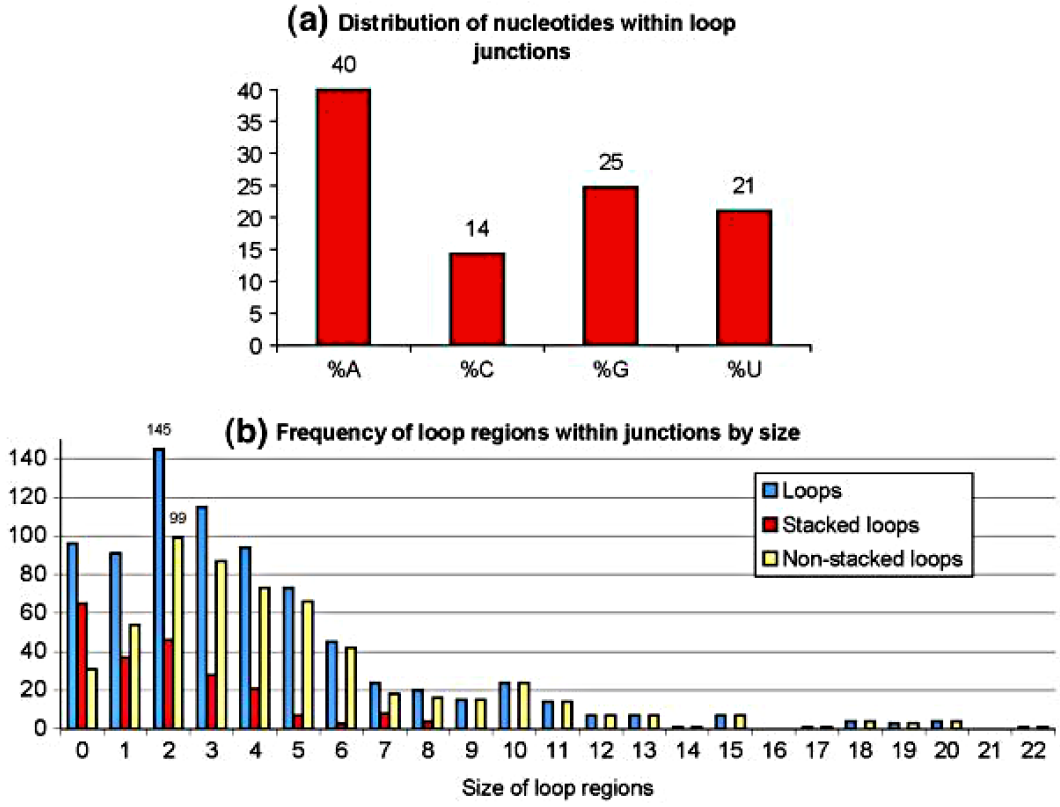


Figure 2.6: Junction statistics. **(A)** Proportion of nucleotides at the single stranded loop elements within junctions. **(B)** Frequency distribution of loop regions within junctions arranged by size. Values for any loop, for loops between coaxial stacking and for loops between helices forming no coaxial stacking are given in blue, red and yellow respectively.

red), and loops between helices where no coaxial stacking is present (unstacked loops shown in yellow). In general, a large number of loops range in size from 0 to 6 with a peak at 2, while the less frequent cases are loops of sizes 14 to 22. Figure 2.6B also shows (in red) that coaxial stacking occurs preferentially in helices adjacent to loops of smaller size, and no stacking is observed for helices between loops of size greater than 8. Coaxial stacking of helices adjacent to loops of size 6 or 7 occurs often due to many non-canonical base pair interactions, which in turn stack with such helices, or also due to the presence of pseudoknots forming at the loop. While a preference for coaxial stacking formation between loops of small size can be noted, there are several cases in which helices with a small loop size do not stack. Particularly, Figure 2.6B shows a peak at 2 corresponding to loops between unstacked helices (99 out of 143). Many reasons could explain the absence of coaxial stacking in these cases, for example the influence of external forces such as pseudoknot formation, long-range tertiary interactions, and protein binding.

In agreement with the work by Elgavish *et al.* [29], non-canonical base pairs involving A-G occur frequently at the end of helices, particularly a *trans* A-G Hoogsteen-Sugar or *cis* A-G Watson-Watson base pairs. These, along with standard Watson-Crick G-C base pairs forming a pseudoknot, are the most frequent interactions observed at the end of helices in junctions. When a non-canonical base pair A-G *trans* Hoogsteen-Sugar is formed, it often stacks to a *trans* A-U Hoogsteen-Watson base pair. These two base pairs are recurrent interactions observed in many junctions and become parts of larger 3D motifs such as the sarcin/ricin [83, 85] or UA-handle motifs [54]. But they can also form as independent and stable sub-motifs, often binding to RNA or proteins, and assisting in the formation of coaxial stacking between helices.

Other important base pair interactions found in junctions are the Sugar-Sugar base pairs, which can form A-minor motifs [105] and often combine with coaxial helices forming higher order motifs [141] (A-minor/coaxial helix). In addition, when long-range interactions occur in junctions, a vast majority of A-minor motifs are formed between loop regions flanking helices (e g., hairpins and internal loops), while the helical receptors are located near the end of helices [141]. Other base pair interactions also occur and are composed mostly of purine-purine interactions. Long-range interactions such as A-minor are important elements because they stabilize helical arms in junctions and allow the proper function of RNA molecules.

2.2.3 Novel tertiary motif for perpendicular helical arrangements

One of the most common elements in the ribosome, highlighted by the structural biologists, is the interaction of RNA double helices via minor grooves. Examples of such interactions are A-minor [105], ribose zipper [131], G-ribo [130] and along-groove packing motif (AGPM) [35, 36], also known as p-interaction [103]. The interactions presented here describe yet another strategy used for packing minor grooves of rRNA helices against each other.

Helices in junctions often align their axes more or less perpendicular to each other via helix-helix interactions along their minor grooves (Figure 2.7A). Because the minor groove in A-RNA has a slightly concave shape, the sugar-phosphate backbone of each helix can pack along the minor groove of the other helix. We previously reported perpendicular interactions in 4-way junctions where the AGPM motif is present [73] (WC G-U interaction in blue shown

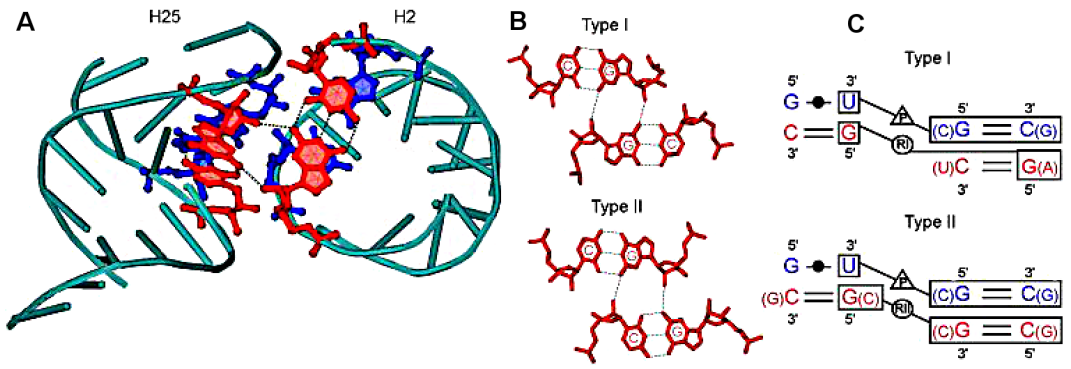


Figure 2.7: **(A)** Perpendicular alignment between helices H_2 and H_{25} in the *T. Thermophilus* 23S rRNA structure (PDB entry 2J01). Residues in blue correspond to the AGPM motif (G_{539}-U_{554} and G_{17}-C_{523}) and residues in red correspond to the ribo-base interaction type I (C_{540}-G_{553} and C_{18}-G_{522}). **(B)** Ribo-base interaction types I and II. **(C)** Consensus motif for the perpendicular interaction of helices composed of four stacked base pairs at each helix.

in Figure 2.7A). A full analysis based on all junctions allows us to recognize two new interactions which often cooperate with AGPM motifs. The combined interactions are composed of four WC base pairs, forming an angle of approximately 60° between their corresponding base pair planes, and occurring when helices are closely packed. Because these new interactions involve ribose-base interactions, we denote them as ribo-base type I (RI) and ribo-base type II (RII) interactions (see Figure 2.7).

The ribo-base type I is characterized by a 2-fold symmetry between two canonical WC base pairs connected by hydrogen bonds interactions between the O_2' of a G residue of the first base pair, and an N_2 of a G (or N_3 of an A) residue of the second base pair, and between O_2' of a G (or A) residue of the second base pair, and N_2 of a G residue of the first base pair (see Figure 2.7B).

Ribo-base type I occurs between a G residue of the first base pair and a purine (A or G) of the second base pair. Interestingly, when it appears next to an AGPM motif, a WC C-G base pair appears stacked below the WC G-U wobble base pair. Indeed, this base pair signature is even more conserved than the WC G-C receptor of the G-U wobble in the AGPM motif (Table A.3, Appendix A).

The ribo-base type II consists of a roto-reflection symmetry (rotation by 180° followed by a reflection around its axis) where two WC base pairs interact by hydrogen bonds between the O'_2 of a G residue of the first base pair with an N_2 of a G (or N_3 of an A) residue of the second base pair, and between O'_2 of a C or U residue of the second base pair with N_2 of a G residue of the first base (Figure 2.7B). When appearing next to the AGPM motif, the C-G base pair stacked below the G-U base pair can be replaced by a G-C base pair, as long as a substitution from C-G to G-C (or A-U) on the receptor base pair of the second stack occurs (see Table A.3, Appendix A).

We found 45 instances of ribo-base interactions, mostly located in homologous regions of the ribosomal RNAs considered, and most of them form next to the AGPM motif. While most cases occur between helical elements in junctions, other instances also occur in pseudoknots or near internal loops. Sequence and secondary structure signature consensus elements for these motifs are shown on Figure 2.7C, where the ribo-base interactions appear next to AGPM. There are, however, cases where AGPM motifs with no ribo-base interaction appears or ribo-base interactions in non-AGPM patterns. These cases usually occur when WC base pairs are replaced by other base pairs such as *cis* Watson-Watson AG, or when the G-U wobble is replaced by a WC A-U base pair (Table A.3, Appendix A). Furthermore, crystallographic data from a hammerhead ribozyme (PDB entry 1HMH) and tRNA-Gly (PDB entry 1VAL) shows type I and II

interactions forming between a pair of helices which are possibly tightly packed during to the formation of the crystal.

Analogous to AGPM motifs [35, 36, 103], ribo-base interactions bring together helical elements and stabilize RNA molecule for proper function. Another possible role is to act as a mechanism for promoting RNA-protein interactions of neighboring purine nucleotides. Klein and coworkers [144] reported that proteins L18e and L15 in the *H. Marismortui* have a high structural homology in the C-terminal domains and both interact with the 5-way junction 1S72_657 (Figure 2.2F), forming a near identical nucleotide and amino acid composition. Both proteins L18e and L15 each interact near ribo-base interactions type I (C_{658}-G_{747} with C_{685}-G_{661}, and C_{696}-G_{689} with C_{741}-G_{730} respectively). A close examination of both cases reveals purine bases that expose their hydrophobic surfaces at the protein-RNA interaction site. In other instances, when pairs of helices are closely packed through AGPM and ribo-base interactions, the AGPM/ribose-base motif appears near the end of helices flanking a *trans* AG Hoogsteen-Sugar base pair interactions. This allows these purine bases to expose their hydrophobic surfaces for possible RNA-protein interactions.

2.2.4 Folding similarity among different degree of junctions

With the available 3D structures of large RNA molecules such as ribosomal RNAs [15, 106, 144], group I introns [1, 40, 44] and RNase P structures [57, 70], it is now evident that there is a high degree of structural conservation in tertiary structures between homologous RNAs. This fact reflects the similarity among junction architectures despite differences in secondary structure. For instance,

Krasilnikov and coworkers [70] reported 3D structural similarities in the S domain of RNase P between an internal loop in RNase P type A and a 4-way junction in RNase P type B. Also, most transfer RNA structures are composed of a 4-way junction (e.g. 1EFW_6 in Table A.2, Appendix A), but the example shown in Figure 2.3B illustrates a tRNA with a 5-way junction conformation. Another interesting example is found in the group I introns (see Figure 2.7A), where a 3-way junction (1U6B_45 in Table A.1, Appendix A) in the *Azoarcus* intron [79] and a 5-way junction (see 1Y0Q_43 in Table 2.1 and Figure 2.3B) in the *Twort* intron [40] align their corresponding helices P3, P4 and P6 with a high degree of similarity (RMSD 1.09Å) despite differences in their secondary structure. This structural similarity is in agreement with the observations that group I introns contain conserved core elements formed by junctions, which provide structural stability with the help of conserved peripheral elements by forming long-range contacts [80].

Moreover, the modular architecture of folded RNAs implies that distances between interacting parts are conserved in functionally homologous molecules [87]; thus, similarities in junctions can be made apparent by observing network interaction diagrams and their 3D motifs. For example, in the large subunit of the ribosomal RNA, a 5-way junction in *H. Marismortui* (see 1S72_657 in Table 2.1 and Figure 2.3F) is structurally similar to the 4-way junctions found in homologous counterparts in *T. Thermophilus*, *E. Coli* and *D. Radiodurans* (2J01_600, 2AW4_600, and 1NKW_608 in Table A.2, Appendix A). In all cases, four helices interact in pairs via perpendicular motifs caused by ribo-base interactions with AGPM. Similarly, the core junctions whose diagrams are shown in Figure 2.5BC present a highly conserved structural similarity between the 9-way junction found in the *T. Thermophilus* and the 10-way junctions found in both

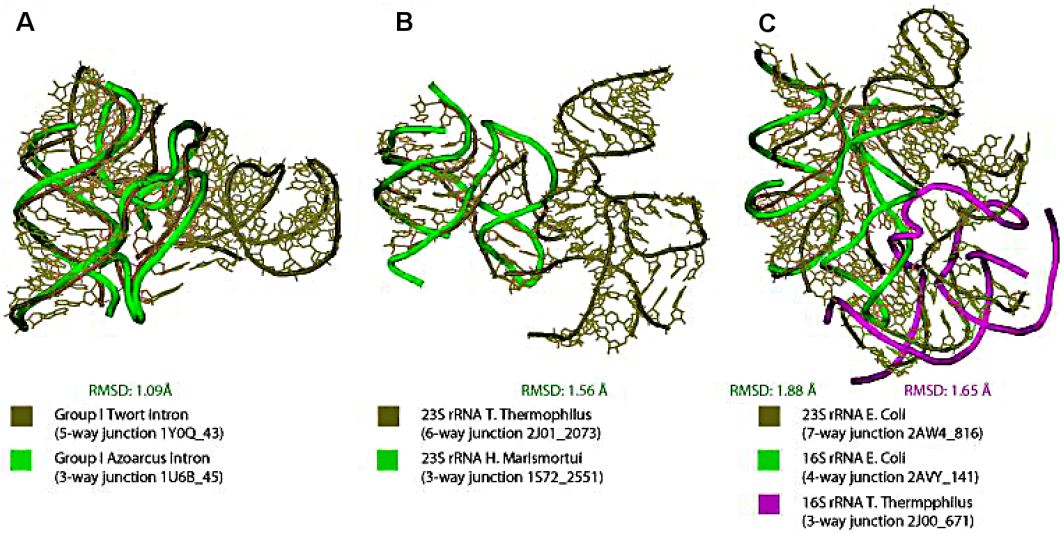


Figure 2.8: Structural similarity between **(A)** homologous and **(B-C)** non-homologous junctions. **(A)** Alignment between the *Azoarcus* intron (olive green) and the *Twort* intron (bright green). **(B)** A 6-way junction (olive green) in the 23S rRNA presenting structural similarity to a 3-way (bright green) in the 16S rRNA. **(C)** A 7-way junction (olive green) in the 23S rRNA presenting structural similarity to a 3-way (magenta) and 4-way junction (bright green) in the 16S rRNA.

the *E. Coli* and *D. Radiodurans*. These observations suggest that the extra helices that are "left out" might have formed later in evolution for particular advantages in species.

Strikingly, a structural similarity of junctions with diverse degree of branching was also observed in non-homologous elements where junctions with a larger degree of branching arrange their helical elements to form "sub-junctions" of smaller degrees. For instance, the 6-way junction 2J01_2073 arranges helices H_1, H_2 and H_3 locally similar to 3-way junctions of the C family. Elements in family C consist of one coaxial stacking, and a helix aligning parallel to the

coaxial helix, by allowing the single strand connecting the coaxial helix to the parallel helix to structure like a hairpin using the standard U-turn. The 6-way junction also forms a U-turn hairpin within the loop $J_{6/1}$ between helices H_1 and H_6. Figure 2.8B shows a pairwise structural alignment (RMSD 1.56Å) between this 6-way junction and the 3-way junction 1S72_2551 (Table A.1, Appendix A) of the family *C*. Similarly, the U-turn hairpin motif is also found in the 4-way junction 2AW4_1832 (Table A.2, Appendix A) within the loop $J_{3/4}$, forming a sub-3-way junction element between helices H_2, H_3 and H_4 (helices also labeled 65-67 by Leffers *et al* [79]). Another example is found in helical elements H_1-H_4 in the 7-way junction, shown in Figure 2.5A, which can be decomposed into a 4-way junction of the *cH* family [73] while helices H_5-H_7 can be associated to a 3-way junction of the *C* family [87] as observed in Figure 2.8C. Here, both the 4-way junction 2AVY_141 from Table A.2 in Appendix A and the 3-way junction 2J00_671 from Table A.1 in Appendix A superimposed with the 7-way junction 2AW4_816 (RMSDs 1.88Å and 1.65Å respectively).

2.3 Discussion

RNA junctions are important structural elements that serve as major architectural components in RNA. While most junctions found in solved crystal structures are formed by a small number of helical branches, higher order junctions with as many as 10 branching helices exist. Junctions organize their helical elements using various common interactions, such as long-range interactions, coaxial stacking, and many 3D motifs.

Our analysis of higher order junctions using network interaction diagrams is a complementary and compatible approach to the classification of RNA 3 and 4-

way junctions given by the Westhof [87] and Schlick [73] groups, which organize elements according to their helical configurations. Our work also complements other studies. For instance, the SCOR [67] database lists examples of coaxial helices as elements of tertiary motifs. Similarly, RNA junctions contained in the RNAJunction [11] database have been grouped by standard nomenclature [90] based on the size of each loop region. However, similar junctions from homologous RNAs can differ by single insertions of deletions in the loop regions, leading to different classifications under the standard nomenclature.

In the present analysis, we considered higher order junctions from 5 to 10 helices, and compared coaxial stacking and base pair configuration properties to those noted in lower order junctions. We described statistical properties of helices and loop regions for all these RNA junctions and introduced a new motif composed of ribo-base interactions and the AGPM, which is involved in perpendicular helical arrangements. We noted the folding similarity that exists among junctions with different degrees of branching.

In agreement with previous works [87, 73, 135], the data from Figure 2.6B indicate a preference for coaxial stacking formation for helices whose common single-stranded loop is small in size. However, there are several cases where helices with a small loop between them do not stack. The reasons for the absence of coaxial stacking are diverse. Often, elements in the loop regions within junctions form non-canonical base pairs, which in turn can help reduce the spatial distance between helical arms and facilitate coaxial stacking. In many cases of the family C of the 3-way junctions [87], a small U-turn motif forms at the end of a helix [87], possibly preventing a coaxial stacking on the caped helix. In addition, proteins can disrupt coaxial helices when their presence alters helical orientations. The 4-way junction 1S72_1743 (Table A.2, Appendix A) found

in the *H. Marismourtui* 23S rRNA, contains a pair of helices (labeled 62-63 by Leffers *et al* [79]) with no single stranded region between them, but the helices are distorted by the protein L19e, thus preventing the formation of coaxial stacking.

Furthermore, in some cases, even if the size of a loop $J_{i/i+1}$ is small, the size of neighboring loops $J_{i-1/i}$and $J_{i+1/i+2}$ can be equal or smaller, as observed in elements of 4-way junction families *H* and *cH* [73] (Table A.2, Appendix A). This can lead to an interconversion of stacking conformers or to a competition for coaxial stacking conformers, which can ultimately be decided by long-range interactions. Indeed, experiments for the hammerhead ribozyme [110] and hairpin ribozyme [137] have shown that loop-loop interactions act as important elements in the function of these ribozymes, by stabilizing the correct conformation of these junctions. In particular, A-minor motifs occurring within the junction help stabilize the structure, and avoid interconversion of different configurations.

Although in general, due to the conformational flexibility and dynamic character of junctions, a continuum of junction conformations might be possible, our compilation of RNA junction domains based on available structures illustrates Natures strong preferences for the arrangement of RNA helical elements in parallel and perpendicular patterns, while keeping the helical axes coplanar. As recently discussed in an essay [22], most RNA structure and folding data comes from in vitro experiments, where high ionic concentrations can compensate for the lack of in vivo folding factors such as ligands and RNA chaperones. Differences between in vitro and in vivo folds of RNA are still being investigated.

Long-range interactions that stabilize helical elements are very diverse, but often involve Sugar-Sugar interactions in the form of A-minor motifs. Other

interactions such as base-ribose and long-range stacking interactions are also observed. One advantage of studying junctions with different number of helices is that it allows recognition of important repeating motifs such as the sugar-edge interactions (A-minor), sarcin/ricin, and trans AG Hoogsteen-Sugar interactions. These sets of non-canonical base pairs play important roles in RNA′s structure and therefore function.

Ribo-base interactions are novel helix-helix interactions found in perpendicular helical conformations. They belong to the same family of helix packing interactions such as the G-ribo [130], A-minor [105], AGPM [36], and ribose zipper [131]. Because ribo-base interactions often appear next to the along-groove pacing motif (AGPM), both motifs form parts of a larger motif (AGPM/ribo-base), whose main function is to pack together helical elements and stabilize RNA molecule for proper function. Such motifs can also act as RNA-RNA or RNA-protein binding promoters by helping their flanking *trans* AG Hoogsteen-Sugar base pair interactions to expose their hydrophobic surfaces for binding.

As more interactions involving RNA base and ribose are discovered, one can foresee the need to extend the current RNA base pair classification given by Leontis and Westhof [86] to include ribose-base interactions.

We encountered many examples of higher order junctions that arrange their helical elements similar to lower order junctions. The junction examples belong to both homologous and non-homologous RNAs. One can then ask: how are higher order junctions formed? We propose that some junctions with a high degree of branching are formed from insertions and unions of smaller order junctions under evolutionary pressure; the optimal junction sites for insertions and unions likely correspond to regions that would not dramatically change its internal tertiary structure conformation. Our analysis also suggests that higher

order junctions can be decomposed into smaller "sub-junctions". Ultimately, a better understanding of junction decompositions can help predict RNA three-dimensional structures and functions.

Chapter 3

Predicting Helical Topologies in RNA Junctions as Tree Graphs

3.1 Introduction

Exciting recent discoveries have made it clear that RNA functions much like a master programmer—far beyond information transfer and protein synthesis [25, 41, 138, 101].[1] Indeed, RNA′s regulatory roles encompass RNA splicing, protein regulation, small-metabolite sensing, RNA interference, and RNA modifications among others. Intimately connected with these gene altering and editing roles are the structural properties of RNAs because they dictate the dynamics of RNAs as well as interactions with other molecules. The close connection between structure and function of RNAs is evident from the many recent studies of RNA tertiary motifs, as well as advances in various aspects of RNA structure; these advances have in turn stimulated efforts in the struc-

[1]This chapter is based on one of our published articles [199].

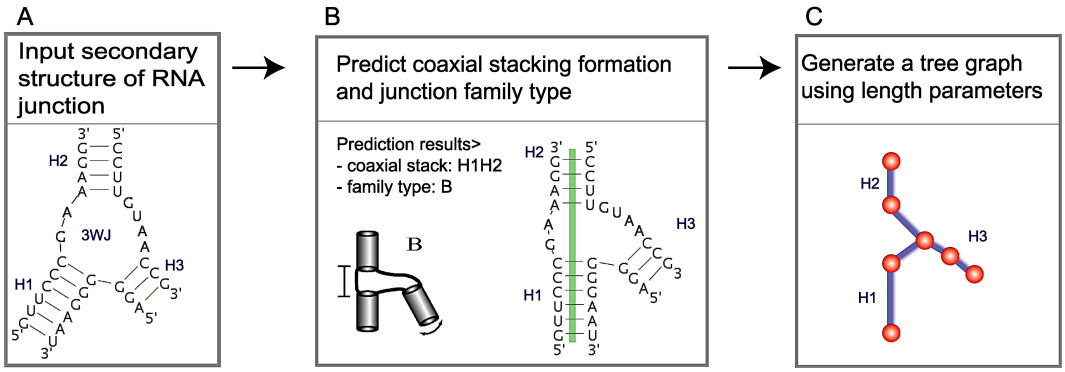


Figure 3.1: RNAJAG starts from an RNA secondary structure **(A)**, uses Junction-Explorer to predict coaxial stacking and junction family types **(B)**, and constructs a scaled tree graph using length parameters **(C)**

ture prediction of RNA (see [21] for de novo RNA structure prediction, and [74, 193, 75, 121, 124] for recent reviews on these topics of 3D structure modeling and prediction).

Here we introduce a new module, called RNAJAG (RNA-Junction-As-Graph), to represent RNA junctions as tree graphs in 3D space and generates a helical arrangement ensemble that approximates plausible 3D structures (see Figure 3.1 for the computational procedure). This module is an updated version of our previous Junction-Explorer program [77], based on the random forests data mining algorithm and uses various geometric and energetic parameters for training (e.g., free energies, loop sizes between junctions, and adenine content). RNAJAG proceeds in two steps: 1) Junction-Explorer is implemented to determine the junction topology, as well as coaxial stacking patterns between helical elements of the target RNA junction for a given 2D structure; we generate the 2D structures by making a consensus on the annotations from FR3D [122], MC-Annotate [196], and RNAVIEW [142]; 2) Using the results in step 1, a 3D

tree graph is constructed by using scaling parameters to determine the length of every edge representing helical axis as well as distance parameters to position the edges in the junction domain. Details are provided below for these two steps followed by the analysis tools needed to assess our predicted graphs, namely converting crystal structures to graphs for measuring various geometric features.

3.1.1 Features of RNAJAG

We present a novel application of graph theory to represent and model helical arrangements in RNA structures. We aim to efficiently sample the 3D conformational space and predict global orientations of RNA junctions, which are important structural elements that form when three or more helices come together in space. As input, we use knowledge of the secondary structure, which can be predicted from the sequence by using programs such as Mfold [146] and RNAfold [47] based on the dynamic programming algorithm first proposed by Nussinov [195, 194], or can be extracted from multiple sequence alignments [46] or from experimental techniques such as RNA probing [24], crystallography, and NMR (resources available in databases such as RNA STRAND [3] and Rfam [42]). The output is a graph model of the predicted junction topology.

Junction topology predictions using Junction-Explorer. Our analysis of RNA junction topologies [73, 72] is built upon previous topology analysis of 3-way junctions by Westhof and co-workers [87], who categorized three major families A, B, and C (Figure 1.3A in Chapter 1). For 4-way junctions, we identified nine major families: H, cH, cL, cK, π, cW, ψ, cX, and X by coaxial stacking patterns and helical configurations (Figure 1.3B in Chapter 1). Helices

within RNA junctions prefer to arrange in parallel and perpendicular patterns, and conformations are stabilized using common 3D motifs like coaxial stacking, loop-helix interactions, and helix-packing interactions. Because the axes of helices in junctions tend to be coplanar [71], we represent junctions using planar tree graphs.

Junction-Explorer [77] uses a data mining approach known as random forests, which relies on multiple decision trees trained here using feature vectors (extracted from the 2D structures of solved RNAs used as the training dataset) for loop length, sequence, and other variables specified for any given junction; to determine the 2D information from the training dataset of 3D structures, we use three different programs—FR3D [122], MC-Annotate [196], and RNAVIEW [142]—and curate the 2D structures to contain only three base pairing types (A-U, G-C, or G-U). We found some cases where programs yield different 2D structures; in such cases, we select the 2D structure with the lowest free energy among these programs as evaluated by the formation of A-U, G-C, or G-U base pairs. To simplify the parsing of an RNA secondary structure into junctions, pseudoknots are automatically removed during the search. Similarly, because we aim to present a computational tool to predict helical arrangements within junctions based solely on a secondary structure, no knowledge from tertiary contacts (including pseudoknots) is introduced in an input secondary structure. Junction-Explorer uses these properties of RNA junctions as a function of sequence content and loop size to predict coaxial stacking patterns and junction family types. For example, a correct prediction of both the family type and coaxial stacking topology for the RNA in Figure 3.1B) is family B and H_1H_2 stacking; family B with H_1H_3 stacking or family A with H_1H_2 would be incorrect in part. Our updated version of Junction-Explorer uses an

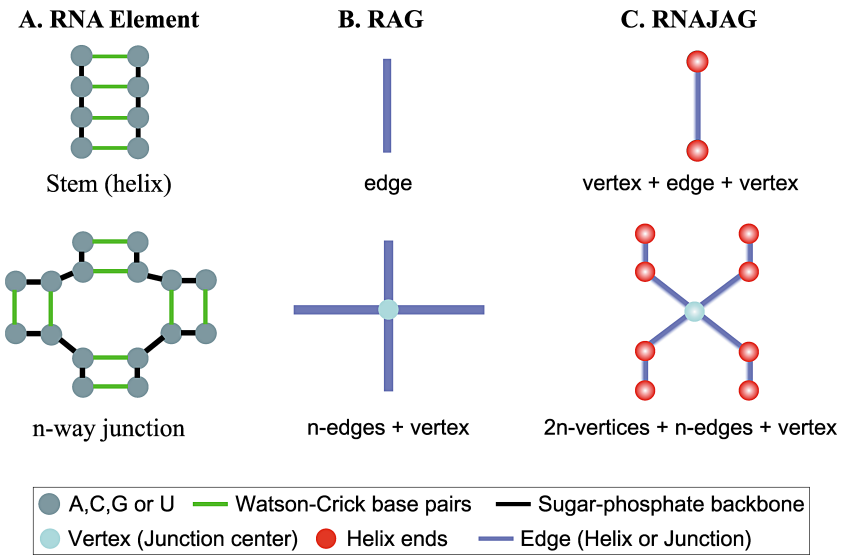


Figure 3.2: RNA graph representations. **(A)** RNA junction elements in a secondary structure. **(B)** RAG tree representation, which describes a helix as an edge and a loop as a vertex. **(C)** RNAJAG tree representation, which defines a helix using an edge and a loop and helix ends using a vertex.

experimental dataset and a standard statistical analysis procedure. Our previous non-redundant junction dataset [77] was updated to include the most recent solved structures found in the PDB database as of October 2012. This dataset includes 130 3-way junctions, and 114 4-way junctions. With the exception of a few 3-way junctions with no coaxial stacks, most new junctions fit within the junction family classifications reported by the Westhof and Schlick groups [72, 73].

Graph representation. Our previous graph theory work considered RNA-As-Graphs [104] to represent RNA secondary structures from a topological perspective [37, 53]. A RAG graph defines trees by representing helices as edges,

and loop domains (hairpins, internal loops, and junctions) as vertices [38] (Figure 3.2A-B). This simple and intuitive representation provides the mathematical tools to estimate the RNA structural space as well as to predict yet unknown motifs [63].

In this work, we add further detail to the tree graphs to represent junction structures. We refine the RAG tree graphs by adding vertices at the terminal base pairs of a helix to represent helices of different lengths. We also include a vertex in the center of the junction domain to capture the junctions spatial properties. In addition, we consider edges connecting the vertices at the end of helices, and edges to connect the end of a helix with the vertex in the center of the junction (Figure 3.2C). We illustrate how to translate RNA structures into RNA graphs, as well as the differences between RAG and RNAJAG, with two examples—a helix and a 4-way junction (Figure 3.2). This new graph representation captures properties of the helical organization for any degree of RNA junctions in 3D space.

3.1.2 RNA graph analysis methods

We utilize two comparison methods—RMSD and Maximum Angle (MaxAngle)—to assess the quality of predicted graphs with respect to the native structures. The RMSD and MaxAngle [104] are useful for measuring global and local similarity of graphs, respectively.

The RMSD measures an average distance of vertices between superimposed graphs, defined as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} \bar{V}_i - V_i}{n}},$$

where n is the total number of vertices, and $\bar{V}_i$ and V_i $(1 \leq i \leq n)$ are vertices

in the reference and predicted graphs, respectively.

To compare a pair of graphs, we translate these graphs into the origin, calculate an optimal rotation matrix using the singular value decomposition (program JAMA, adapted from a java matrix package (http://math.nist.gov/javanumerics/jama)), and superimpose them by a rotation matrix.

MaxAngle finds a maximum angle by calculating an angle of aligned two vectors of edges in the reference and predicted graphs defined by

$$MaxAngle = \max_i \frac{\bar{E}_i \cdot E_i}{|\bar{E}_i| E_i},$$

where i is the number of edges, and $\bar{E}_i$ and E_i are vectors of edges from the reference and predicted graphs, respectively.

3.1.3 Building atomic models using graphs

Our general idea is to use a threading-like procedure to determine the atomic coordinates of the graphs predicted by RNAJAG based on a search for graph similarities in 3D-RAG, an extension of the RAG database. 3D-RAG contains 3D atomic models extracted from high-resolution RNA crystal structures from the PDB databank; atomic structures are linked to corresponding 3D graphs. Figures B.2 and B.3 in Appendix B illustrate the build-up and search procedure of the 3D-RAG database (unpublished). The 3D graphs are classified based on RAG motif IDs, which reflect topological properties of secondary structural elements. We construct all-atom models in three steps (see Figure B.3, Appendix B). First, we identify a motif ID of the target graph. Second, we compare the target graph to all 3D graphs catalogued with the same motif ID in 3D-RAG based on a standard RMSD calculation. Third, we select the graph with the lowest RMSD, extract its all-atom 3D coordinates, and verify that it

contains the same number of nucleotides as the target sequence. The bases are then altered to match the target sequence as needed, while keeping the backbone intact.

If we do not find any structure match for the entire predicted RNAJAG graph, we partition the target graph into subgraphs and follow the procedure described above for each subgraph. We then assemble all the atomic fragments of the subgraphs to form a final all-atom RNA model. Energy minimization may be implemented in the future to relax the structure.

3.1.4 Overview of Results

RNAJAG predicts tree graphs of RNA junctions for a given secondary structure (see Figure 3.1 for the computational procedure). It expands upon our program Junction-Explorer in several important ways; first, RNAJAG generates a candidate junction graph model with specific helical arrangements (on top of family type/stacking orientation); second, the predicted graph incorporates native-like RNA junction features such as interhelical distances obtained from analysis of hundreds of solved RNA junction structures; third, the graph serves as basis to build all-atom models. Results show that RNAJAG reproduces native-like folds of helical arrangements in most junctions tested in the cross validation procedure (3 and 4-way junctions). Specifically, comparisons between our predicted tree graphs and the graphs obtained from solved crystal structures yield RMSD (root-mean-square deviation) values within range of 2-11Å (3-way) and 2-26Å (4-way), for all corresponding junctions. Importantly, the graph output of RNAJAG can be utilized to build coarse-grained or all-atom models and extend the approach to higher-order junctions. In addition, RNAJAG allows determin-

ing helical packing arrangements in junction domains (e.g., coaxial stacking) for larger RNAs, which is one of the main limitations among current RNA 3D prediction methods.

3.2 Results

To represent RNA junctions, we construct 3D tree graphs with the structural configuration consistent with the three and nine types of recognized junction families (see Figure 1.3 in Chapter 1). Because these junction families describe helical arrangements in parallel, perpendicular, and diagonal arrangements, we consider only the graphs that are restricted to these configurations.

3.2.1 Translation of RNA crystal structures to graphs

To evaluate the accuracy of our approach for predicting helical arrangements via tree graphs, we generate a set of graphs obtained from solved crystal structures according to the definition of tree graphs described above. Thus, a helical element in an RNA junction is defined only if at least two consecutive Watson-Crick base pairs (G-C and A-U, and G-U) are present. As described above, we represent each helix by two vertices and one edge: the vector origin (O') of each vertex is determined by three steps: 1) find the midpoint M of $C1'$ atoms between the purine ((A)denine and (G)uanine) and pyrimidine ((C)ytosine and (U)racil) of the terminal base pairs of a helix; 2) consider the orthogonal projection from M to the line connecting the C8 and C6 atoms of the purine and pyrimidine, respectively; 3) scale the vector projection by 4Å as proposed by Schlick [123] (see Figure 3.3A). This definition for positioning a vertex is applied

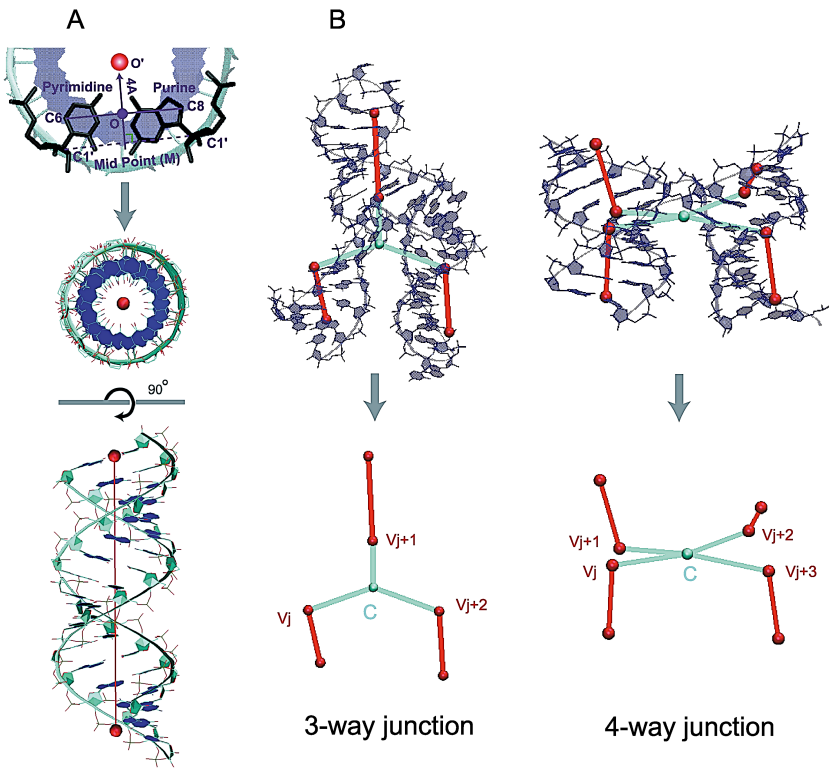


Figure 3.3: Graph representation of a helix and RNA junctions. **(A)** definition of coordinates for the origin (O′) of base pairs (see [123]) and a global helical axis for A-form RNA, from the top and the side. **(B)** Graphs of RNA junctions are obtained by translating helical branches into vertices and edges, and locating the center vertex C of each RNA junction (colored cyan); the center vertex C of an n-way junction is positioned as the average of adjacent vertices of C (v_i, $i = 1,..., n$, for n-way junction) at helix ends.

to both terminal base pairs of a helix. An edge is then added to connect the two adjacent vertices. Note that this edge aligns with the axis of the helix.

We extend this graph definition for helices to describe RNA junctions. For instance, an n-way junction translates into $2n$+1 vertices—$2n$ vertices for n helices and one vertex for a junction centroid—and $2n$ edges; the junction centroid is an average of adjacent vertices V_i (i = 1,..., n). Figure 3.3B illustrates examples of 3 and 4-way junctions and their translation into tree graphs; red edges represents helices while cyan edges illustrates the edges connecting the center of each junction to the helix edges. By converting a set of solved crystal structures into our graph notation, we can derive knowledge-based information about the spatial arrangements of helices within junctions.

3.2.2 Distance parameter calculations using graphs

To determine the distance parameters to scale RNA graphs properly, we analyze structural data of 224 junctions collected from a non-redundant dataset of 47 solved crystal RNA structures (see Figure B.1, Appendix B) and calculate the distances between coaxial helices, parallel, perpendicular, and diagonal helical arrangements in all 3 and 4-way junction elements of our graphs (Figure 3.4). We classify a diagonal topology when the helix axis roughly forms a 45° angle with respect to the axis of stacked helices. Using linear regression we determine the distance between coaxial helical stacks by $s0 = (2.75L + 3.91)$Å (R^2 = 0.84), where L is the number of nucleotides between the helical elements forming coaxial stacks and R^2 describes how well the linear regression fits the dataset (Figure B.1A, Appendix B); the distances between parallel, perpendicular, and diagonal helical arrangements within junctions are determined by the position

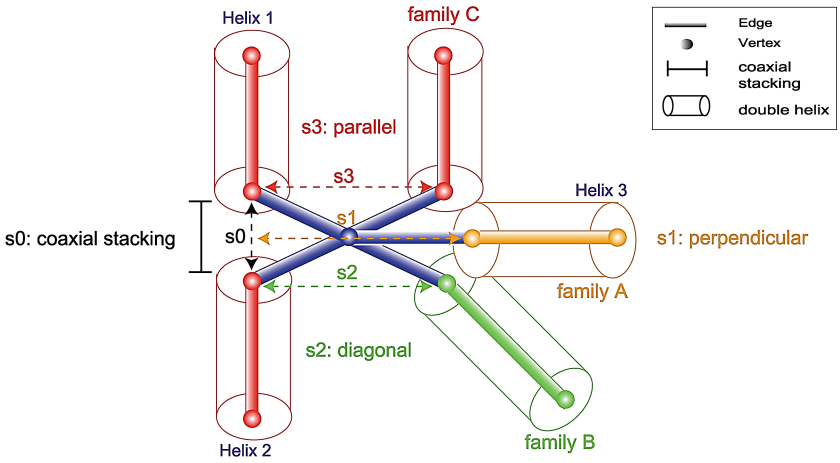


Figure 3.4: Scaling parameter calculations using graphs translated from crystal structures. The diagram shows the scaling distance parameter calculations for 3-way junctions where the scaling parameters s0, s1, s2, and s3 denote the distances between coaxial helices, perpendicular, diagonal, and parallel helical arrangements, respectively.

of unstacked helices with respect to the coaxially stacked helices (see s1, s2, and s3 in Figures 3.4 and B.1) and reported as average values (with standard deviations) of 20.48 ($\pm$ 5.25)Å, 19.95 ($\pm$ 2.71)Å, and 21.17 ($\pm$ 5.20)Å, respectively (see Figure B.1 in Appendix B for the distance distributions). In addition, we estimate the length of edge parameters representing the helical axis based on the distance of solved helical elements found within our non-redundant dataset. The helix length parameter is given by $2.87(b-1)$Å, where b is the number of base pairs and 2.87Å corresponds to the base rise [123].

3.2.3 Relation between graph and atomic models

To analyze the relation between root-mean-square deviation (RMSD) for tree graphs as opposed to atomic models, we calculate RMSDs for 13 all-atom models predicted from MC-Sym [109], NAST [55], and FARNA [23] against their corresponding all-atom native structures (33 calculations in total). This dataset of 13 structures composed of 3 or 4-way junctions was selected because both secondary and tertiary structures have been experimentally determined and they represent diverse features: the lengths vary from 51 to 117 nucleotides, and the topologies are diverse, including pseudoknots and loop-loop interactions. In addition, while some structures have been solved in the presence of proteins, others are structurally stable (e.g., tRNA), or rearrange upon binding to a substrate (e.g., ribozymes, riboswitches). We then build the tree graphs associated with these predicted atomistic models and compare these graphs to the corresponding graphs obtained from native structures (as described above). When performing a linear regression analysis using the RMSD values, we observe a positive correlation between all-atom and graph models (Figure 3.5). Thus, assessing graphs using the RMSD method is not equivalent to all-atom RMSD calculations but indicates similar trends.

3.2.4 RNAJAG prediction performance

To assess general RNAJAG performance, we consider the set of 200 junction domains (100 each for 3-way and 4-way systems) from high-resolution crystal structures as prediction targets. Results in Table B.2, Appendix B and Figure 3.6 (RMSD distributions) show that RNAJAG reproduces well native-like RNA folds in most of the 3 and 4-way junctions tested in the cross valida-

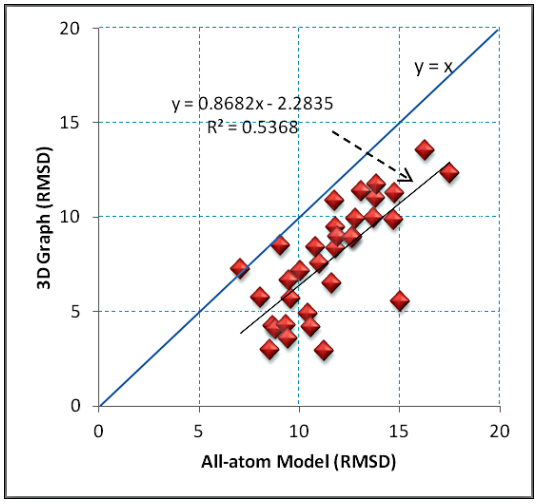


Figure 3.5: Statistical analysis of RMSDs for graphs with respect to their atomic models using a linear regression. Overall, a positive trend between all-atom models and graphs is observed with a slope value of 0.86.

tion procedure. As the module RNAJAG consists of two components—junction topology prediction and graph modeling, we discuss the two parts in turn.

Overall, for the first component—junction topology—results indicate that the junction topology predictor module of RNAJAG, Junction-Explorer, identifies topologies and stacking patterns reasonably well for most of the test examples. Specifically, the module achieves accurate coaxial stacking prediction (95/100 for 3-way and 92/100 for 4-way) as well as junction family type (94/100 for 3-way and 87/100 for 4-way). Interestingly, most of the incorrect predictions for 4-way junctions correspond to families π and X, which are junction topologies rarely encountered. Other cases involving unusual inter or intra-molecular interactions (e.g., D-loop/T-loop interaction) are beyond the capability of our data mining approach and can lead to erroneous topology predictions.

Our second component, graph modelling, builds a candidate model graph

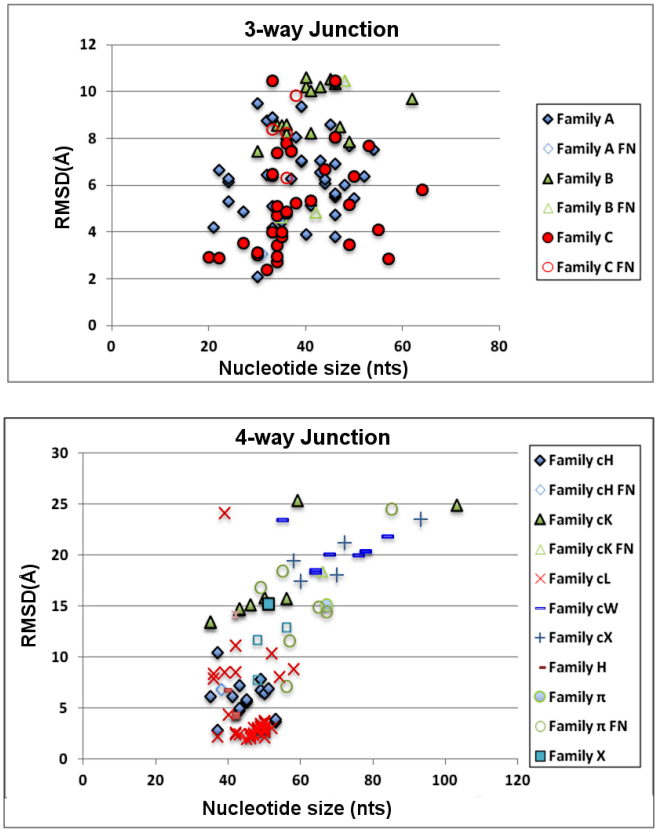


Figure 3.6: Distribution of RMSD scores for 3-way junctions (top) and 4-way junctions (bottom). The RMSD comparison is computed between the RNAJAG graphs and the graphs obtained from the PDB structures corresponding to the target RNA. Values are color-coded according to their correctly predicted family topology (solid colors), as well as the failed family predictions (false negatives with same shape but no filling).

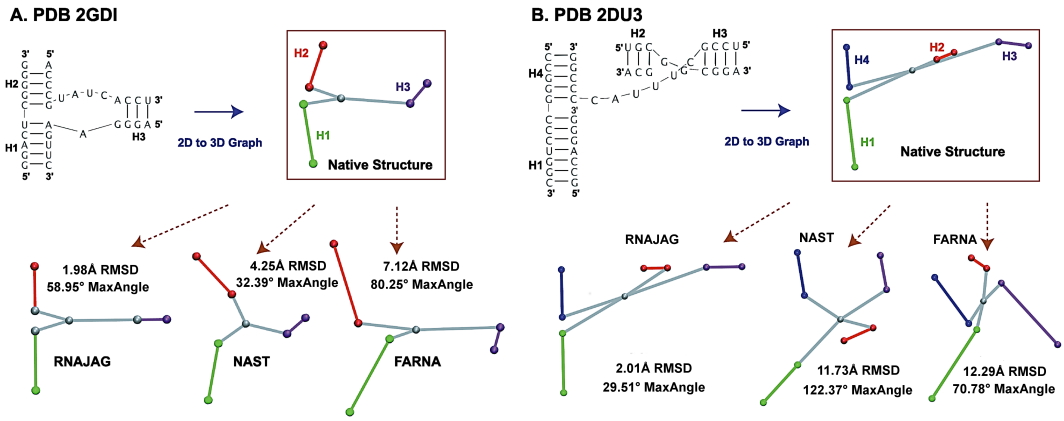


Figure 3.7: Prediction results of 3D modeling programs. Starting from the bottom left, a list of predictions from each program is presented by increasing RMSD values against the native structure in the counterclockwise direction. **(A)** 3-way junction of the TPP riboswitch (PDB 2GDI) with family type *A* and coaxial stacking between helices H_1 and H_2. Based on these examples, RNAJAG predicts most accurately followed by NAST, and FARNA. **(B)** 4-way junction of tRNA (PDB 2DU3) with family type cL and coaxial stacking between helices H_1 and H_4, and H_2 and H_3. After RNAJAG, NAST predicts most accurately followed by FARNA.

compatible with the predicted junction topology as described under Methods. These scaled tree graphs are generated and compared using RMSD and MaxAngle to those graphs from the corresponding native crystal structures. While RMSD is a global measure of graph similarity, MaxAngle, defined by a maximum angle of two aligned edge vectors (See Figure 3.7), is a local measure of accuracy that can help understand specific graph differences. For all 200 junctions considered, comparisons between our predicted and native tree graphs for all corresponding junctions yield RMSD values within range of 2-11Å (3-way) and 2-26Å (4-way). The RMSD values are presented and grouped by successful or missed junction family predictions in Figure 3.6. Interestingly, we note that for junctions corresponding to family C, our method produces reasonably graph junction models, while RMSDs for junctions belonging to family B perform poorly. A possible explanation is that for junction members of the family B, there is a high variability of the spatial arrangement between the coaxial stacking and its third helix. The parallel helical packing from junction elements of family C, on the other hand, tends to make a small variation because the coaxial stacking and its third helix often form long-range contacts. Similarly, we can observe that 4-way junction families of types cL, cH present better RMSD scores because these families are among the most abundant, and also present less variability in their inter-helical distances due to long-range contacts formed at the point of strand exchange.

We now analyze these RNAJAG results for a set of 13 representative RNAs of diverse sizes and functions (Table 3.1) by the same cross-validation procedure (leave-one-out). Correct junction topology classification is critical to achieve native-like graphs. Among the correct predictions for the junction topology in 3 and 4-way junctions are, for instance, the riboswitch (PDB 2GDI) and tRNA

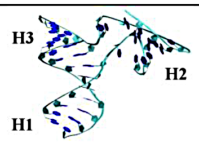
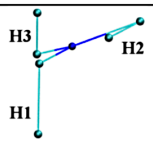
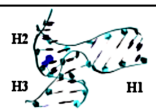
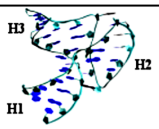
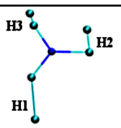
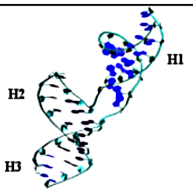
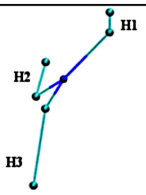
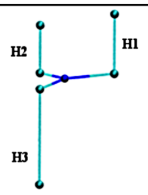
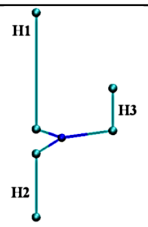
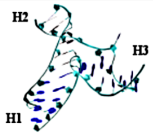
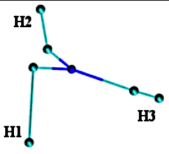
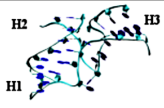
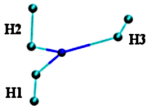
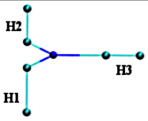
Table 3.1: List of 13 representative RNA junctions from the PDB database. Each junction is listed with its junction family and coaxial stacking arrangement. RNAJAG achieves graphs with RMSD values below 11Å and 13Å for 3 and 4-way junctions, respectively.

PDB	Nts	RNA Type	Degree	Native Structure		RNAJAG			
				Coaxial Stacks	Family Type	Coaxial Stacks	Family Type	RMSD (Å)	Max Angle (°)
2FK6	52	tRNA	3WJ	H_1H_3	C	H_1H_3	A	4.01	166.43
1DK1	57	rRNA	3WJ	H_2H_3	A	H_2H_3	A	6.16	65.70
1MMS	58	rRNA	3WJ	H_1H_3	C	H_1H_3	C	4.13	16.36
3EGZ	65	Riboswitch	3WJ	H_2H_3	C	H_2H_3	C	6.59	46.63
2QUS	64	Ribozyme	3WJ	H_1H_3	C	H_1H_2	C	10.40	159.06
2OIU	51	Ribozyme	3WJ	H_1H_2	A	H_1H_2	A	2.12	28.98
3D2G	77	Riboswitch	3WJ	H_1H_2	A	H_1H_2	A	2.07	45.90
2HOJ	78	Riboswitch	3WJ	H_1H_2	A	H_1H_2	A	2.18	52.36
2GDI	80	Riboswitch	3WJ	H_1H_2	A	H_1H_2	A	1.98	58.95
1LNG	97	7S.S SRP	3WJ	H_1H_3	C	H_1H_2	A	9.04	62.26
1MFQ	117	7S.S SRP	3WJ	H_1H_3	C	H_1H_3	C	5.26	16.21
2DU3	71	tRNA	4WJ	H_1H_4,H_2H_3	cL	H_1H_4,H_2H_3	cL	2.01	29.51
2GIS	94	Riboswitch	4WJ	H_1H_4,H_2H_3	cL	H_1H_4,H_2H_3	cL	12.18	74.08

(PDB 2DU3), yielding best RMSD values of 1.98Å and 2.01Å, respectively.

An example of a misclassification involves the tRNA (PDB 2FK6); it was assigned to a family *A*, but the native RNA structure forms a D-loop/T-loop motif (loop-loop tertiary interaction commonly observed in tRNA [197]) outside the junction domain that stabilizes its structural configuration as a family C (see Figure 3.8). Such misclassifications also occur for coaxial stacking; the hammerhead ribozyme (PDB 2QUS) was correctly classified in family type, but the coaxial stacking was predicted as H_1H_2 instead of H_1H_3. Finally, the signal recognition particle (PDB 1LNG) is incorrectly predicted, perhaps due to the small loop size differences, 1 *nt*, between H_1H_2 and H_1H_3 (see Figure 3.8).

Most RMSD values fall below 7Å except for the three examples (ribozyme (2QUS), SRP (1LNG), and riboswitch (2GIS)) that are within the range of

		Native Structure		RNAJAG
PDB	Degree	All-atom	Graph	Graph
2FK6	3WJ			
1DK1	3WJ			
1MMS	3WJ			
3EGZ	3WJ			
2QUS	3WJ			
2OIU	3WJ			
3D2G	3WJ			

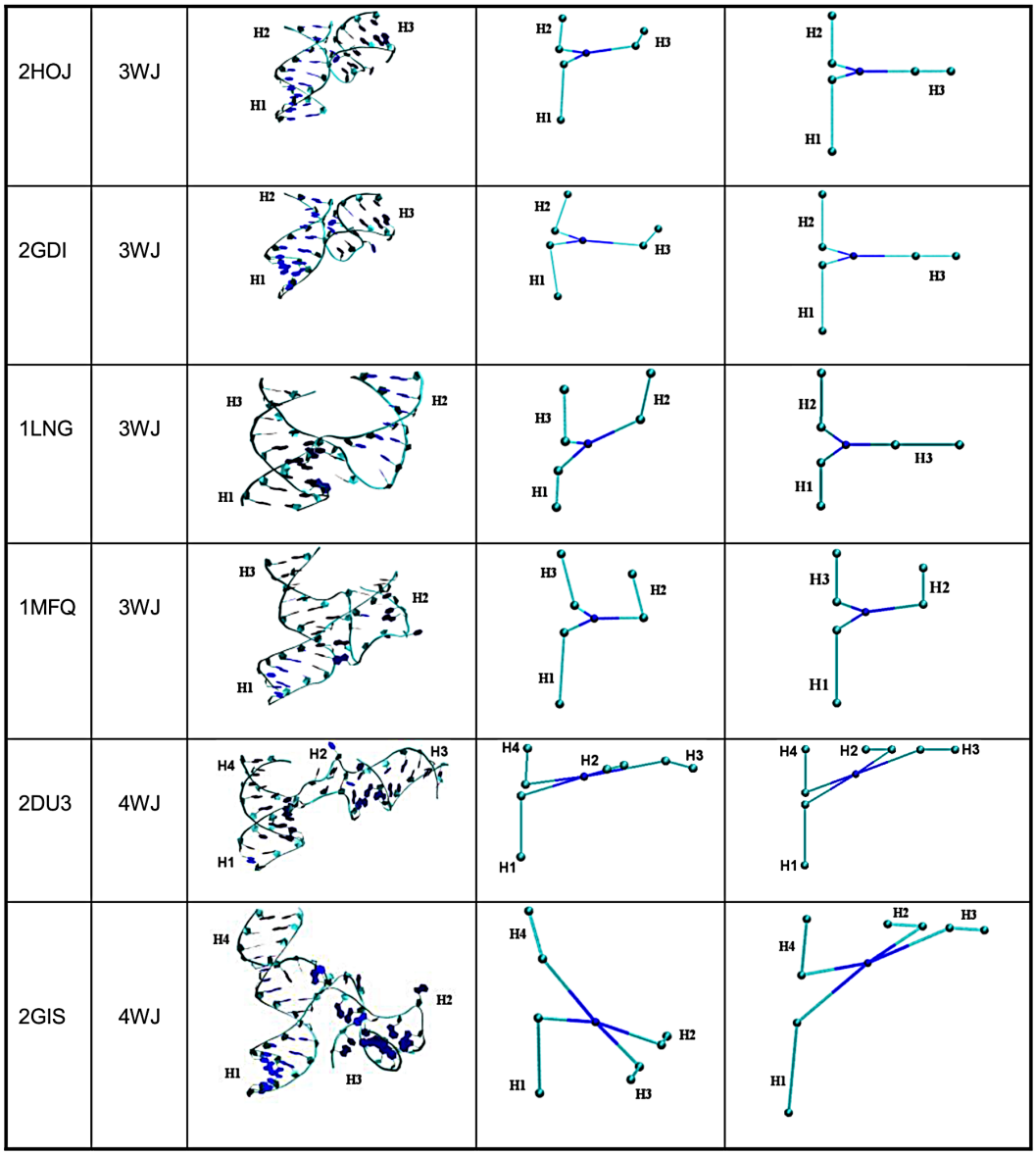


Figure 3.8: Graphs of the 13 representative RNA junctions. In each column from left to right, PDB entry, junction type, native structure, graph from native structure, and graph from RNAJAG are shown.

9 to 13Å. Similarly, most MaxAngle values fall below 75°, except for the two examples (tRNA (2KF6) and ribozyme (2QUS)) that have values higher than 159° due to incorrect topology predictions. The graphs (corresponding to RNAs listed in Table 3.1) are shown in Figure 3.8 for both the native structures and RNAJAG models.

3.2.5 Computational performance of RNAJAG relative to other RNA folding programs

To compare the performance of RNAJAG with other programs, we made use of programs such as MC-Sym [109], NAST [55], and FARNA [23] to produce models from a selected set of 13 RNA junctions. The junctions consist of 3 and 4-way junctions and represent diverse features including nucleotide length and topology. To make comparisons at the graph level, we translate all predicted atomistic models into tree graphs using our graph definition (Figure 3.3), and compute RMSD and MaxAngle against the corresponding graphs of native structures (graph from predicted structure vs. graph from crystal structure). The results are presented in Table 3.2 and the distributions in Figure B.4, Appendix B.

Although comparative RMSD values with respect to graphs and atomic models are not interchangeable, they are closely correlated. Our statistical analysis uncovers the relationship between atomic models and their translated graphs, indicating that atomic models are well described in highly coarse-grained models (Figure 3.5).

We observe that both the RMSD and MaxAngle values range widely depending on the program. Specifically, RNAJAG produces a wider range of

Table 3.2: Comparison between RNAJAG and other tertiary structure prediction programs. Only the junction domain is considered for the RMSD and MaxAngle calculation using graph representation. The best RMSD and MaxAngle values for each structure are highlighted in bold on background. We denote N/A for those structures that other programs failed to predict using secondary structure information.

	RMSD (Å)				MaxAngle (°)			
PDB	RNAJAG	MC-Sym	NAST	FARNA	RNAJAG	MC-Sym	NAST	FARNA
2FK6	**4.01**	8.51	N/A	11.38	166.43	108.31	N/A	**49.10**
1DK1	6.16	5.74	**4.06**	6.63	65.7	129.42	**46.19**	140.39
1MMS	4.13	9.85	**2.89**	9.46	**16.36**	32.46	96.15	112.55
3EGZ	6.59	**5.53**	5.69	9.90	46.63	50.32	**36.04**	72.07
2QUS	10.40	**8.34**	10.86	8.94	159.06	130.49	59.80	**42.45**
2OIU	**2.12**	4.21	N/A	8.39	**28.98**	71.55	N/A	84.00
3D2G	**2.07**	N/A	N/A	3.56	**45.90**	N/A	N/A	88.84
2HOJ	**2.18**	N/A	N/A	4.17	**52.36**	N/A	N/A	74.07
2GDI	**1.98**	N/A	4.25	7.12	58.95	N/A	**32.39**	80.25
1LNG	9.04	**6.47**	7.55	8.92	62.26	80.16	**57.62**	73.84
1MFQ	**5.26**	9.93	11.01	8.88	**16.21**	82.10	83.05	78.30
2DU3	**2.01**	N/A	11.73	12.29	**29.51**	N/A	122.37	70.78
2GIS	12.18	13.51	N/A	**11.27**	**74.08**	101.37	N/A	122.67

RMSD values varying from 1.9-12.2Å, with the largest values occurring mostly when coaxial helices or junction family (or both) are inaccurately predicted. In tandem, the best prediction values are observed when RNAJAG correctly classifies both the junction family type and coaxial stacking formation. The RMSD values for MC-Sym range from 4.2-13.5Å, NAST from 2.9-11.7Å, and FARNA from 3.5-12.3Å. By considering the number of predicted structures with best RMSDs over these 13 test cases, RNAJAG outperforms with 7 predictions followed by MC-Sym, NAST, and FARNA for 3 or less. MC-Sym and NAST often fail to predict structures, possibly due to some complications with the fragment insertion or assembly as reported in our previous study [74]. Although FARNA

performs structure predictions least accurately, the program produces a model for all the structures along with RNAJAG.

To complement the RMSD measures, we also use MaxAngle to assess a local agreement of edges in the predicted graphs. The MaxAngle values for RNAJAG range from 16.2-166.4°, but mostly less than 65° with only three exceptions. Again, the largest (worst) values occur when RNAJAG fails to achieve the correct junction family and/or coaxial stacking patterns. The MaxAngle values for MC-Sym range from 32.4-130.5°, NAST from 32.4-122.4°, and FARNA from 42.4-140.4°. Overall, RNAJAG performs better on 7 of the 13 predictions, followed by NAST, FARNA, and MC-Sym for 4 or less.

Figure 3.7 presents two cases of graph comparisons between the native structure and graphs predicted by RNAJAG and the other programs to illustrate where predictions deviate from the experimental structure and from each other. The first example (Figure 3.7A) considers the 3-way junction structure of the TPP riboswitch (PDB 2GDI). When the RNAJAG graph is compared to the native one, RMSD and MaxAngle values of 1.98Å and 58.95°, respectively, are obtained. Interestingly, RNAJAG produces the best graph model with the lowest RMSD value, but not the lowest MaxAngle value; NAST yields a graph with the best MaxAngle value of 32.39°. Note that the graph conformations of RNAJAG for 3-way junctions are predefined by the major junction family types (Figure 1.3A in Chpater 1) whereas NAST has much larger conformational space to explore, thus leading to a better fit of H_3 to the native structure in this case. Our graph representation also gives ideal alignments for the coaxial helices, which is not always the case for graphs obtained from native structures, possibly due to helical rearrangements outside the junction domain.

The second case is the 4-way junction obtained from a Cys-tRNA transfer

RNA (PDB 2DU3). In contrast with other programs, RNAJAG generates the typical L-shape with similar proportions to the native state (Figure 3.7B), without knowledge of the D-loop/T-loop interaction occurring outside the junction domain, and yields the lowest RMSD (2.01Å) and MaxAngle (29.51°) among the programs. Considering the RMSDs, NAST follows RNAJAG, with 11.73Å, and it is followed by FARNA (12.29Å). MC-Sym was unable to generate a model in these examples, possibly due to the insufficient number of cyclic motif fragments to insert.

In both prediction cases, RNAJAG configures most edges similar to the native structures; however, the scaling of the loop region in the tRNA (Figure 3.7B) is slightly inaccurate and would require additional information (e.g., tertiary motifs) for proper orientation.

3.2.6 Building all-atom models using graphs

Of course, predicted model graphs are only a starting point. Ultimately, a protocol to build atomic models is required. Using the threading/build-up procedure described in Methods, we illustrate this idea for two mid-sized ($\sim$50 *nts*) junction structures (see Figure B.2 and B.3, Appendix B for technical details).

The 3-way junction, guanine riboswitch RNA, is 53 *nts* long (PDB entry 3RKF) and belongs to the family type *C*. RNAJAG correctly predicts both the junction family type and the coaxial stacking and yields a graph with RMSD value of 4.32Å with respect to the graph of its native structure (See Table 3.3 and Figure 3.9).

We superimpose the predicted graph against all the graphs of the same motif ID family (namely (4, 2)) available in the 3D-RAG database, and order all these

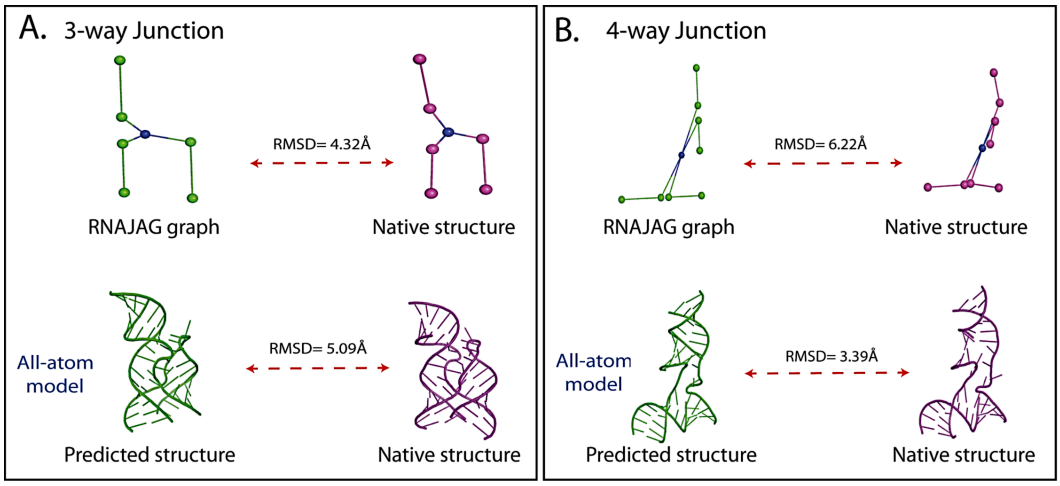


Figure 3.9: Derived all-atom models from predicted RNAJAG graphs using 3D-RAG threading for: **(A)** 3-way junction of a guanine-riboswitch RNA (PDB entry 3RKF) and **(B)** 4-way junction of a tRNA of Staphylococcus aureus (PDB entry 1QU2).

matches based on their RMSDs to the target graph. We extract the all-atom coordinates of the lowest RMSD graph (4.41Å), and create a model by mutating the bases to match the query sequence. We obtain an RMSD value of 5.09Å for the all-atom model junction region compared to its native structure.

The 4-way junction topology of the tRNA of Staphylococcus aureus, 50 *nts* long (PDB entry 1QU2), is correctly predicted by RNAJAG. It generates graph with 6.22Å RMSD compared to the graph of its native structure (See Table 3.3 and Figure 3.9).

Similar to the 3-way junction, we search the 3D-RAG database for graph similarities in the same motif ID family (5, 3). We verify the 2D structure and construct an atomic model by mutating the bases of the extracted structure to match the query sequence. We achieve an all-atom model with RMSD of

Table 3.3: All-atom modeling examples built from graphs. Both examples show comparable RMSD values (computed for all atoms except hydrogen) to the native structures.

PDB	Junction type	NTs	Graph (RMSD)	Atomic model (RMSD)
3RKF	3-way junction	53	4.32Å	5.09Å
1QU2	4-way junction	50	6.22Å	3.39Å

3.39Å against the junction native structure.

3.3 Discussion

With the continuous discovery of novel RNAs, it is imperative to advance computational methods to determine RNA structure and thus help in understanding RNA function. A major limitation in the field of RNA structure is the size of RNA molecules that can be accurately predicted. Indeed, the structural complexity grows rapidly as molecular size increases.

RNA junctions are important structural components that are often difficult to determine at both the secondary and tertiary structure levels. To address this problem, we introduced here a new graph theoretic approach that is applied to model RNA junctions in 3D space. The simplicity of using tree graphs to represent RNA junctions allows us to sample the minimal conformational space, particularly on the assembly of helical elements. Although our tree graph notation cannot represent pseudoknots, the proximity in 3D space of edges representing helices in junctions can suggest the formation of long-range interactions (pseudoknots, kissing hairpins, loop-receptors, etc. [141]).

RNAJAG is the new module that predicts and builds helical models for RNA junctions as tree graphs and consists of two components—junction topology prediction and graph modeling. Using an updated version of Junction-Explorer [77], we determine both the junction family type and coaxial stacking patterns. Based on these prediction results, an RNA graph, consisting of vertices and edges, is then constructed using length parameters describing spatial arrangements of helices in junctions. Note that the accurate prediction performance of Junction-Explorer is a critical step in RNAJAG as the tree graph generation depends sensitively on the outcome of Junction-Explorer.

Overall, RNAJAG reproduces reliable helical arrangements of the junctions with competitive RMSD values, in the range of 2-11Å (3-way) and 2-26Å (4-way) (see Table B.2, Appendix B). In addition, the predicted graphs described here are comparable or better than other RNA folding programs. Note that RMSDs for RNAs are generally much larger than scores from protein predictions [75, 108] and also have a larger volume per unit mass. Thus, while 6Å RMSD is generally considered poor for proteins, it is a good prediction for RNAs. For atomic models, other measures besides RMSDs have alternatively been proposed to better assess RNA predictions [108, 198]. This is partly because nucleotides have a larger molecular size than proteins (while the diameter of a α-helix is 12Å, a typical A-DNA helix has a diameter of 23Å). The results from Table 3.3 show that our approach provides the largest number of best predictions, 7 for both RMSD and MaxAngle measures among compared graphs. Specifically, RNAJAG gives top 7 RMSD values compared to 3 or less out of 13 graphs with respect to MC-Sym, NAST and FARNA. Similarly, RNAJAG yields the top 7 MaxAngle measures compared to 4 or less for MC-Sym, NAST and FARNA.

Accurate predictions of Junction-Explorer in most instances make RNAJAG

competitive with other programs. On the other hand, incorrect determinations of coaxial stacks and/or junction family types in a minority (20%) of the cases (Table 3.1) lead to dramatic deterioration of accuracy. The wide range of RMSD and MaxAngle values may reflect this possibility as reported in Table 3.3.

Our resulting tree graphs hold promise for further refinement of RNA structures. For example, our graphs can be used as starting templates to build coarse-grained or full atomic models using a threading/build-up procedure to link subgraph components and atomic structure (Figures B.2 and B.3, Appendix B). For these two examples, accurate all-atom models are achieved with RMSD values of 5.09Å and 3.39Å for 3 and 4-way junctions, respectively (Figure 3.9 and Table 3.3). Current work is focusing on generalizing this approach.

Although the tree graphs and all-atom models are not comparable, our statistical analysis shows that the RMSD measures of these two distinct models are positively correlated (Figure 3.5); a tree graph model is an oversimplified representation of the atomic RNA structure where helical elements and loop regions are mapped by a finite number of edges and vertices. Generally speaking, lower RMSD values for atomic models can be obtained compared to graph models. Additionally, we use MaxAngle to evaluate the quality of predicted local helical arrangements.

In this work we have primarily focused on pseudoknot-free 3 and 4-way junctions. These junctions represent over 80% of RNA junctions found in all available crystal structures to date [77]. RNAJAG can potentially be extended to predict higher order junctions since Junction-Explorer is capable of predicting coaxial stacking patterns for any junction order. For example, 5-way junctions can be partitioned into various possibilities of 3 and 4-way junctions [73], and thereby model the subset of junctions using RNAJAG.

Though our promising approach could be easily adapted to large RNAs with multiple junctions, several challenges remain with respect to the prediction accuracy of both the junction family and coaxial stacking configurations. For example, when loop-loop interaction motifs (e.g., PDB 2FK6) form outside the junction domain, they lead to unpredictable junction configurations. We also cannot account for protein-RNA interactions or solvent effects, challenges to all other tertiary structure prediction programs.

Finally, RNAJAG considers a limited range of the conformational space [72, 87] since we only consider parallel, perpendicular, and diagonal helical arrangements. These orientations make graph generation very rapid; however, describing the dynamic nature of RNA structures requires flexible models, which can be addressed using coarse-grained or atomic models.

Additional ongoing work involves determining the optimal helical positions of the internal loops as well as the helical elements connecting these loops for large RNAs. Internal loops flanked by two helices can also be represented using tree graphs; therefore, preferred structural arrangements based on loop size and sequence content for them will improve the overall models. Ultimately, a pipeline that starts from our tree graphs and results in all-atom models can be envisioned. Combined with successful predictions of helices and internal loops, junction arrangement predictions could eventually provide a novel hierarchical approach to build tertiary RNA models for large RNA molecules.

Chapter 4

Candidate RNA Structures for Domain 3 of the Foot-and-Mouth-Disease Virus Internal Ribosome Entry Site

4.1 Introduction

4.1.1 Translation initiation mechanisms of FMDV IRES

The foot-and-mouth-disease virus (FMDV) belonging to the picornavirus family is the contagious agent of foot-and-mouth-disease, a severe plague for animal farming.[1] The viral replication of FMDV begins with a translation initiation by forming a specific RNA structure called internal ribosome entry site (IRES).

FMDV IRES consists of ∼450 nucleotides and can fold in multiple stem-

[1]This chapter is based on one of our published articles [148].

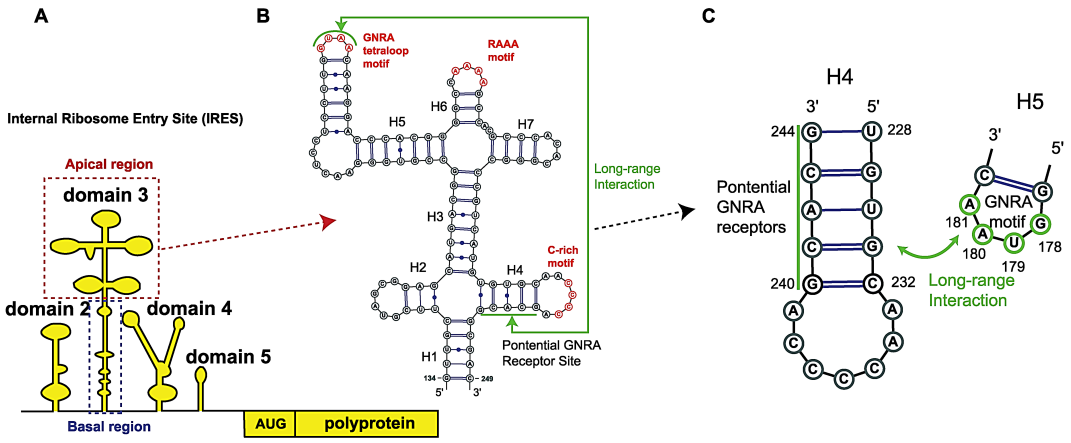


Figure 4.1: Global organization of FMDV IRES and a secondary structure of truncated domain 3 including conserved RNA motifs for RNA-RNA long-range interactions. **(A)** Schematic representation of the viral genome organization including four subdomains of FMDV IRES. **(B)** Secondary structure (deduced from RNA structure probing experiment) of the truncated FMDV IRES domain 3 which consists of a pair of 4-way junctions and is a self-folding region containing conserved GNRA and RAAA motif at the apical region for RNA-RNA long-range interactions. **(C)** Potential long-range interactions between helices H_4 and H_5. $G_{240}CACG_{244}$ residues in helix H_4 is a potential receptor site of the $G_{178}UAA_{181}$ tetraloop.

loops organized in five domains (Figure 4.1A). These domains can host binding proteins such as eukaryotic initiation factors (eIFs) and IRES transacting factors (ITAFs), which play crucial roles in IRES-directed translation [69, 94, 114, 92, 129, 2, 107]. Among these, the third domain is the largest and contains structural elements critical for IRES activity [93].

4.1.2 4-way RNA junctions in domain 3 and their potential roles in IRES activity

Domain 3 consists of basal and apical regions (Figure 4.1A). The basal region consists of a long internal loop, and the apical region contains multiple 4-way junctions. Recent biochemical data have suggested that it is the apical region that contributes significantly to the structural organization and stability of domain 3, as well as to the critical function of IRES activity [93, 34, 31, 32].

Specifically, the apical region of domain 3 includes two conserved GNRA and RAAA motifs [93]. The GNRA (N is any nucleotides; R is A or G) tetraloop motif is common in folded RNA [18]; the loop-helix interactions combine base-pairing and stacking to define a tertiary contact that stabilizes the global fold of an RNA molecule. In the IRES domain 3 (see Figure 4.1), the GNRA motif is situated at the apex of a stem-loop motif [33, 28]. Biochemical studies demonstrate that this motif is critical for IRES function [93, 120]. RNA probing experimental data further show that the RAAA motif also contributes to enhance IRES activity via RNA-RNA long-range tertiary contacts, but only in the presence of GNRA tetraloop-receptor long-range interactions [34].

Deciphering the contribution of domain 3 to IRES-driven translation has been challenging. Based on the potential capacity for inter and intramolecular RNA-RNA interactions, it has been proposed that this domain stabilizes the entire IRES element [34, 97]. More recently, structural analysis based on SHAPE probing and microarray data confirmed domain 3′s role in the organization of other domains [31, 32].

The GNRA motif in helix H_5, along with its potential distal binding region in helix H_4 for intramolecular RNA-RNA interactions, is located in the 4-way

RNA junctions of FMDV IRES domain 3 (Figure 4.1). RNA junctions in general provide a hub for different double-stranded helical arms to come together [90]. Thus, junctions occur in many RNAs, including, for example, the hepatitis C virus IRES for the translation initiation [58]. Because the global conformation of RNA is thought to be largely determined by topological constraints encoded at the secondary structure level [4], an understanding of the three-dimensional (3D) structural aspects of RNA junctions in IRES′s domain 3 is essential to decipher the mechanism of IRES-driven translation.

4.1.3 Challenges in multiple RNA junction structure predictions

Among recurrent structural elements (or motifs) common to RNA junctions, coaxial stacking is prominent. Coaxial stacking between two continuous helices stabilized by base stacking in a shared single strand [97] is a major determinant of three or higher-order junctions of RNA. Recent studies of RNA junctions have identified structural patterns in coaxial stacking that define different RNA junction family types [72, 73, 87]. These classifications also link the RNA junction family type to a nucleotide length in a single strand; fewer nucleotides in the single strand between two helices increase the probability of forming coaxial stacking. Although other factors such as protein binding can also alter these noted patterns of coaxial stacking arrangements, the above correlation holds in general particularly for self-folding RNA molecules, including transfer RNA (tRNA) [65], P4-P6 domain of the *Tetrahymena* group I ribozyme [14], hepatitis C virus IRES [58], and FMDV IRES domain 3 [93]. The coaxial stacking motif often cooperates with other tertiary motifs including A-minor and loop-helix

interaction to enhance the stability of RNAs.

Currently, computational programs cannot predict multiple RNA junction structures well, though there are many useful 3D prediction programs as recently surveyed [75, 74]. Very recently, our RNA junction structure prediction program Junction-Explorer, based on data mining and bioinformatics, was shown to predict the topology of individual RNA junction domains with about 70% or higher prediction accuracy [76].

4.1.4 Overview of Results

To construct plausible structures for the two consecutive 4-way junctions in domain 3 of IRES, we devise a divide-and-conquer approach that combines various effective computational techniques. We began with the IRES secondary structure determined by RNA probing [33, 113]. Considering RNA-RNA long-range interactions involving the GNRA motif, we partitioned RNA into subsystems and then modeled each RNA junction topology on the basis of knowledge from 4-way RNA junction classification coupled with Junction-Explorer. Further analysis produced four viable candidates for 3D models constructed using MC-Sym [109].

Subjecting these four candidate models to MD simulations allowed identification of the most energetically favorable and stable conformational states in the presence of the GNRA tetraloop-receptor long-range interactions. The dynamics data also suggested specific tertiary interactions and helical rearrangements. Only one model emerged as viable, revealing not only the specific binding site for the GNRA tetraloop, but also helical junction arrangements that enhance the stability of domain 3 further. We propose this structure, compatible with

available experimental data, as a feasible tertiary structure for the apical region in FMDV IRES domain 3.

4.2 Computational Methods

4.2.1 RNA target structure

Domain 3 of FMDV IRES is a self-folding RNA that is 214 nucleotides long. We consider the sequence of the FMDV C-S8 IRES and model the apical and basal region separately; the apical region contains two consecutive 4-way junction structures, which consists of 116 nucleotides (G_{134} to C_{249}) and the basal region is a long internal loop containing 98 nucleotides (G_{86} to U_{133} and C_{249} to C_{299}).

4.2.2 RNA sequence conservation analysis

To assess the significance of the structural key elements involved in long-range RNA-RNA interactions in domain 3, we perform sequence alignments of many IRES sequences and analyze the 4-way junctions, focusing on the sequence conservation of the GNRA loop and its binding receptors. 318 FMDV IRES sequences are collected from the GenBank database [8] using the standard Nucleotide Blast webserver with a query sequence of the FMDV C-S8 IRES. Incomplete and identical sequences we removed, and the remaining 318 sequences we aligned using the ClustalW program [78]. We use sequence logos to analyze patterns in aligned RNA sequences. The RNA sequence logos consist of stack of four letters (measured in 2 bits)—A, U, G, and C—at each position in a sequence. While the overall height of the stack indicates a degree of sequence conservation, the height of each letter within the stack shows a relative frequency

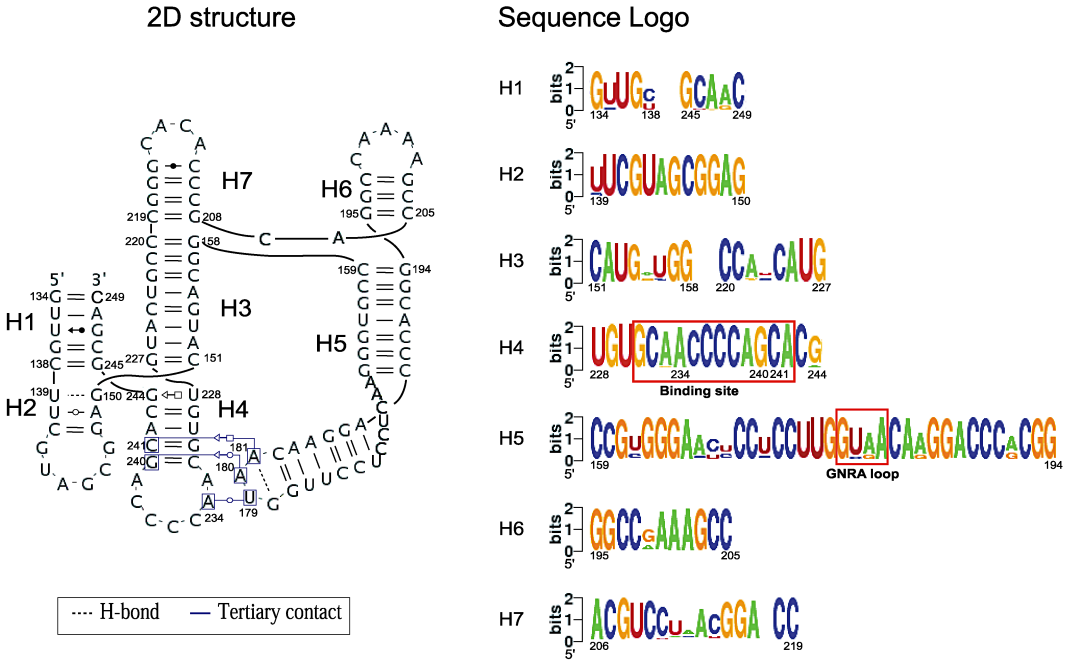


Figure 4.2: Sequence conservation analysis using logo for sequences of the apical region in domain 3 derived from 318 FMDV IRES systems. The RNA sequence consists of four letters (bases)—A, U, G, and C, and the sequence logos consist of a stack of the four letters quantified by 2 bits (1 bit can describe two possible values) at each position in a sequence. The 2 bit high vertical bar is a measure of the relative frequency of the four letters. While the overall height of a stack indicates a degree of sequence conservation, the height of each letter within the stack shows the relative frequency at each position. At each nucleotide position, most to least frequent bases are placed from top to bottom. Overall, sequence of the junctions is largely conserved. Notably, the bases (in a red box) involved in RNA-RNA long-range interactions between H_4 and H_5 are highly conserved, especially the binding receptor (A_{234}, G_{240}, and C_{241}) of the GNRA loop in H_4 which is nearly perfectly conserved.

at each position. The logos are generated using the RNALogo webserver [16]. See Figure 4.2 for the RNA sequence logos of the 4-way junctions.

4.2.3 Modeling and simulation of the apical region

Divide-and-conquer approach to model multiple RNA junction topologies

To tackle multiple consecutive 4-way RNA junctions, we use a divide-and-conquer approach by partitioning the large complex. Each 4-way junction is analyzed with regards to the loop size of single strands between helices; this analysis is coupled to the Junction-Explorer program to help determine coaxial stacking patterns and helical arrangements. Junction-Explorer is based on the random forests data mining algorithm [12] and uses various geometric and energetic parameters as "feature vectors"(which contain information on free energies, loop sizes between junctions, and adenine content) for training. Using the predicted topology for each 4-way junction, we search for all possible combinations of the multiple 4-way junctions to produce combined structures. These potential topologies for the secondary structures are then refined further by incorporating experimental data as constraints (Figure 4.3).

3D modeling of multiple RNA junction structures

Using state-of-the-art 3D modeling programs we build RNA 3D models of FMDV IRES domain 3 combined with experimental data. We primarily use MC-Sym, which utilizes a fragment-based library to obtain all possible structures of RNA junctions [109]. To complement the modeling results, we explore conformational space using NAST, a knowledge-based coarse-grained simulation

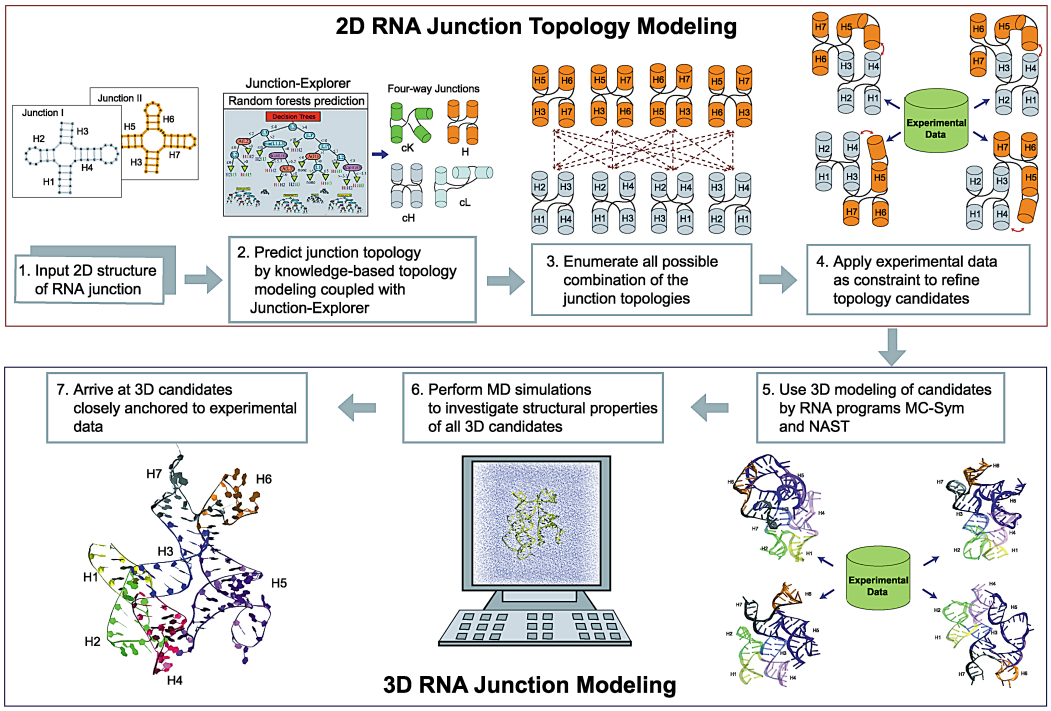


Figure 4.3: Computational procedure for modeling multiple 4-way RNA junction structures. **(1)** Multiple 4-way RNA junction structures are separated into individual RNA junction as input. **(2)** Each junction is analyzed for 2D helical arrangements based on coaxial stacking and junction family type in conjunction with Junction-Explorer. **(3)** These topologies are processed to enumerate all possible junction combinations. **(4)** Available experimental data are applied as constraints to refine the topology candidates. **(5)** 3D models based on the topology candidates are developed using computational programs. **(6)** MD simulations are performed to arrive at **(7)** potential 3D structure.

tool [55]; these two programs have been shown to perform well in predicting native RNA structures [74].

We hypothesize that fewer nucleotides between helices should naturally restrict the orientational flexibility at some degree yielding coaxially stacked helices. Thus, we first model Junction I and II following the 5′ to 3′ direction without constraints for coaxial stacking arrangement. This yields thousands of structures for each junction. Because helical elements in RNA junctions tend to form coplanar arrangements [58], we ranked the predicted structures for coplanarity and collected the best 1,000 structures. These junctions are then assembled by imposing a distance constraint for potential long-range interactions from experimental data (see "Modeling Atomic Junction Structures for the Apical Region" in RESULTS section for more details).

Using NAST, the Nucleic Acid Simulation Tool, we performed a coarse-grained MD simulation for 40 ns (10×10^6 time step) with one tertiary contact between A_{180} and C_{232}/G_{240}. For the 10,000 coarse-grained templates generated, we filter the templates by a potential energy with the cutoff energy of 1,000 kJ.

Molecular dynamics simulations

Each system was solvated with the explicit TIP3P water model in a water box of dimension 10Å on each side. Simulations were performed using the Amber Parmbsc0 force field [111, 17] with sodium ions to neutralize the system charge.

We minimize the system in two steps, first over the water and ion molecules holding domain 3 fixed and, second, with all constraints removed. The minimization was performed using the Powell conjugate gradient algorithm. The initial equilibration was achieved over 60 ps at constant temperature (300 K) and pressure (1 atm), respectively. Pressure was maintained at 1 atm using

the Langevin piston method, with a piston period of 100 fs, damping constant of 50 fs, and piston temperature of 300K. Temperature coupling was enforced by velocity reassignment every 2 ps. Both minimization and equilibration are performed using the NAMD program [112].

For the production run, we simulated a conventional MD trajectory for 100 ns with the Parmbsc0 force field using the NAMD package. The system was simulated at constant temperature (300K) and volume using weakly coupled Langevin dynamics of non-hydrogen atoms, with a damping coefficient of c = 10 ps^{-1} with a 2-fs time step maintaining bonds to all hydrogen atoms rigid. Non-bonded interactions are truncated at 12Å and 14Å for van der Waals and electrostatic forces, respectively. Periodic boundary conditions are applied, and the particle mesh Ewald method is used to calculate electrostatic interactions.

All simulations using the NAMD package were run on IBM Blue Gene/L supercomputer at the Computational Center for Nanotechnology Innovations (CCNI) based in Rensselaer Polytechnic Institute, NY.

Stability of the simulated structures

Our simulated structures contain 116 nucleotides of which 82 involve base pairs. To justify the stability of simulations maintaining secondary structure in trajectories, we computed average distance of base-paired residues in helices (5 base pairs in H_1, 3 in H_2, 8 in H_3, 5 in H_4, 13 in H_5, 3 in H_6, and 4 in H_7) and measured RMS deviations (RMSD) considering all residues and only base pairs with reference to the starting structure as we hypothesize unpaired residues contribute to increase RMSD.

The average distances of all helices in our simulations are below 3Å and most of them are lower than 2.5Å (Figure C.1A, Appendix C). In addition,

RMSD values (considering all except hydrogen atoms) are for the entire system (116 residues) including the paired 82 residues indicating overall stability (Figure C.1B, Appendix C).

Very recently, problems in χ torsion angles of MD simulations for RNA systems have been reported (namely, ladder-like structures that loose the helical twist of A-form RNA conformation, especially in long RNA MD simulations) [102, 119], and improved force fields have been introduced [145, 143]. We have carefully monitored these potential problems but have not observed in our dynamics data.

4.2.4 Entire sequence modeling including the basal region

We model the basal region containing 98 nucleotides (G_{86} to U_{133}, C_{249} to C_{299}) based on 2D information of FMDV C-S8 IRES domain 3 using MC-Sym. Evaluating all the 717 structures based on RMS deviation and clustering analysis yields four candidate models that were chosen from the first four large clusters containing at least 10 structures (Figure C.2, Appendix C) and Figure 4.10). Since overall shapes of these four candidates were similar, we chose a representative model from the largest cluster (Figure 4.10A) to build a complete 3D model of domain 3. Structures of the apical and basal region were merged using a python library of modeRNA [121]. Both minimization and equilibration were performed on the entire domain 3 following the protocol in the "Molecular Dynamics Simulation for the Apical Region" section above.

4.3 Results

4.3.1 Modeling and prediction of the apical region

Sequence Conservation Analysis

Sequence similarity provides evidence for structural conservation and hence essential biological function. Sequence logos of the aligned 318 sequences of FMDV IRES systems suggest that the apical region is largely conserved (Figure 4.2), implying that its secondary structure is constrained under an evolutionary pressure to carry an important biological function for non-canonical IRES-mediated translation initiation. In particular, conservation of the potential binding receptors (G_{229} to C_{232} and G_{240} to C_{243}) of the GNRA loop is near perfect (316 out of 318 sequences); the sequence logos of H_4 marked in the red box (Figure 4.2) indicate that the entire hairpin including the binding nucleotides are conserved almost fully. We also observe that the GUAA sequence appears most frequently with 233 instances (73.2%) followed by GUGA (17%), GCAA (7%), GCGA (2.5%), and GAGA (0.3%) (Table C.1, Appendix C).

Multiple 4-way junction topology prediction

We partition domain 3′s 2D structure into two 4-way junctions and list all possible junction topologies. We denote the two 4-way junctions as Junction I and II following the 5′ to 3′ direction (Figure 4.4). As loop size dictates orientation and flexibility of helices in RNA junctions [73, 87], we build candidate topological models accordingly. Because very few nucleotides are present between helices in both junctions, we consider two coaxial stacking patterns, parallel to each other with a possible crossing at the point of single strand exchange.

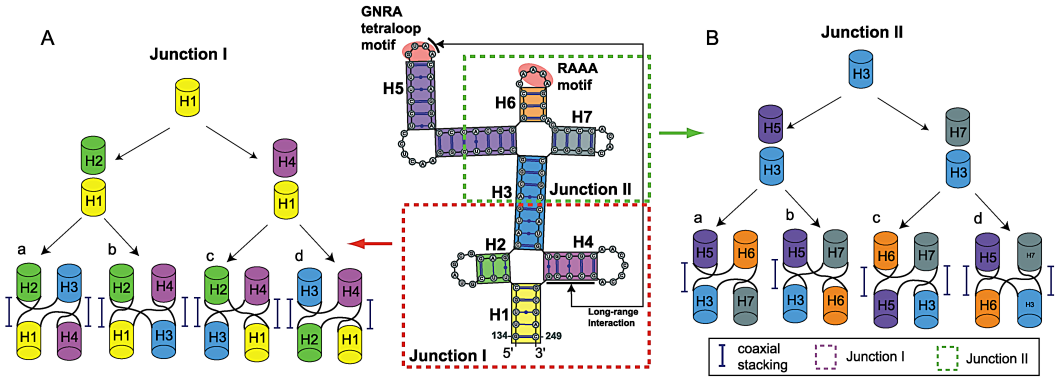


Figure 4.4: Piecing the possible helical arrangements for the two 4-way junctions in domain 3 of IRES. **(A)** Possible junction topologies of junction I with two coaxial stacking following the 5′ to 3′ direction: a) H_1H_4 and H_2H_3 without crossing in the single-stranded region, b) H_1H_4 and H_2H_3 with crossing, c) H_1H_2 and H_3H_4 with crossing, d) H_1H_2 and H_3H_4 without crossing. **(B)** Predicted junction topologies of junction II with one or two coaxial stacking following the 5′ to 3′ direction: a) H_3H_5 and H_6H_7 without crossing, b) H_3H_5 and H_7H_6 with crossing, c) H_5H_6 and H_7H_3 without crossing, d) H_6H_5 and H_7H_3 with crossing.

For Junction I, two types of pairwise coaxial stacking patterns are likely (because no nucleotides are present between helices). Helix H_1 can coaxially stack with either H_2 or H_4 (Figure 4.4A). This results in two coaxial stacking: H_1H_2 with H_3H_4 or H_1H_4 with H_2H_3, as shown in the figure.

Similarly, for Junction II we consider H_3H_5 with H_6H_7 or H_3H_7 with H_5H_6 (Figure 4.4B). However, we speculate that the latter pattern is more likely due to the presence of two nucleotides in a single strand loop between H_6 and H_7 whereas no nucleotides are in other single strands between coaxially stacked helices H_3H_7, H_5H_6 and H_3H_5 (see enlarged view in the middle of Figure 4.4): a strong preference for coaxial stacking has been observed with a smaller loop

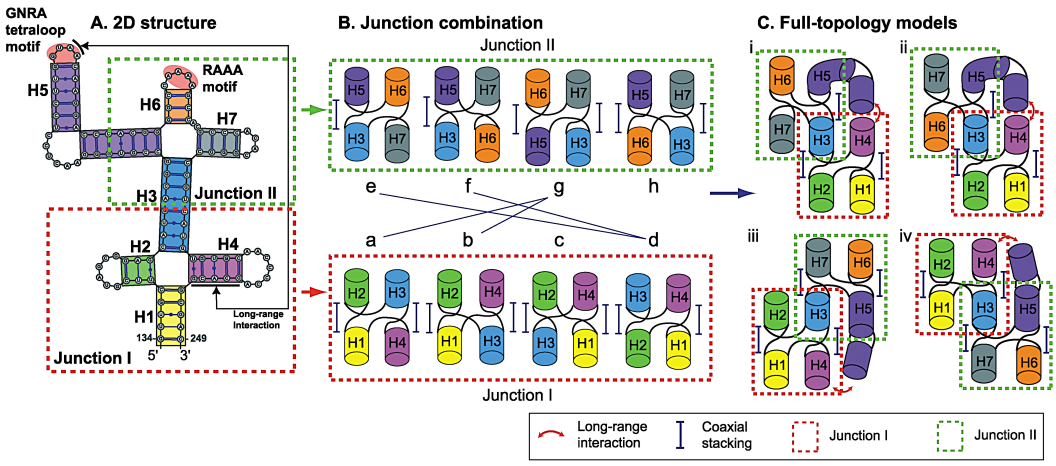


Figure 4.5: Candidate models derived from combinations of two 4-way Junctions I and II of Figure 4.4. **(A)** Secondary structure of domain 3 in FMDV IRES. **(B)** Combinations of the two 4-way junctions considered. To accommodate the long-range interactions between helices H_4 and H_5, helix H_3 must be either parallel or perpendicular to the helices H_4 and H_5 in space. Two arrangements, c and h, of junctions I and II do not satisfy the long-range interactions and are thus eliminated. **(C)** Four complete junction topology models where Junction I (dotted red box) and II (dotted green box) are stitched via helix H_3 considering the GNRA tetraloop long-range interaction between H_4 and H_5.

size [73, 87, 64]. The Junction-Explorer program also predicts a pair of coaxial stacking formation for both 4-way junctions, parallel to each other. On the basis of these combined models, we arrive at four candidate helical arrangements for each junction (Figure 4.4) that correspond to *H* and *cH* family types containing two coaxially stacked helices based on our 4-way junction classification study [73]. Note that only three 4-way junction families *H*, *cH* and *cL* contain two coaxially stacked helices and are distinguished by the angle between the two stacked helices with roughly 0, 180, and 90 degree, respectively. To achieve

a particular configuration for each family, different lengths of single strands between stacked helices are required. While the shape of family H and cH can be achieved with a relatively short single strand, family cL requires a long single strand. Because the junctions in domain 3 contain two nucleotides in single stranded regions at most, we do not consider the cL family as a candidate. Next, we consider these combinations of configurations compatible with experiment.

Considering the nine major junction family types in 4-way RNA junctions [73], the number of ways to pair two 4-way junctions is (9 family types $\times$ 2 different helical arrangements)^2 = 324 when no other information is considered. Using the two possible family types for each junction (Figure 4.4AB), the number of likely conformations becomes (2 family types $\times$ 2 different helical arrangements)^2 = 16 (Figure 4.5). We further consider the potential RNA-RNA long-range interactions between GNRA motif and its distal receptor region from experiment [34] to eliminate some of these 16.

Given that the GNRA tetraloop and its potential receptors are located in helices H_4 and H_5 (Figure 4.5A), we can eliminate some helical arrangements. To make the long-range interactions possible, both H_4 and H_5 are required, in positions either parallel or perpendicular with respect to H_3. Thus, the two configurations c and h in Figure 4.5B can be eliminated because the bridging helix H_3 between Junction I and II is diagonal to H_4 and H_5 (see the helical arrangements in third row (for c) and fourth column (for h) in Figure C.3, Appendix C); these two models are not eligible to make tertiary contacts between H_4 and H_5 due to either the orientation of these two helices that are opposite one another ((c,e), (c,f), (c, h), (a,h), and (b,h)) or some steric clashes ((c,g) and (d,h)) (Figure C.3, Appendix C). In addition, the five pairs of helical arrangements (a, e), (a, f), (b, e), (b, f), (d, g) in Figure 4.5B can also be excluded because

helices H_3, H_4, and H_5 are aligned in the same direction (either to parallel or perpendicular). Four viable models remain Figure 4.5C); the junctions in the final topology models correspond to either H or cH family types, and thus H/H or H/cH combinations are possible overall. See Figure C.3 in Appendix C for all 16 combinations elaborated from Figure 4.5B.

Modeling atomic junction structures

Mutational analysis proposed a non-specific receptor site, $G_{240}CACG_{244}$ in H_4 of Junction I, for the $G_{178}UAA_{181}$ tetraloop of H_5 in Junction II [34]. The two adenosines in the GUAA tetraloop prefer to interact with a pair of C/G base pairs or alternatively a combination of C/G and G/C base pairs. In the potential receptor site, we identify a combination of C_{232}/G_{240} and G_{231}/C_{241} base pairs that was reported as receptors of GUAA loop by an in vitro selection experiment [20]. Note that the C_{232}/G_{240} pair is highly conserved in 130 FMDV sequences while the G_{231}/C_{241} pair is invariant [13] and thus probably significant for correct RNA folding.

Because the energetics of tertiary interactions have not yet been considered, at this stage we model the RNA-RNA long-range interactions by imposing a loose distance constraint of 10Å between helices H_4 and H_5, specifically between A_{180} and $C_{232}G_{240}$ using C1′ atoms.

Sampling these constrained models using MC-Sym reduces the number of models to 267: in Junction I, 160 of these contain stacked helices of H_1H_2 with H_3H_4 while the remaining 107 structures contain stacking of H_1H_4 with H_2H_3. These numbers may reflect the preference of RNA′s helical arrangements in Junction I. Up to now, the coaxial stacking pattern—H_1H_2 and H_3H_4—appears to dominate the possibilities when long-range interactions are considered. After

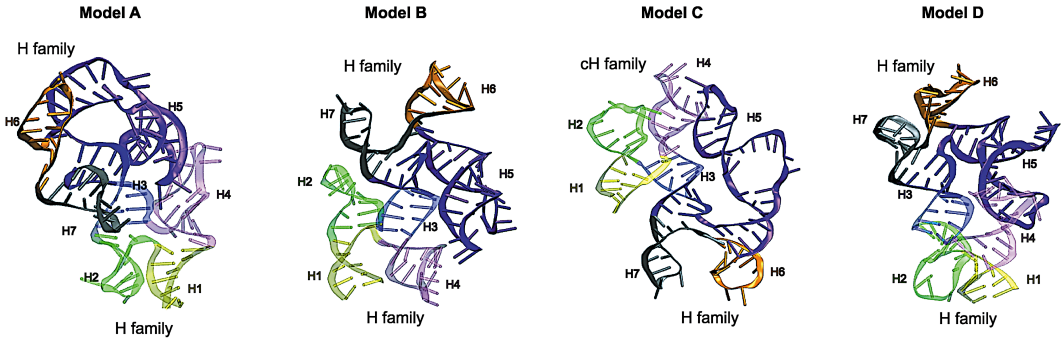


Figure 4.6: Candidate 3D models of the FMDV IRES domain 3. Models A, B, and C correspond to our junction topology models (Figure 4.5C) whereas Model D is new. All structures have two coaxial stacking both in Junction I and II. A combination of H/H or H/cH, but not cH/cH family types is observed in the junctions.

evaluating all the 267 structures by structural similarity based on RMS deviation, clustering analysis, and visual inspection, we arrive at the consensus with initial junction topology models (Figure 4.6). Some variations in H_6 and H_7 may occur due to flexibility introduced by two unpaired nucleotides in the single-stranded region.

Similarly, when we consider another invariant—G_{229}/C_{243} base pair—near the junction core for tertiary interactions, we obtain 52 viable structures; about one-fifth of the structures targeting the C_{232}/G_{240} base pair. We speculate that the relatively small number of sampled structures targeting near the junction core may explain unfavorable potential binding receptor. In fact, the representative 3D structures show that helical arrangements in Junction II are rather distorted than structured.

We also arrive at a new candidate model, a combination of the junction topologies in Figure 4.5B (d) and (g) predicted by MC-Sym (see Fig-

ure C.3(d,g)*, Appendix C). Although excluded while modeling junction topologies, this model may be possible due to a versatile nature of RNA molecules.

The four resulting candidate 3D models are shown in Figure 4.6 A, B, and C correspond to the junction topology models in Figure 4.5C (i), (iii), and (iv), respectively, while D is a new 3D model. Note that we select four 3D models from six clusters considering similarity of overall helical arrangement. All structures have two coaxial stacking in Junction I and II. Interestingly, the 3D model corresponding to the topology model in Figure 4.5C (ii) is not predicted by MC-Sym; this model, different from the three other topology models, has a crossing at the point of single strand exchange in Junction II. We speculate that this particular helical arrangement in Junction II makes it difficult to satisfy the distance constraint criteria in 3D space.

We also explored different RNA conformations by simulating models over 40 ns by one-bead coarse-grained MD simulations using NAST. These simulations yield three representative conformations where helical arrangements are identical in Junction I, but some variations in Junction II (data not shown). Overall, these simulations lead further support to the models in Figure 4.6B and C.

Assessment of structural properties using molecular dynamics simulations

We use MD to supplement the structural studies above and to further explore the feasibility of our structural candidates. Despite algorithmic approximations as well as force field imperfections, MD is widely used to provide further insights into atomic-level interactions and energetic aspects that are not readily revealed from other techniques [124]. Hence, we perform 100 ns MD simulations for all four candidate structures (Models A-D) in Figure 4.6; specifically to in-

vestigate structure stability and potential long-range interactions suggested by experimental data.

4.3.2 Long-range interactions including a novel tertiary contact revealed by MD simulation

Experimental data have proposed intramolecular long-range interactions in domain 3 of FMDV IRES [33, 34]. Although this tertiary contact is required for efficient IRES activity, the specific binding receptor of GNRA tetraloop is yet unknown. To explore this, we track for each trajectory the distances between the GUAA hairpin in H_5 and each of potential target receptors in H_4, specifically between $A_{180}A_{181}$ and $G_{240}CACG_{244}$ (including their complementary residues). Only for Model C we detected two receptor candidatesG_{231}/C_{241} and C_{232}/G_{240} base pairsinteracting with $A_{180}A_{181}$ residues in GUAA tetraloop (Figure 4.7). The trajectory for Model C shows that the two adenosines retain a distance below 3Å. In contrast, only the first adenosine A_{180} of Models A, B, and D retain a distance below 4Å during the initial 12, 15, and 26 ns respectively. In Model C, the average distance between C_{232}/G_{240} pair and A_{180} is 2.1 $\pm$ 0.59Å while C231/G241 pair and A_{181} is 2.0 $\pm$ 0.20Å. These findings suggest that the C_{232}/G_{240} and C231/G241 pairs may be the target receptors of A_{180} and A_{181} residues, respectively.

To further explore the tertiary interaction of model C we consider the Leontis/Westhof nomenclature [86] and analyze the three edges—Watson-Crick, Hoogsteen and Sugar edge—for potential hydrogen bonding interactions. The measured minimum distances between the Sugar edge of each C_{232}/G_{240} and C_{231}/G_{241} base pair with three edges of each A_{180} and A_{181} over the 100 ns

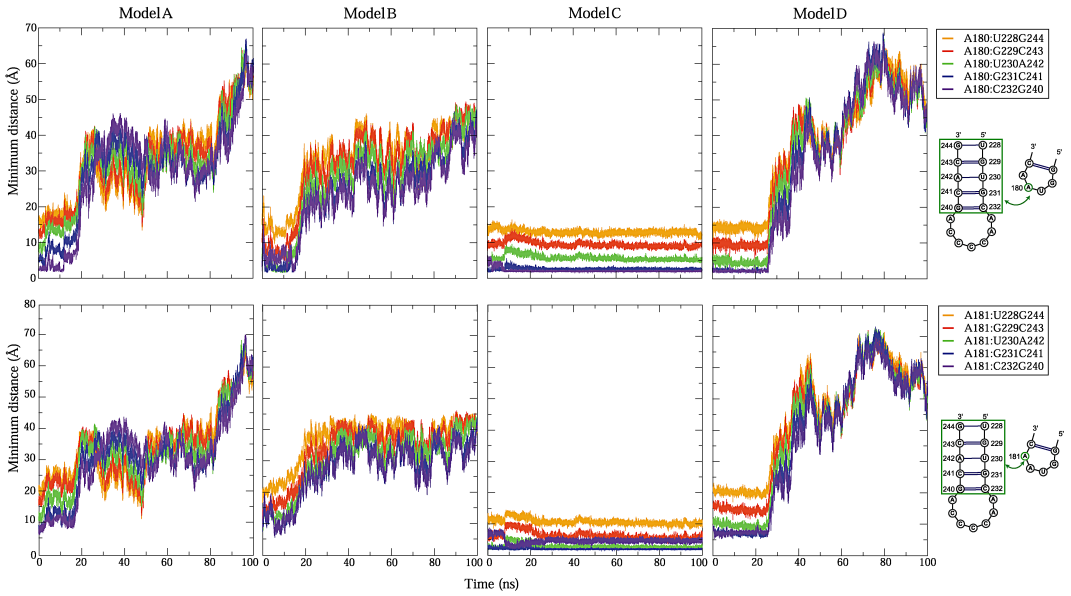


Figure 4.7: RNA-RNA long-range interactions identified by distance measures of atoms between two adenosineA_{180} and A_{181}in the GUAA tetraloop and its potential receptors during the MD trajectories.

time course in Figure 4.8 show tightly formed hydrogen bonding interactions for the Sugar edge/Watson-Crick between the C_{232}/G_{240} pair and A_{180} and Sugar edge/Hoogsteen edge tertiary interactions between G_{231}/C_{241} pair and A_{181}. In addition, we observe tertiary contacts between U_{179} and A_{234} residues via *trans* Watson-Crick/Watson-Crick edge interactions at ∼22 ns. These long-range interactions occur sequentially: at ∼7 ns, ∼20 ns, and ∼22 ns, involving A_{180}, A_{181}, and U_{179}, respectively (Figure 4.8A). These cooperative long-range interactions help stabilize the IRES domain 3.

The corresponding time-averaged secondary structure from the 100 ns dynamics data underscores these three long-range interactions involving the GUAA tetraloop (Figure 4.8B). The $A_{180}A_{181}$ residues in the GUAA tetraloop form hydrogen bonds via non-canonical base pairing interactions with the C_{232}/G_{240} and

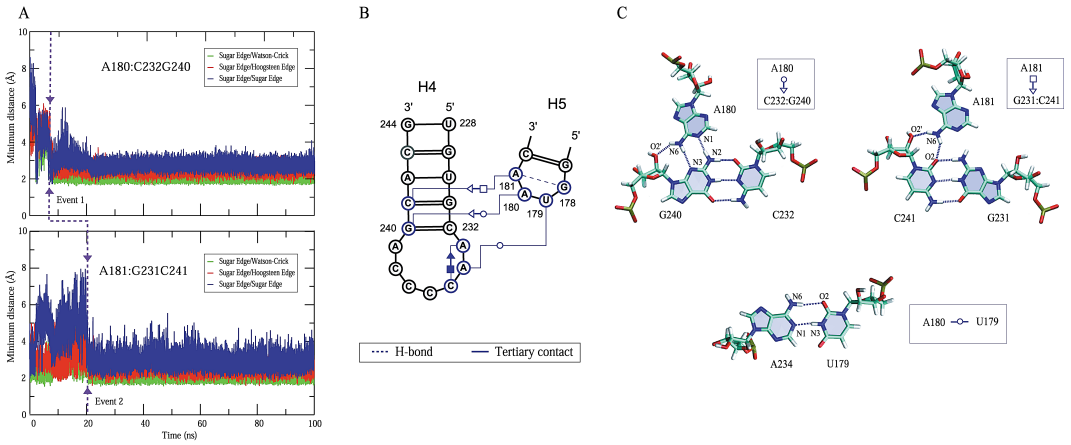


Figure 4.8: Intramolecular RNA-RNA long-range interactions involving GUAA hairpin loop during the Model C MD simulation. **(A)** a minimum distance of atoms for the GUAA tetraloop-receptor long-range interactions. **(B)** long-range interactions in a time-averaged secondary structure obtained from the dynamics data. **(C)** atomic details of these three tertiary contacts involving U_{179}, A_{180} and A_{181} residues in the GUAA loop and their binding receptors.

G_{231}/C_{241} base pairs, respectively; specifically, *trans* Sugar edge/Watson-Crick edge where N_1 and N_6 atoms of A_{180} interact with N_2, N_3 and O'_2 atoms of G_{240} and Sugar edge/Hoogsteen edge interaction where N_6 atom of A_{181} forms hydrogen bonds with O_2 and O'_2 atoms of C_{241} (Figure 4.8C). The U_{179} and A_{234} residues interact via *trans* Watson-Crick/Watson-Crick edge interactions involving N_1 and N_6 atoms of A_{234} with N_3 and O_2 atoms of U_{179}. As demonstrated by in vitro selection experiment [20], this U_{179}:A_{234} tertiary contact promotes the loop-helix long-range interactions.

4.3.3 Contribution of long-range interactions to the structural organization in domain 3

Because the single-stranded region is more dynamic and flexible than double-stranded helices, we speculate that the five hairpins and one long internal loop present in the system may contribute to the overall exhibit of the structure. Thus, the tertiary contacts in domain 3 of IRES may restrict these fluctuations and therefore help recruit ribosomes for viral protein synthesis. Below we further analyze dynamics data for all four models to discern the contributions of the long-range interactions to structural stability as well as organization based on root-mean-square (RMS) fluctuation and radius of gyration (Rg).

In the RMS fluctuation plot (Figure 4.9A), we observe six peaks which correspond to hairpins and an internal loop. Among them, two highest peaks are from hairpins located in the helices H_4 and H_5. Interestingly, these helices involve in the long-range interactions and have been emphasized for its important role in IRES activity.

Overall, the RMS fluctuations of all four models follow a similar trend, albeit at different scales. Overall, Model A, B and D fluctuates widely with the values from $\sim$2.5 to $\sim$21Å , whereas Model C ranges between $\sim$2 and $\sim$7.5Å which is about a four-fold decrease. Notably, the GUAA tetraloop in H_4 fluctuates between 12 $\sim$ 18Å for models A, B, and D, whereas Model C experiences only about $\sim$2.5Å deviation; this underscores the potential stabilizing role of the tertiary contacts. The long-range interactions appear to stabilize not only adjacent stem-loops, but also the entire structure of IRES domain 3.

From the combined data above, involving bioinformatic, experimental, and MD modeling data, we propose a theoretically feasible tertiary structure for the

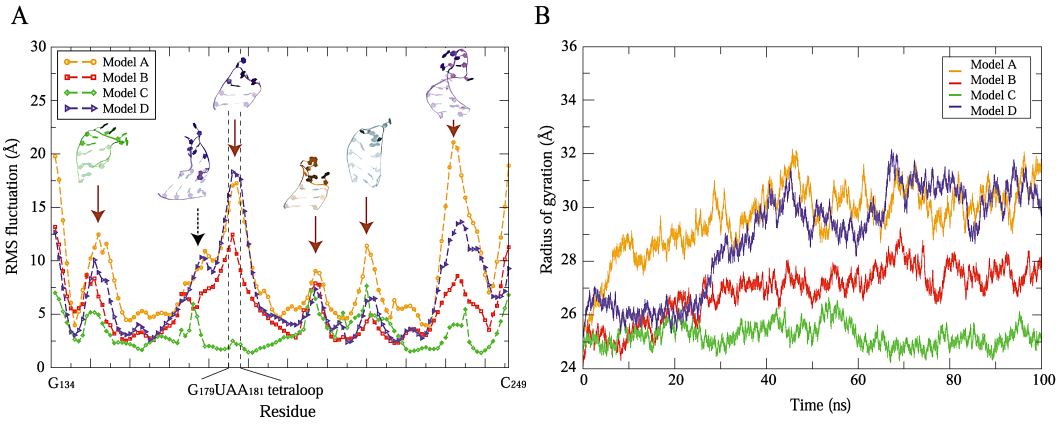


Figure 4.9: Root-mean-square (RMS) fluctuations **(A)** and radius of gyration (Rg) **(B)** measures for four candidate 3D models. In the RMS fluctuations, high peaks (dotted black arrow for internal loop and solid brown arrows for hairpins) correspond to unpaired regions shown as solid color in the structures.

apical region in FMDV IRES domain 3 (Figure 4.11). Here the non-canonical long-range interactions occur between the GUAA tetraloop in helix H_5 and the distal region in helix H_4. The overall configuration is highly structured; each 4-way junction contains two coaxial stacks parallel to each other and Junction I has a crossing at the point of strand exchange. Junctions I and II are classified as family type *cH* and *H*, respectively, according to the nomenclature in [73]. The three helices H_3, H_4, and H_7 are coaxially stacked together. Each of these two 4-way junctions is nearly planar and these two planes are perpendicular to each other.

4.3.4 Modeling of entire domain 3

The basal region contains a long internal loop formed by 98 nucleotides (G_{86} to U_{133} and C_{249} to C_{299}). Sampling this region using MC-Sym produces 717

models from which four representative structures were selected from the four largest clusters (Figure C.2, Appendix C and Figure 4.10). The overall shape and orientation of the helical axis in these four structures are relatively similar, with differences explained by the flexibility of bending in unpaired bases. Thus, our candidate model for the basal region is chosen from the largest cluster (Figure 4.10A), and then the apical region model (Figure 4.11) was combined to it to complete a 3D model of the entire domain 3; the minimum distance between the basal (C_{250}) and apical (G_{194}) regions is 23.5Å; the orientation of the basal region (turned away from the apical region) suggests that the former region is unlikely to be involved in RNA folding of the junctions of the latter region (Figure 4.12).

4.4 Discussion

Picornavirus IRES elements are considered as efficient regulatory RNAs which make possible initiation of translation for viral RNAs. FMDV requires RNA binding proteins such as translation initiation factors (eIFs) and IRES trans-acting factors (ITAFs) that can affect IRES activity; for example, domain 2, 4 and 5 provide binding sites for cellular proteins including PTB, eIF4G, eIF3 and eIF4B [96].

The FMDV IRES domain 3, often denoted as a central domain, consists of two structural elementsa long internal loop in the basal region and 4-way junctions in the apical region; each of which is ∼50% of the entire sequence. Sequence logos of the 318 aligned FMDV IRES sequences show that the apical region of domain 3 is highly conserved (Figure 4.2); in particular, hairpin loop H_4 which contains the potential binding receptors of the GNRA tetraloop is

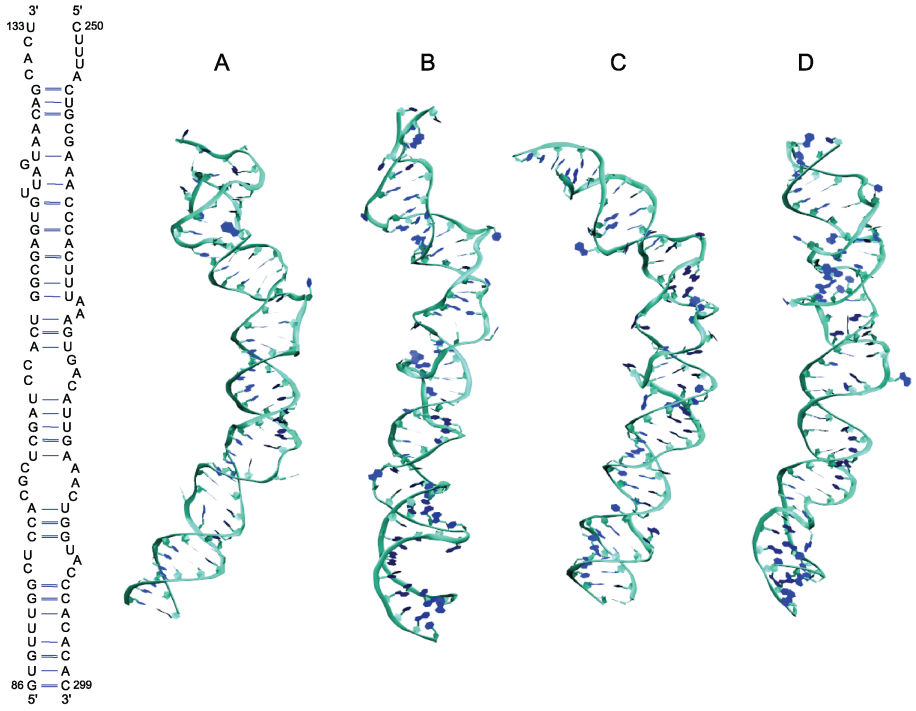


Figure 4.10: Candidate 3D models of the basal region in FMDV IRES domain 3. These four representative models are selected from the first four large clusters containing **(A)** 594, **(B)** 64, **(C)** 23, and **(D)** 13 structures, respectively (see (Figure C.2 B, Appendix C).

nearly perfectly conserved. The GUAA loop in H_5 is also strongly preferred in FMDV IRES systems. This analysis suggests that these structural elements provide an important role in maintaining the functional 3D structure of FMDV IRES domain 3.

It was determined by biochemical experiments that the apical region is a self-folding structural element due to the intramolecular RNA-RNA interactions involving the crucial GNRA motif [33, 116, 126]. This region has thus been suggested to contribute significantly to the structural organization and stability of domain 3, and to the critical function of IRES activity [93, 34, 31, 32]. IRES-mediated translation initiation is closely linked to structural organization in domain 3, specifically the apical region formed by two 4-way junctions enabling the RNA-RNA intramolecular interactions. Thus, we focused on the apical region of domain 3 to decipher the spatial arrangement of the RNA fold that is a prerequisite essential step to understand the initiation mechanism of translation.

Grounded in our recent RNA 4-way junction classification study and the Junction-Explorer program [73, 76], we have constructed possible junction topologies for domain 3 where a pair of coaxial stacks are arranged parallel to each other in the presence and absence of a crossing at the point of strand exchange (Figure 4.11). Utilizing only the information for the helical arrangements—*H* and *cH* family types, we built 16 candidate topologies (Figure 4.5) and these were reduced to four after applying constraints from experimental data regarding the GNRA tetraloop-receptor long-range interactions. Our next step in modeling was employing MC-Sym to explore conformational space using experimental data. The combined data from junction topology and 3D modeling produced four representative structures where three of the four confirmed the junction topology models. We speculate that the excluded model

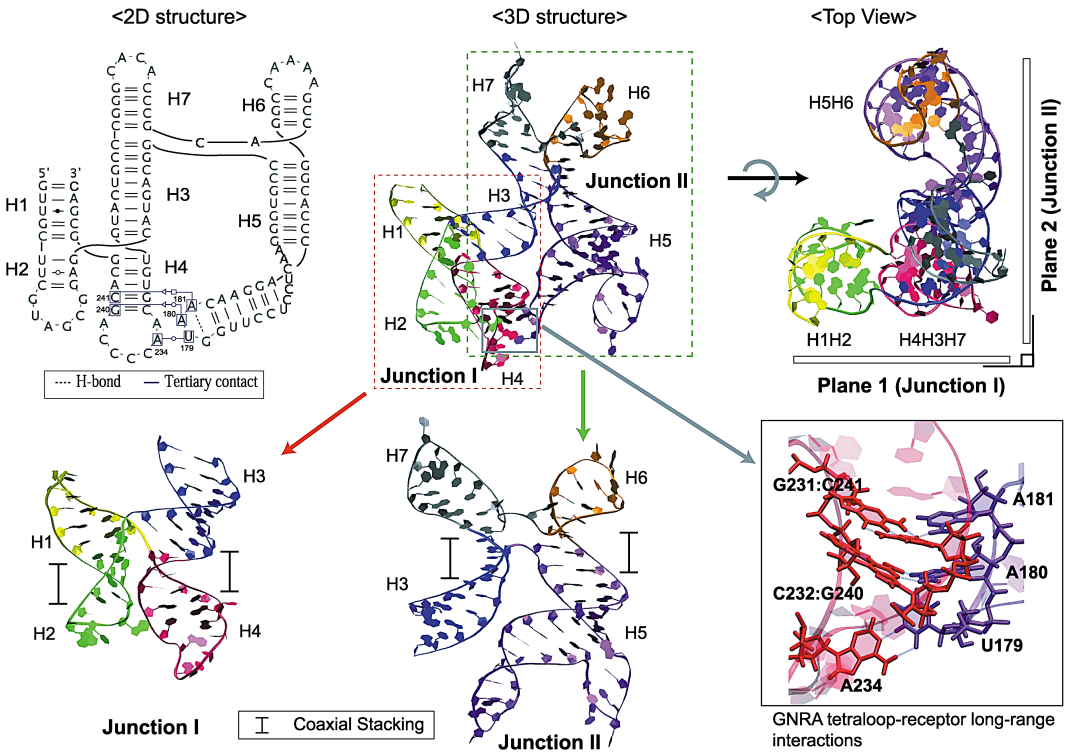


Figure 4.11: Time-averaged tertiary structure of domain 3 taken from the 100 ns dynamics data (top middle) where the long-range interactions occur between helices H_4 and H_5 (details shown at bottom right). Both Junctions I and II contain two coaxial stacking, parallel to each other and Junction I with a crossing in the single-stranded region (bottom left and middle for Junction I and II, respectively). Both junctions are planar locally and are arranged in a perpendicular orientation to each other globally (top right); note that the three helices H_4, H_3, and H_7 are coaxially stacked all together.

does not satisfy geometric criteria due to steric clashes. Using MD simulations, we attempted to identify geometrically accessible binding receptors to GNRA tetraloop for the long-range interactions. Among the five residues in the potential receptor site, the bases near the hairpin loop emerged viable over one near the junction core in helix H_4; they also have great potential to form long-range interactions with the GNRA tetraloop in H_5.

Specifically, the MD simulations revealed a GUAA tetraloop binding site in addition to a novel tertiary interaction in model C (Figure 4.6). The two adenosines A_{180} and A_{181} form hydrogen bonds with the receptors C_{230}/G_{242} and G_{231}/C_{241} base pair, respectively. The dynamics data also suggest a U:A tertiary contact which enhances the structural stability (U_{179} in GUAA tetraloop interacts with A_{234} in a hairpin loop of H_4), an interaction also observed by an vitro selection experiment [20]. Interestingly, these tertiary interactions form sequentially.

A previous study suggested that RNA-RNA long-range interactions involving an RAAA motif occurs in the presence of GNRA tetraloop long-range interactions as well as Mg^{2+} ions [34]. We have not observed this RAAA motif, but speculate that the distant contacts associated with the RAAA motif might occur when the current RNA system (G_{134}...C_{249} residues) is extended to include 25 more residues (U_{121}...G_{133}, U_{250}...A_{261}). Such an extended system has greater potential for the long-range interactions (Figure C.4, Appendix C). Due to current limitations of the force field for treating divalent ions [98, 128, 118, 117], we have not attempted to include magnesium ions in our system.

In picornavirus, types 1 and 2 of IRES species exist. FMDV IRES belongs to type 2 whereas poliovirus IRES to type 1. Although the overall contents of sequence and 2D structure are different, these two IRES systems share the

GNRA motif. NMR data of stem-loop in domain IV revealed the L-shaped conformations that are required in order to provide a protein binding site. However, little is known how the L shape is achieved and maintained. In FMDV IRES domain 3, helix H_5 contains similar 2D structure of loop B which includes the GNRA motif. In our dynamics simulations, we observe that H_5 forms an L shape configuration in the presence of long-range RNA-RNA interactions. The overall shape of H_5 agrees well with the NMR data (Figure C.5, Appendix C). However, the shape of H_5 in the absence of long-range interactions is variable, with potential diverse phases (S-shaped or U-shaped). Thus, we speculate that long-range interactions involving the GNRA motif have a role in stabilizing the L-shaped loop B in poliovirus IRES.

Based on the above modeling of the apical, self-folding region of IRES domain 3 containing 4-way junctions and the experimental data discussed above [33, 116, 126] combined with our extended modeling of the entire sequence of FMDV IRES domain 3, we hypothesize that the influence of the basal region on structural stability and organization of the junctions is not primary. This is because the basal region is set apart from the junction domains in the apical region with a minimum distance of 23.5Å (Figure 4.12).

Although our combined modeling strategy involves many proven approaches and is closely anchored to available experimental data, it is not possible to rule out other plausible overall 3D structures. Further studies using the candidate models for long-time MD studies or with advanced sampling techniques and investigation of potential receptors for RAAA motif may be useful. Yet, the overall 3D configuration reached in Figure 4.11 and the suggested long-range interactions in the central domain of FMDV IRES provide insights into the potential role of the long-range interactions for structural stability and organi-

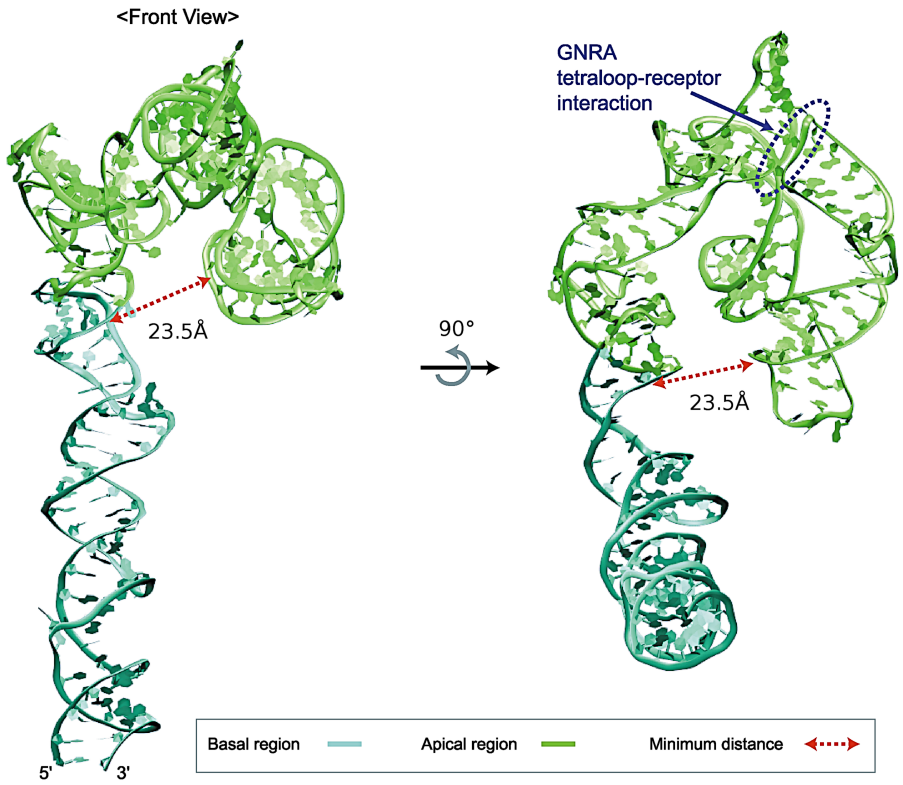


Figure 4.12: 3D model of the entire sequence in domain 3. Domain 3 consists of basal and apical regions; the corresponding structural elements are an internal loop and 4-way junctions, respectively. The minimum distance between the two regions is 23.5 Å that the basal region is not likely involved in RNA folding of the junctions. 3D model of the entire sequence in domain 3. Domain 3 consists of basal and apical regions; the corresponding structural elements are an internal loop and 4-way junctions, respectively. The minimum distance between the two regions is 23.5 Å that the basal region is not likely involved in RNA folding of the junctions.

zation of IRES domain 3 and thus may help in further analysis of the structure, mechanism, and function of viral RNAs. Ultimately, structures may lead to the development of antiviral drugs that inhibit IRES activity and thus virus multiplication.

Chapter 5

Interconversion between Parallel and Antiparallel Conformations of a 4H RNA junction in Domain 3 of Foot-and-Mouth-Disease Virus IRES Captured by Dynamics Simulations

5.1 Introduction

5.1.1 Dynamic characteristics of 4-way RNA junctions

RNA junctions play crucial roles in directing the overall folding of RNA molecules as well as in a variety of biological functions. In particular, there

has been great interest in the dynamics of RNA junctions. A prominent example is the 4-way junction of hepatitis C virus (HCV) internal ribosome entry site (IRES), a specific RNA structure for internal translation initiation. The junction in HCV IRES is important in the overall IRES structure conformation. Two different conformations of the RNA junction were reported—parallel and antiparallel structures—by both crystallography and single-particle cryo-EM techniques [58, 185]. Later, Lilley *et al.* [192] studied the IRES RNA junction using comparative gel electrophoresis and fluorescence resonance energy transfer (FRET) and showed that two different conformations can interconvert via continuous transitions (Figure 5.1C). Such dynamic characteristics of 4-way junctions often have functional significance. For example, the 4-way junction in U1 snRNA plays a crucial role in organizing the whole RNA molecule via RNA-RNA interactions [184, 186]; the junction in the hairpin ribozyme forms a catalytic site for the RNA self-cleavage reaction [187]; and the junctions in viral mRNAs are essential for translating the maturation protein-encoding gene [188]. All these 4-way junctions contain fully base-paired four helical arms with no additional nucleotides at the point of strand exchange, an architecture termed 4H junction [90], as shown in (Figure 5.1).

5.1.2 Folding pathways of 4H RNA junctions

Such 4H junctions often appear in self-folding RNA molecules. The 4H junction adopts a compact fold with pairwise coaxial stacking of helices [183, 184, 137] and is known to fluctuate between multiple conformations (e.g., parallel, antiparallel structures) during the search for the most stable native structure [192, 200, 48, 201]. These conformational states consist of different helical stacking

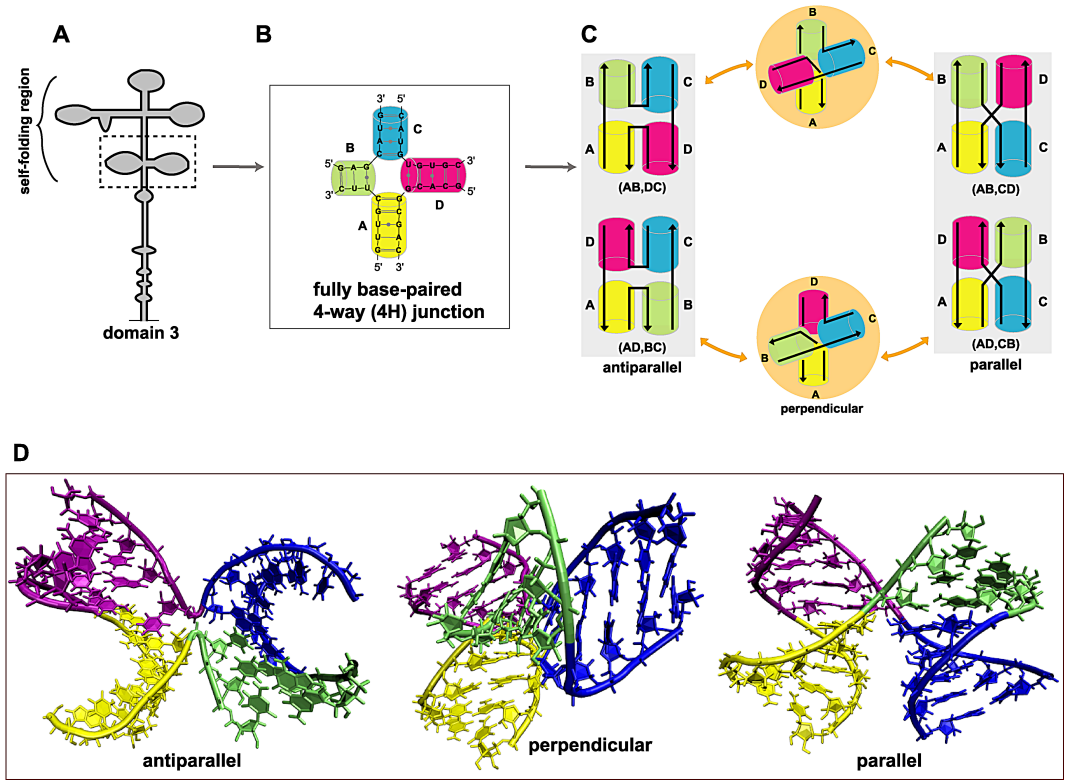


Figure 5.1: A fully base paired 4-way junction with possible conformations. **(A)** Schematic representation of domain 3 in FMDV IRES with a fully base paired 4-way junction (4H junction) in the inset. **(B)** Secondary structure of the 4-way junction in domain 3, deduced from RNA structure probing experiment. **(C)** A possible pathway of the 4H junction that can interconvert between parallel and antiparallel conformations via a perpendicular intermediate with alternative stacked conformers. **(D)** A possible pathway of the 4H junction that can interconvert between parallel and antiparallel conformations via a perpendicular intermediate with alternative stacked conformers.

conformers, depending on the local sequence content at the branch point and ionic strength, with either parallel (AB, CD or AD, CB) or antiparallel (AB, DC, or AD, BC) arrangements (Figure 5.1). While the mechanism of interconversion is not fully understood, the experimental data suggest two possibilities. One intermediate involves a helical rearrangement by (partial) unstacking of the helices, and another possibility is a rotation between helical axes while maintaining the stacked conformers intact [48].

5.1.3 Investigation of structural properties of the 4H RNA junction using molecular dynamics simulations

Molecular dynamics (MD) simulations are a well-established method to investigate structural properties of biomolecules at an atomic level. Previous modeling and dynamics studies of large RNAs with junctions include riboswitches to investigate conformational dynamics upon substrate binding [172, 202, 203], ribosomal subunits to explore dynamic properties with respect to the biological functions [175, 204], and viral RNAs to predict and characterize structural models [205, 148]. Domain 3 in FMDV IRES is the largest structural element containing multiple 4-way junctions. Its apical region, a self-folding RNA molecule, directs adjacent stem-loops for correct RNA folding [31]. Here we investigate the ambient fluctuations of a free 4H junction found in FMDV IRES domain 3 (Figure 5.1), by MD simulations. The sequence of the 4H junction is highly conserved, implying that its secondary structure is constrained under evolutionary pressure to deliver important biological functions [148]. Indeed, the 4H junction provides potential binding motifs in helix D for RNA-RNA and

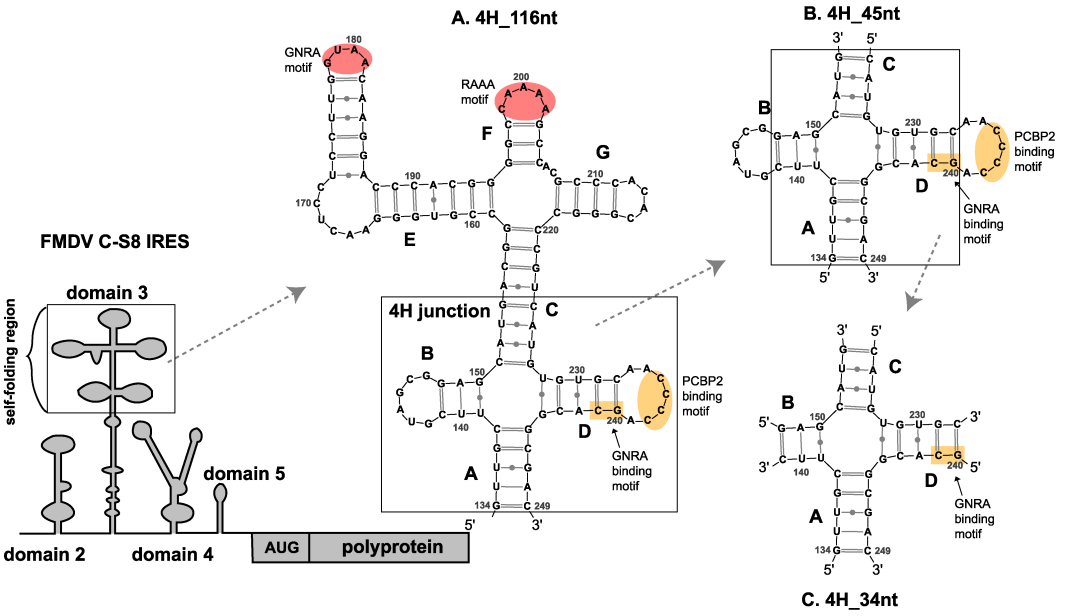


Figure 5.2: A secondary structure of domain 3 in FMDV C-S8 IRES. Three systems of truncated domain 3 with different sizes are prepared with a sequence length of 116 *nt* **(A)**, 45 *nt* **(B)**, and 34 *nt* **(C)**.

RNA-protein interactions, crucial for IRES activity, involving the GNRA (N is any nucleotides; R is A or G) tetraloop and polyC binding protein (PCBP2), respectively (Figure 5.2) [34]. Thus, the dynamic characteristics and folding pathways of this 4H junction are important for understanding the junction's role in the folding and activity of domain 3.

5.1.4 Overview of Results

Our simulations capture the transition dynamics and folding pathway of this IRES-associated 4H junction in domain 3. We observe a concerted, virtually barrier-free, transition from antiparallel (AD, BC) to perpendicular (AD⊥BC), and from perpendicular (AD⊥BC) to parallel (AD, CB) conformations, driven

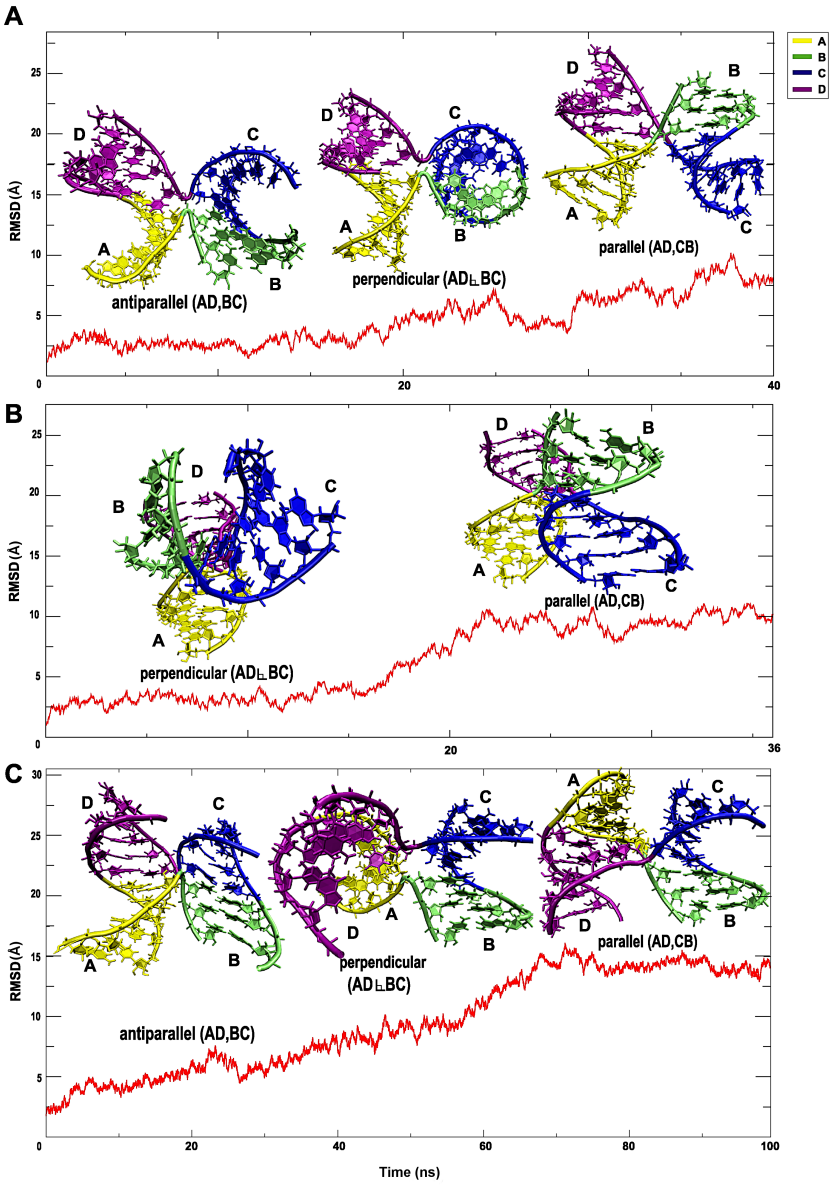


Figure 5.3: Conformational change of the 4H junction in FMDV IRES domain 3. While keeping pairwise coaxial stacking of helical arms intact, a transition from antiparallel or perpendicular to parallel states (simulation name: *Perp_34nt* **(A)**, *Perp_45nt* **(B)**, and *Anti_116nt_2* **(C)**) was observed, driven by a rotation between the helical axes.

by a rotation between axes of the coaxially stacked helices (Figure 5.3). During these interconversions, the pairwise coaxial stacking of helices remains intact. Our captured transitions in the MD trajectories exhibit various inter-helical angles and involve a perpendicular intermediate, less stable than the two other conformations. Because the GNRA tetraloop-receptor long-range interactions are important for the folding of IRES, the transient perpendicular intermediate connecting the parallel and antiparallel configurations may be beneficial for overall IRES structure organization since the maintained coaxial stacks may help direct essential tertiary-contact formation.

5.2 Computational Methods

5.2.1 RNA target sequence and 3D structure modeling

Domain 3 of FMDV IRES is a self-assembling RNA that is 214 *nt* long. Using the sequence and secondary structure of truncated domain 3 in FMDV C-S8 IRES, we modeled 3D structures that include the 4H junction with three different system sizes—34, 45, and 116 *nt* (Figure 5.2)—following the same modeling procedure described in Chapter 4.

In brief, we developed a computational divide-and-conquer strategy for modeling candidate tertiary structures for the IRES RNA [148]. We began by modeling junction topology candidates and then built atomic 3D models consistent with available experimental data using MC-Sym [109], which utilizes a fragment-based library to obtain all possible RNA structures. Because fewer nucleotides between helices restrict the structural flexibility yielding coaxially stacked helices, no constraints were applied to yield stacked conformers. We modeled

the 4H junction (G_{134},..., G_{145}, C_{224},..., C_{249}) by following the 5′-to-3′ direction without and with hairpin loops in the helices B and D to produce two different RNA systems (34 and 45 *nt*). To select final candidates corresponding to the three different conformations (parallel, perpendicular, and antiparallel), we used a clustering analysis followed by a visual inspection of representative structures from each cluster. We generated the remaining structural elements (A_{146},..., U_{223}), composed of a 4-way junction and a helix (U_{172},...,A_{187}), connected by a long bulge (A_{166},...,C_{171}). Specifically, the junction and helix were modeled separately by following the 5′-to-3′ direction, and then assembled via the long bulge. The large RNA system (116 *nt*) was modeled by combining these two structural entities—4H junction and the remaining structure—by imposing a distance constraint of 10Å (between A_{180} and C_{232}-G_{240} using C1′ atoms) for the GNRA tetraloop-receptor long-range interactions. Similar to the smaller junction systems above, final candidate structures were selected based on a clustering analysis and visual inspection. See Chapter 4 for full details.

5.2.2 Studied RNA systems

We prepared three different sets of RNA systems differing by sizes and helical arrangements (Table 5.1 and Figure 5.2). The first set, 34 *nt*, contains parallel (system name: *Para_34nt*), perpendicular (system name: *Perp_34nt*), and antiparallel (system name: *Anti_34nt*) configurations. The second set, 45 *nt*, consists of parallel (system name: *Para_45nt*), perpendicular (*Perp_45nt*), and antiparallel (system name: *Anti_45nt*) configurations. The third set, 116 *nt*, includes parallel (system name: *Para_116nt*) and antiparallel (system name: *Anti_116nt_1* and *Anti_116nt_2*) configurations with different stacking conform-

Table 5.1: List of simulations of 4H RNA junctions in FMDV IRES domain 3. The name of simulated structure is based on the three different systems shown in (Figure 5.2).

Simulation name	size (nt)	Starting structure configuration	Force field	Trajectory length (ns)	Conformational change
Anti_34nt	34	antiparallel	bsc0χ_{OL3}	92	No transition
Perp_34nt	34	perpendicular (AD,BC)	bsc0χ_{OL3}	40	Fluctuation between antiparallel and perpendicular, and transition from perpendicular to parallel
Para_34nt	34	parallel (AB,CD)	bsc0χ_{OL3}	219	No transition
Anti_45nt	45	antiparallel	bsc0χ_{OL3}	104	No transition
Perp_45nt	45	perpendicular (AD,BC)	bsc0χ_{OL3}	36	Transition from perpendicular to parallel
Para_45nt	45	parallel (AD,BC)	bsc0χ_{OL3}	30	No transition
Anti_116nt_1	116	antiparallel	bsc0	100	No transition
Anti_116nt_2	116	antiparallel (AD,BC)	bsc0	100	Transition from antiparallel to perpendicular, and from perpendicular to parallel
Para_116nt	116	parallel (AB,CD)	bsc0	100	No transition

ers, either D with A or B with A.

The RNA systems in the first set are composed of four helical arms without hairpin loops on helices B and D, whereas the second set contains hairpin loops that may form loop-helix or loop-loop tertiary interactions. The third set contains the 4H junction plus additional structural elements that could establish tertiary contacts in various forms including the long-range interactions involving the GNRA and RAAA motifs (Figure 5.2).

5.2.3 Molecular dynamics simulations

We solvated each system with the explicit TIP3P [206] water model in a water box of dimension 10Å on each side using tLeap from the AmberTools package [207]. Simulations were performed using the Amber parmbsc0 and

parmbsc0χOL3 force fields [111, 17, 145, 143] with sodium ions to neutralize the system charge (Figure D.1, Appendix D).

The choice of a force field for RNA is often crucial to achieve meaningful and reliable trajectories. We test two latest Amber force fields, parmbsc0 and parmbsc0χOL3, for RNA–the latter representing an improved version of parmbsc0 for χ torsion angles. We found both force fields perform equally well for our simulated systems, not exhibiting any χ torsion angle related problems [102, 119]. However, we observe a base pair disruption at the helix end of B, formed by three base pairs without a hairpin loop, in our smallest RNA systems (*Anti_34nt* ($\sim$54 ns) and *Para_34nt* ($\sim$195 ns) in Table 5.1). This likely occurs because a helix composed of ≤ 3 base pairs may be too small to maintain the overall structural stability corresponding to the secondary structure.

We minimized each system in two steps, first over the water and ion molecules holding domain 3 fixed and, second, with all constraints removed. The minimization was performed using the Powell conjugate gradient algorithm [208]. The initial equilibration was achieved over 60 ps at constant temperature (300 K) and pressure (1 atm), respectively. Pressure was maintained at 1 atm using the Langevin piston method, with a piston period of 100 fs, damping constant of 50 fs, and piston temperature of 300 K. Temperature coupling was enforced by velocity reassignment every 2 ps. Both minimization and equilibration were performed using the NAMD program [112].

For the production run, each system was simulated at constant temperature (300K) and volume using weakly coupled Langevin dynamics of non-hydrogen atoms, with a damping coefficient of c = 10 ps^{-1} with a 2-fs time step maintaining bonds to all hydrogen atoms rigid. Non-bonded interactions were truncated at 12Å and 14Å for van der Waals and electrostatic forces, respectively. Peri-

odic boundary conditions were applied, and the particle mesh Ewald method was used to calculate electrostatic interactions.

All simulations using the NAMD package were run on the local clusters at New York University and the IBM Blue Gene/L supercomputer at the Computational Center for Nanotechnology Innovations (CCNI) based in Rensselaer Polytechnic Institute, NY.

5.2.4 Principal component analysis (PCA)

To identify the most significant conformational degrees of freedom of a system, dynamics trajectories of 4H junctions were analyzed using PCA [209]. PCA describes the overall dynamics of systems with collective essential motion. The approach is based on the positional n×n (where n=3×number of atoms N) covariance matrix, C, defined as

$$C = \left[\left(r_i - \langle r_i \rangle \right) \left(r_i - \langle r_i \rangle \right) \right],$$

where r_i and r_j are position vectors of two atoms i and j in the fitted structure and the angular brackets ($\langle ... \rangle$) denote the average over all sampled conformations.

By diagonalizing the covariance matrix C, the eigenvectors, V, and their corresponding eigenvalues, λ, are obtained defined as

$$V^T C V = \Lambda, \ or \ C V_n = \Lambda_n V_n,$$

where Λ is the diagonal matrix, diag(λ_1, λ_2,..., λ_{3N}), with eigenvalues λ_i and n=1,2,...,3N.

To remove rotational and translational motions of the trajectory, we use the least squares method to fit the trajectories to its initial configuration as

a reference structure. Each eigenvector V_n defines the direction of motion of N atoms as an oscillation about the average structure $\langle X \rangle$. The normalized magnitude of the corresponding eigenvalue is a measure of the amplitudes of motion along the eigenvector V_n as calculated by $\lambda_i \setminus \sum_i \lambda_i$ and organized in decreasing order. Thus, λ_i represents the largest positional fluctuation and λ_n the least.

5.3 Results

5.3.1 Global dynamic motions of the 4H junction

The 4H junction in FMDV IRES domain 3 (Figure 1) is defined by the base sequence—C_{138}, U_{139},..., G_{149}, C_{150}, ..., G_{227}, U_{228},..., G_{244}, and G_{245}—for the four helical arms at the junction center. We label each helical arm by A through D following the $5'$ to $3'$ direction that consists of canonical Watson-Crick base pairs and three G-U wobble pairs (Figure 5.1). Our starting 3D models of parallel, perpendicular, and antiparallel configurations contain pairwise coaxial stacking of helical arms as all 4H junctions studied experimentally [184].

The major folding pathway of the 4H junction suggested by experimental data [200, 48] involves fluctuations between parallel and antiparallel configurations with two possible intermediates: one is via a helical rearrangement caused by partial or full unstacking of the helices due to insufficient cation binding to the junction; the other is via a rotation between axes of two stacking conformers (Figure 5.1C).

In our collective dynamics data for nine different systems (see Table 5.1), various helical arrangements of the junction including parallel, perpendicular,

and antiparallel are sampled. These conformations are all connected via a rotation between coaxially stacked helices which exhibits various inter-helical angles. Specifically, three simulated systems with different sizes of 34, 45, and 116 *nt* (simulation name: *Perp_34nt*, *Perp_45nt*, and *Anti_116nt_2* listed in Table 5.1) exhibit such transitions within ∼6 ns: *Perp_34nt* shows fluctuation between antiparallel and perpendicular configurations, and a transition of perpendicular to parallel configurations followed by fluctuation between perpendicular and parallel states; *Perp_45nt* exhibits a transition from perpendicular to parallel conformations; and *Anti_116nt_2* fluctuates gradually from an antiparallel to a perpendicular intermediate followed by a transition from perpendicular to parallel states (Figure 5.3).

The other six simulated systems remain in one conformation. The junctions are likely stabilized by coaxially stacked helices or tertiary interactions. Namely, the systems of *Anti_34nt*, *Para_34nt*, and *Anti_45nt* appear to be stabilized by pairwise coaxial stacking, while *Para_45nt*, *Anti_116nt_1*, and *Para_116nt* exhibit both coaxial stacking and RNA-RNA tertiary interactions involving helices B and D that restrain the junction. We analyze further the conformational changes below.

5.3.2 Dominant motion captured by principal component analysis (PCA)

PCA of the dynamics trajectories of the 4H junction captures the dominant collective motion that occurs during the conformational changes. The first four eigenvalues (denote as PC1,..., PC4) of the PCA capture 91% of the overall motion: PC1, 65%; PC2, 20%; PC3, 4% and PC4, 2%. Figure 5.4 shows

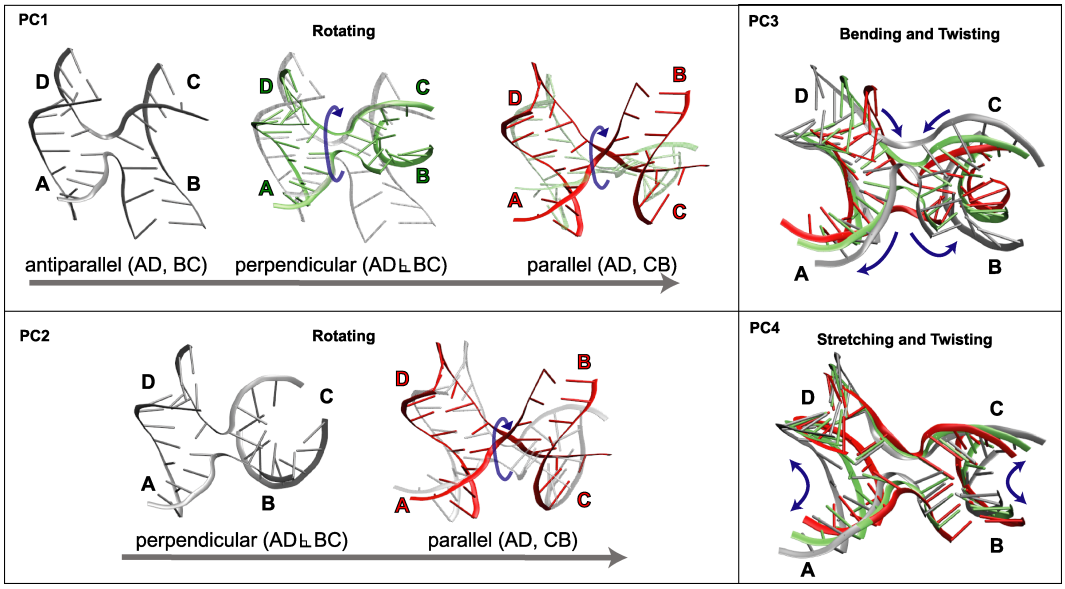


Figure 5.4: Major motions captured by PCA. PC1 and PC2 describe the global rotational motions of the 4H junction that transit structures from antiparallel to parallel forms and from perpendicular to parallel conformations, respectively; PC3 and PC4 capture the local bending and stretching motions of stacked helices, respectively.

that PC1 characterizes the transition of three different conformations (parallel, perpendicular, and antiparallel) achieved by a rotation of one stacked conformer against the other. PC2 describes a rotational motion similar to the PC1, but characterizing only the transition between perpendicular and parallel states. PC3 and PC4 capture local motions such as bending, stretching, and twisting within the stacked helices.

5.3.3 Analysis of conformational changes using various geometric measures

The leading dynamic motion of the 4H junction captured by PCA involves a key rotational motion of helical axes. We quantify this motion by following the rotational angle by measuring the pseudo-dihedral angle describing the relative orientation of residues—G_{137}, G_{229}, U_{226}, and U_{140}—considering two base pairs at the inter-helical interface (Figure 5.5A), as well as inter-helical distances using a pair of these residues during the simulation time (Figure 5.6A)

Assessment of conformational changes by a pseudo-dihedral angle

We measure the pseudo-dihedral angle θ, defined by the phosphate backbone atoms of G_{137}(P)-G_{229}(P)-U_{226}(P)-U_{140}(P) (Figure 5.5A), to illustrate the conformational change. The three systems (Figure 5.5B-D) exhibit similar θ distribution for the transition from perpendicular to parallel configurations. All the systems sample the perpendicular (θ: 7–16°) and parallel (θ: 55–60°) states. The system *Perp_34nt* (Figure 5.5B) fluctuates between 35° and −15° over the first 19 ns, sampling a few antiparallel configurations followed by a rapid angle change from −15° to −56° within ∼6 ns (from 19 ns to 25 ns) and arriving at a parallel configuration (25 ns). The system samples the perpendicular intermediate (27 ns) again and remains at the parallel state over the next 10 ns. *Perp_45nt* (Figure 5.5C) fluctuates with θ between 12° and −27° over the first 15 ns, fluctuating within a perpendicular state. Within the next 3 ns (from 15 ns to 18 ns), the system transitions from perpendicular to parallel states (18 ns) with the minimum θ of −56° and remains there over the next 18 ns. *Anti_116nt_2* (Figure 5.5D) gradually decreases during the antiparallel state

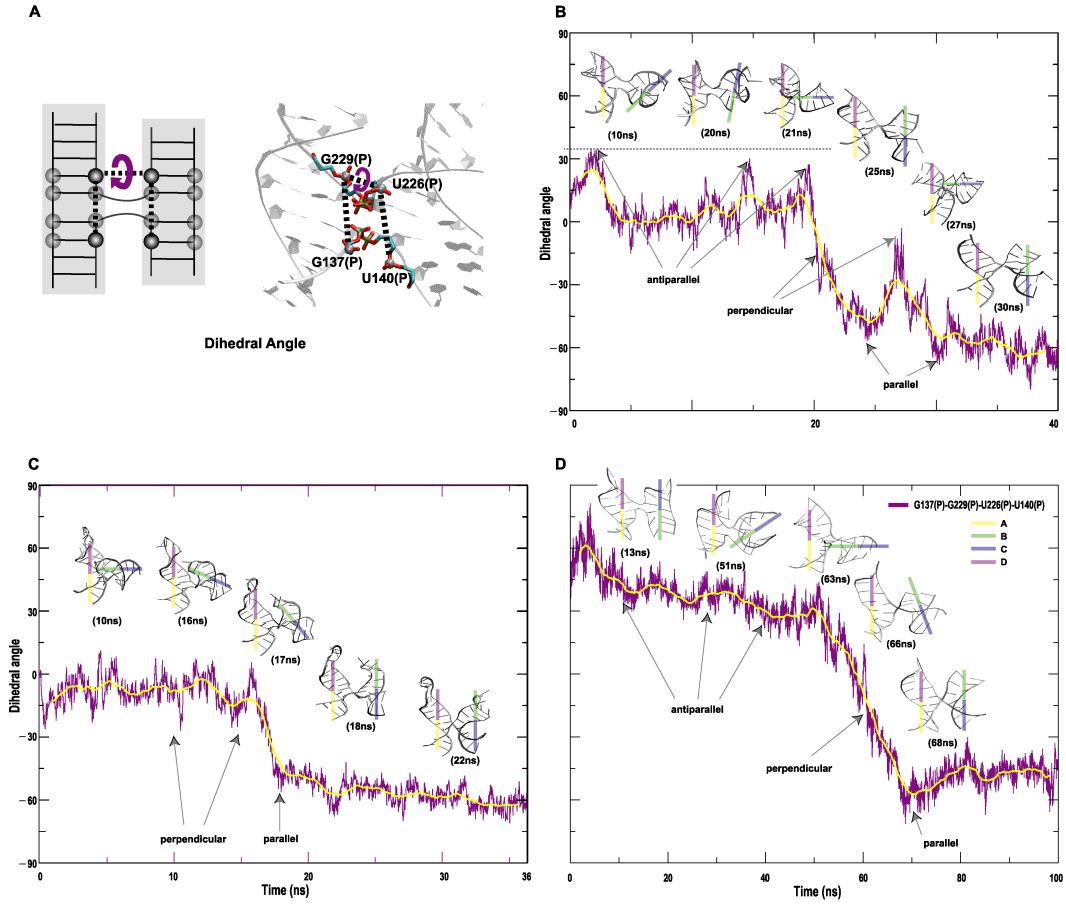


Figure 5.5: Conformational changes of the 4H junction described by a pseudo-dihedral angle between coaxially stacked helices using phosphate backbone atoms of the four residues G_{137}, U_{140}, U_{226}, and G_{229} near the center of 4-way junction. **(A)** Representative definition of a pseudo-dihedral angle *theta* based on the four residues (G_{137}, U_{140}, U_{226}, and G_{229} illustrated as dark balls) near the point of strand exchange involving the inter-helical orientation in 2D structure (left) and the angle *theta* in the 3D structure (right). **(B)** The pseudo-dihedral angle sampled at every 20 ps over the 40 ns (*Perp_34nt*), 36 ns (*Perp_45nt*), and 100 ns (*Anti_116nt_2*) time course with some of the representative structures.

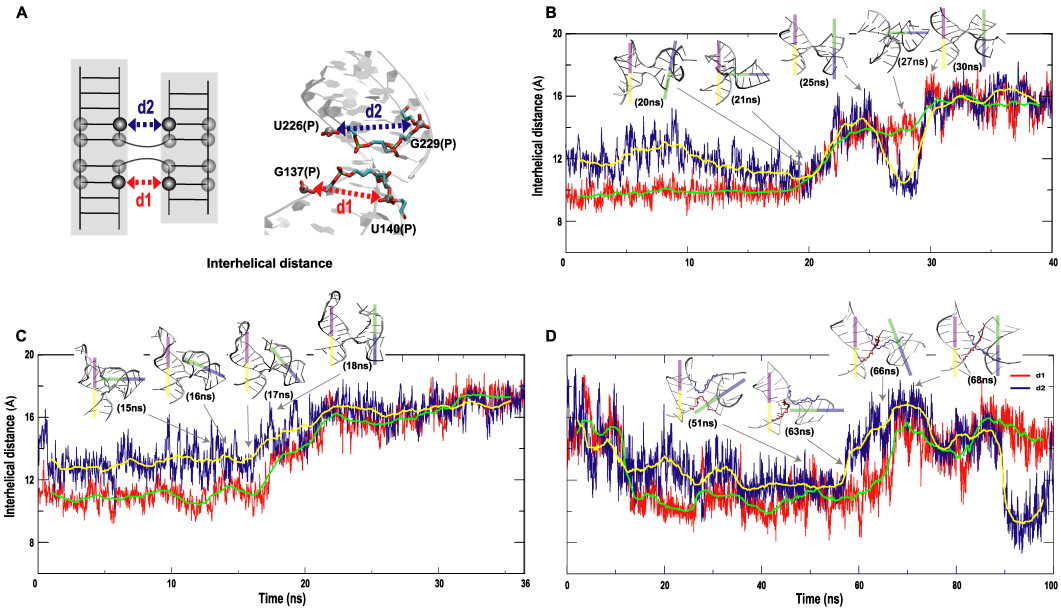


Figure 5.6: Conformational changes of the 4H junction described by inter-helical distance between coaxially stacked helices. **(A)** Representative definition of the inter-helical distances based on the four residues (G_{137}, U_{140}, U_{226}, and G_{229} illustrated by dark balls) near the point of strand exchange in 2D (left) and in 3D (right). **(B)** Inter-helical distances sampled at every 20 ps over the 40 ns (*Perp_34nt*), 36 ns (*Perp_45nt*), and 100 ns (*Anti_116nt_2*) time course with some representative structures around the rapid transition.

from the maximum value of 81° to 30° over the first 50 ns where the average angle is 39.0 ± 7.5°. Then, θ decreases rapidly to −71° over the next 18 ns (from 50 ns to 68 ns), exhibiting the conformational change from antiparallel to parallel configurations via a perpendicular intermediate; specifically, the system arrives at the perpendicular state (63 ns), and transitions from perpendicular to parallel states at 68 ns in ∼5 ns. During the parallel conformation, the average θ is −48.6 ± 4.7°. Overall, the θ distribution shows three distinctive regions where two dominant states, parallel and antiparallel, are bridged by the perpendicular intermediate of the 4H junctions.

Assessment of conformational changes by inter-helical distances

In Figure 5.6 we measure two inter-helical distances d1 and d2 defined by the backbone atoms of G_{137}(P)-U_{140}(P) and U_{226}(P)-G_{229}(P), respectively. Similar to the overall curve of pseudo-dihedral angle, d1 and d2 of the three systems exhibit similar distances for the transition from perpendicular (10–13Å) to parallel (14–16Å) configurations. *Perp_34nt* (Figure 5.6B) shows that d2 converges to ∼10Å while d1 is stable at ∼10Å (19 ns). Both d1 and d2 increase to ∼16Å over the next 6 ns (19 ns to 25 ns). D2 only drops to ∼10Å when the system samples again the perpendicular state at ∼28 ns. *Perp_45nt* (Figure 5.6C) shows some fluctuations of d1 and d2 around 10Å and 13Å (15 ns), respectively. Both d1 and d2 increase to ∼14Å over the next 3 ns (from 15 ns to 18 ns) and transit from perpendicular to parallel configurations. *Anti_116nt_2* (Figure 5.6D) shows that both d1 and d2 decrease from ∼16–17Å to ∼11Å over the first 50 ns, within the antiparallel configuration. Then, both distances increase abruptly, reaching the maximum value of 16Å over the next 18 ns. Interestingly, the d2 distance rapidly decreases at ∼90 ns where we observe a tertiary interaction

between B and D. Overall, the two inter-helical distances behave similarly, but with different degrees of fluctuations.

Correlation between a pseudo-dihedral angle and inter-helical distances

The above pseudo-dihedral angle and inter-helical distances (Figures 5.5 and 5.6) describe the global and local motions with respect to the different conformational states. To analyze these parameters′ contribution to the conformational changes, we examine the correlation between the dihedral angle and inter-helical distances in Figure 5.7.

During the perpendicular state for ∼19 ns, *Perp_34nt* (Figure 5.7A) also samples a few antiparallel states. While both d1 and d2 converge to ∼10Å, θ fluctuates between 35° and −15°. When θ decreases from −15° to −56° and both d1 and d2 increase to ∼16Å, a transition occurs from perpendicular to parallel configurations within ∼6 ns. *Perp_45nt* (Figure 5.7B) fluctuates during the perpendicular state (15 ns) with θ between 12° and −27°, and d1 and d2 each around 10Å and 13Å. With the decrease of θ from −27° to −56°, a transition occurs from perpendicular to parallel configurations within ∼6 ns while both d1 and d2 increase to ∼16Å. During the antiparallel conformation which lasts for about 50 ns, both d1 and d2 of *Anti_116nt_2* (Figure 5.7C) arrive at a local minima of ∼11Åwhile θ gradually decreases to ∼30°. From 30° to −30°, a rapid transition occurs from antiparallel to perpendicular states over 13 ns while both d1 and d2 gradually increase. When θ is between −30° to −71°, the system achieves the parallel conformation with a maximum distance of ∼16–17Å. During the parallel state, d2 fluctuates more than d1 within the range of 7–18Å and 10–17Å, respectively, showing a scattered distribution. Interestingly,

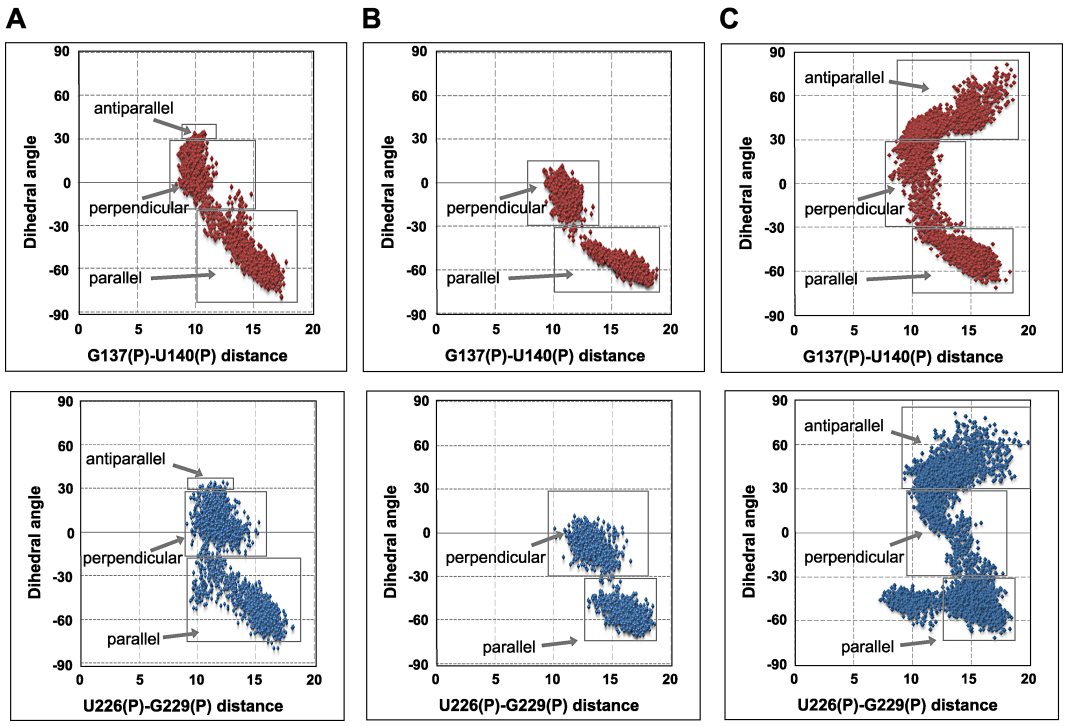


Figure 5.7: Overall distribution of the correlation between pseudo-dihedral angle and inter-helical distances for the conformational change from perpendicular (*Perp_34nt* and *Perp_45nt*) and antiparallel (*Anti_116nt_2*) to parallel. A relation between the dihedral angle and the distances U_{226}(P)-G_{229}(P) and G_{137}(P)-U_{140}(P) for *Perp_34nt* **(A)**, *Perp_45nt* **(B)**, and *Anti_116nt_2* **(C)** is highlighted with solid gray boxes for these events (parallel, perpendicular, and antiparallel configurations).

the overall trend of the correlated distribution for θ versus d1 and d2 are similar. The difference at the parallel state of θ versus d2 is due to the tertiary contact between B and D, initiated at ∼90 ns.

5.3.4 Flexibility of terminal base pairs at the core of the 4H junction

All base-pairs in the 4H junction are involved with base pairing and stacking interactions, between complementary strands and between adjacent bases, respectively. These interactions contribute significantly to RNA structure stability by forming coaxial helical stacks, for example. Previously, a disruption of base stacking in the connecting (or inter-helical) residues at the center of a RNA-DNA hybrid 4H junction was noted as responsible for achieving a different conformational state [201]. Thus, we next analyze these base pairing and stacking interactions at the core of our 4H junction to assess their involvement with the conformational change from antiparallel to parallel configurations (*Anti_116nt_2*).

First, we measure distances of the base pairs at the helix ends—C_{138}-G_{245}, U_{139}-G_{150}, C_{151}-G_{227}, and U_{228}-G_{244}—around the junction center (or branch point) by considering nitrogen and oxygen atoms (Figure D.1, Appendix D). The four terminal base pairs consist of a pair of G-C and G-U bases. The latter (G-U) is thermodynamically less stable than the former (G-C). Figure D.1 in Appendix D shows that G-C base pairs in A and C remain highly stable with the average distance of 2.82 ± 0.07Å whereas G-U wooble base pairs in B and D exhibit small fluctuations with average distances of 2.89 ± 0.12Å and 2.85 ± 0.09Å, respectively. In particular, fluctuations of the G-U base pair in D

indicate that the hydrogen bonds are temporarily disrupted, affecting the base pairing and overall flexibility. All four base pairs maintain well the hydrogen bonds over the time course of the simulation.

Second, we measure base stacking interactions for the two non-consecutive bases C_{138}-U_{228} and U_{139}-G_{227} between AD and BC, respectively (Figure D.2, Appendix D). To consider base stacking interactions, we use geometric criteria of a distance ($\leq$ 5.5Å) and angle ($\leq 30^{\circ}$) between these bases. Overall, the base stacking interactions are well maintained, with only temporarily disruption during the antiparallel state. In particular, the base stacking interactions remain stable during the fast transition, from perpendicular to parallel configurations.

5.4 Discussion

RNA junctions are the largest secondary structural element or motif found in diverse RNA molecules. They are structurally and functionally important, playing central roles in RNA folding. The 4H junction we examine here is a simplest type of a 4-way junction that contains fully base-paired helices often found in self-assembling molecules such as the hairpin ribozyme [187] and viral mRNAs [188]. The junction′s overall shape contributes significantly to biological functions (e.g., splicing, catalyzing, and translation initiation).

The 4H junctions contain pairwise coaxial stacking of helices that adopt well-defined helical arrangements which direct the system to a compact fold [89]. Thus, the folding pathway of such junctions has been under intense study. Gel electrophoresis and (single-molecule) fluorescence resonance energy transfer have suggested two possible pathways between parallel and antiparallel configurations: (1) a transition via a helical rearrangement by disrupting coaxial helical

stacking; (2) a transition driven by a rotation at the center of the junction which maintains the coaxial stacks.

Using MD simulations, we have explored structural properties of the 4H junction, taken from FMDV IRES domain 3. This 4H junction brings together the distant RNA-RNA segments that play crucial roles in the structural stability and organization of entire domain 3, which in turn affects IRES activity. Thus, assembly of the 4H RNA junction is a prerequisite for establishing the folded 3D structure of domain 3 and thus enabling the initiation mechanism of translation in FMDV IRES. Our studies suggest that both parallel and antiparallel configurations of the 4H junction are sampled, with a virtually barrier-free transition between them as deduced experimentally [48]. The transition between parallel and antiparallel conformations occurs via a perpendicular intermediate that maintains the coaxial stacks (Figure 5.1). Because the GNRA motif interacts with the helix D in the 4H junction, a transition that offers various stable configurations via pairwise coaxial stacking of helices is beneficial to initiate the long-range RNA-RNA interactions. Still, we cannot exclude other pathways for the transition.

Analysis of the principal motions indicates that both global and local motions contribute to the above conformational exchanges. The first two largest PCs capture 85% of the dominant motion and characterize the transition between parallel and antiparallel via perpendicular states involving a rotation. The third and fourth largest PCs, in total of 6%, capture local motions within stacked helices (e.g., bending, stretching, and twisting). These motions are described by inter-helical residues connecting the two coaxial stacking helices (Figure 5.5A and 5.6A). Specifically, analysis of inter-helical distances and pseudo-dihedral angles help organize the conformations into antiparallel, parallel, and perpen-

dicular states and reveal the transience of the transition state. The polymorphic nature of the 4H junction without added cofactors is well appreciated in the literature [192, 48] and thought to be advantageous for IRES′s versatile functions. Thus, a modular structural platform that is easily adjusted by the binding of the molecular co-factors suits this large RNA for its complex activity.

The alternative suggested interconversion via a helical rearrangement, including a cruciform intermediate triggered by reduced cation binding at the junction domain, was not observed in our equilibrium trajectories within the 30?100 ns time scale, neutralized by Na^+ ions; the pairwise coaxial stacks of helices remain intact due to effective screening of the strong Coulomb repulsion between RNA junction domains. At present, state-of-the-art nucleic acids force fields for MD simulations describe well monovalent ions and solute-solvent interactions, but not divalent ions [98].

The perpendicular intermediate, in particular, may be advantageous for directing further long-range RNA-RNA interactions via the GNRA or RAAA motifs because it provides a rapid transition that can potentially accelerate assembly of interactions with a possible binding site in the 4H junction. It is possible that while one of the coaxially stacked helices is occupied in tertiary interactions in the perpendicular orientation, the other stacked conformer continues to explore conformational space to find its tertiary interaction partner for further stabilization. Ultimately, we envision that inter and intra-molecular RNA-RNA interactions, possibly involving the hairpin loops in helices B and D, are required to anchor the 4H junction in either the parallel or antiparallel conformation.

Chapter 6

Conclusions

Understanding the nature of complex RNA junction structures is important in RNA structure modeling and prediction since they are the major determinant in the organization of large RNA molecules. In this thesis, we have studied various structural aspects of RNA junctions using non-redundant high-resolution dataset and develop applications to predict 3D RNA structures including regulatory regions in viral RNAs.

We have reported in Chapter 2 that diverse RNA junctions are observed in high-resolution crystal structures containing up to 10 helical arms. Our statistical analysis on structural elements for these RNA junctions shows recurrent tertiary motifs such as coaxial stacking of helices and A-minor interactions, and a new motif for perpendicular helical arrangements. Notably, we observe the folding similarity among different degree of junctions; for example, similar helical arrangements of 3 and 4-way junctions are found in higher-order junctions. This analysis suggests that higher-order junctions can be decomposed into smaller sub-junctions. Ultimately, we hope that a better understanding of the higher-order junction decomposition and recurrent tertiary motifs can help

predict architecture of large RNA 3D structures and the biological functions.

A current major challenge in the field of RNA 3D structure prediction is the reliable prediction accuracy as the RNA system size and structural complexity grow. In Chapter 3, we have demonstrated a novel computational approach to describe helical topologies of RNA 3 and 4-way junctions as tree graphs, called RNAJAG (RNA Junction-As-Graph), grounded in the RNA junction analysis above. RNAJAG reproduces reliable helical junction configurations in 3 and 4-way junctions for a large set of 200 RNA junctions. The remaining challenges for RNAJAG are to deal with higher-order RNA junctions, and to build eventually the detailed atomic models. As described in Chapter 2, higher-order junctions can be partitioned into sub-junctions. Thus, further development of RNAJAG is required to partition, predict, and assemble these junctions. In addition to the threading/build-up procedure noted in Chapter 3, it is also feasible to build all-atom models by mapping predicted graphs to atomic models in space using the jsecondary structure and spatial helical organization information of the junctions.

With the advances in RNA junction analysis, prediction, and modeling, we have proposed in Chapter 4 the candidate RNA junction structures in regulatory regions, called internal ribosome entry site (IRES), of the foot-and-mouth-disease virus (FMDV). Based on all available experimental data, we suggest a plausible theoretical tertiary structure of the apical region in FMDV IRES domain 3 by utilizing various computational approaches—topology modeling, atomic 3D structure modeling, and MD simulations. Together with the dynamics study in Chapter 4, our work provides insights into the potential role of the long-range interactions for structural stability in the central domain of FMDV IRES and thus may offer a further experimental investigation of the structure,

mechanism, and function of the viral RNA. Ultimately, we hope that our findings help the development of antiviral drugs that inhibit IRES activity and thus virus multiplication.

Important biological functions of RNA junctions are often closely linked to the dynamic nature and conformational flexibility. In Chapter 5, we have investigated the structural properties of the 4H junction found in FMDV IRES domain 3, which contains no nucleotides between helices within the junction domain, by molecular dynamics (MD) simulations. Our results suggest that a transition between parallel and antiparallel conformations occurs via a rotation at the axes of coaxially stacked helices. This interconversion exhibits various inter-helical angles including a transient perpendicular intermediate. Our findings suggest the possible conformational pathway of the 4H RNA junction. In particular, the perpendicular intermediate with a rapid transition can potentially accelerate assembly of interactions with a possible binding site in the 4H junction to direct overall RNA folding. We hope that this conformational pathway and detailed mechanism of the conformational change open new ways to think about RNA versatility and to design a novel self-assembling RNA system such as A-minor 4H RNA junctions.

Appendix A

Supplementary Information for Chapter 2

Table A.1: List of RNA 3D structures containing 106 3-way junctions. The name describes the PDB code and the number of the first residue of helix H1 in the junction. The nomenclature is based on [90] and the helices are numbered according to the scheme in Leffers *et al.* [80].

Name	RNA Type	Coaxial stacks	Helical alignments	Family	Nomenclature	Domain	Helix numbers
1NKW_31	23S rRNA *D. Radiodurans*	H1H2		A	HS2HS14HS4	I	H3-4-23
1S72_28	23S rRNA *H. Marismortui*	H1H2		A	HS2HS15HS4	I	H3-4-23
2AW4_31	23S rRNA *E. Coli*	H1H2		A	HS2HS15HS4	I	H3-4-23
2J01_31	23S rRNA *T. Thermophilus*	H1H2		A	HS2HS15HS4	I	H3-4-23
1NKW_54	23S rRNA *D. Radiodurans*	H1H2		A	2HS2HS4	I	H5-6-7
1S72_51	23S rRNA *H. Marismortui*	H1H2		A	2HS2HS3	I	H5-6-7
2AW4_55	23S rRNA *E. Coli*	H1H2		A	2HS2HS3	I	H5-6-7
2J01_55	23S rRNA *T. Thermophilus*	H1H2		A	2HS2HS3	I	H5-6-7
1NKW_1310	23S rRNA *D. Radiodurans*	H1H3		A	HS4HS1HS2	III	H48-60-?
1S72_1403	23S rRNA *H. Marismortui*	H1H3		A	HS4HS2HS2	III	H48-60-?
2AW4_1297	23S rRNA *E. Coli*	H1H3		A	HS4HS1HS2	III	H48-60-?
2J01_1297	23S rRNA *T. Thermophilus*	H1H3		A	HS4HS1HS2	III	H48-60-?
1NKW_1318	23S rRNA *D. Radiodurans*	H1H2		A	HS2HS2HS1	III	H49-59.1-?
1S72_1411	23S rRNA *H. Marismortui*	H1H2		A	HS2HS2HS1	III	H49-59.1-?
2AW4_1305	23S rRNA *E. Coli*	H1H2		A	HS2HS2HS1	III	H49-59.1-?
2J01_1305	23S rRNA *T. Thermophilus*	H1H2		A	HS2HS2HS1	III	H49-59.1-?
1NKW_2072	23S rRNA *D. Radiodurans*	H2H3		A	HS2HS1HS5	V	H75-76-79
1S72_2130	23S rRNA *H. Marismortui*	H2H3		A	HS2HS1HS5	V	H75-76-79
2AW4_2090	23S rRNA *E. Coli*	H2H3		A	HS1HS1HS4	V	H75-76-79
2J01_2090	23S rRNA *T. Thermophilus*	H2H3		A	HS1HS1HS4	V	H75-76-79
1NKW_2788	23S rRNA *D. Radiodurans*	H1H2		A	2HS4HS3	VI	H99-100-101
1S72_2830	23S rRNA *H. Marismortui*	H1H2		A	2HS4HS3	VI	H99-100-101
2AW4_2811	23S rRNA *E. Coli*	H1H2		A	HS2HS4HS3	VI	H99-100-101
2J01_2813	23S rRNA *T. Thermophilus*	H1H2		A	2HS4HS3	VI	H99-100-101
2J00_585	16S rRNA *T. Thermophilus*	H2H3		A	HS1HS1HS3	C	H20-21-22
2AVY_585	16S rRNA *E. Coli*	H2H3		A	HS1HS1HS3	C	H20-21-22
2J00_671	16S rRNA *T. Thermophilus*	H1H2		A	2HS1HS1	C	H22-23-23a
2AVY_671	16S rRNA *E. Coli*	H1H2		A	HS4HS11HS1	C	H22-23-23a
2J00_825	16S rRNA *T. Thermophilus*	H2H3		A	HS2HS1HS5	C	H25-26-26a
2AVY_825	16S rRNA *E. Coli*	H2H3		A	HS2HS1HS5	C	H25-26-26a
2J00_1059	16S rRNA *T. Thermophilus*			A	HS1HS1HS5	3'm	H34-35-38
2AVY_1059	16S rRNA *E. Coli*			A	HS1HS1HS9	3'm	H34-35-38
2GDI_13	Riboswitch TPP *E. Coli*	H1H2		A	2HS2HS2		P1-2-4
2HOJ_13	Riboswitch TPP *E. Coli*	H1H2		A	2HS2HS2		P1-2-4
2CKY_4	Riboswitch TPP *A. Thaliana*	H1H2		A	HS1HS1HS4		P1-2-4
2QBZ_53	Riboswitch M-box *B. Subtilis*	H1H2		A	HS1HS1HS4		P3-4a-5
1U6B_45	GI Intron *Azoarcus*	H2H3		A	HS22HS13		P3-4-6
2A64_12	RNase P type *B. Stearothermophilus*	H2H3		A	HS12HS17		P1-2-19
1NKW_709	23S rRNA *D. Radiodurans*	H2H3		B	HS2HS2HS4	II	H33-34-35
1S72_787	23S rRNA *H. Marismortui*	H2H3		B	HS2HS4HS4	II	H33-34-35
2AW4_696	23S rRNA *E. Coli*	H2H3		B	HS2HS1HS4	II	H33-34-35
2J01_696	23S rRNA *T. Thermophilus*	H2H3		B	HS2HS1HS4	II	H33-34-35
1NKW_1322	23S rRNA *D. Radiodurans*		H2H3	B	HS1HS6HS2	III	H50-51-?
1S72_1415	23S rRNA *H. Marismortui*		H2H3	B	HS12HS2	III	H50-51-?
2AW4_1309	23S rRNA *E. Coli*		H2H3	B	HS1HS6HS2	III	H50-51-?
2J01_1309	23S rRNA *T. Thermophilus*		H2H3	B	HS1HS6HS2	III	H50-51-?
1S72_2328	23S rRNA *H. Marismortui*		H2H3	B	HS1HS1HS4	V	H83-84-85
2AW4_2294	23S rRNA *E. Coli*		H2H3	B	HS1HS1HS4	V	H83-84-85
2J01_2294	23S rRNA *T. Thermophilus*		H2H3	B	HS1HS1HS4	V	H83-84-85
2J00_935	16S rRNA *T. Thermophilus*			B	HS2HS1HS6	3'm	H28-29-43
2AVY_935	16S rRNA *E. Coli*			B	HS2HS1HS6	3'm	H28-29-43
2J00_989	16S rRNA *T. Thermophilus*	H2H3		B	HS6HS2HS4	3'm	H32-33-34
2AVY_989	16S rRNA *E. Coli*	H2H3		B	HS6HS2HS4	3'm	H32-33-34
2J00_1002	16S rRNA *T. Thermophilus*	H1H3		B	HS2HS2HS4	3'm	H33-33a-33b
2AVY_1002	16S rRNA *E. Coli*	H1H3		B	HS2HS1HS2	3'm	H33-33a-33b

2OIU_6	L1 ribozyme *Synthetic*	H1H2		B	2HS2HS3		A-B-C
1NKW_23	23S rRNA *D. Radiodurans*	H1H2		C	HS6HS9HS19	I	H2-3-24
1S72_20	23S rRNA *H. Marismortui*	H1H2		C	HS7HS10HS20	I	H2-3-24
2AW4_24	23S rRNA *E. Coli*	H1H2		C	HS5HS9HS18	I	H2-3-24
2J01_23	23S rRNA *T. Thermophilus*	H1H2		C	HS6HS9HS19	I	H2-3-24
1NKW_307	23S rRNA *D. Radiodurans*	H1H3		C	HS3HS8HS3	I	H18-19-20
1S72_301	23S rRNA *H. Marismortui*	H1H3		C	HS6HS9HS4	I	H18-19-20
2AW4_296	23S rRNA *E. Coli*	H1H3		C	HS3HS8HS3	I	H18-19-20
2J01_296	23S rRNA *T. Thermophilus*	H1H3		C	HS3HS8HS3	I	H18-19-20
1NKW_695	23S rRNA *D. Radiodurans*	H1H2		C	HS5HS3HS6	II	H32-33-35a
1S72_773	23S rRNA *H. Marismortui*	H1H2		C	HS5HS3HS6	II	H32-33-35.1
2AW4_682	23S rRNA *E. Coli*	H1H2		C	HS5HS3HS6	II	H32-33-35a
2J01_682	23S rRNA *T. Thermophilus*	H1H2		C	HS5HS3HS6	II	H32-33-35a
1NKW_1065	23S rRNA *D. Radiodurans*	H1H3		C	HS3HS9HS3	II	H42-43-44
1S72_1158	23S rRNA *H. Marismortui*	H1H3		C	HS3HS8HS2	II	H42-43-44
2AW4_1054	23S rRNA *E. Coli*	H1H3		C	HS3HS8HS2	II	H42-43-44
1S72_1550	23S rRNA *H. Marismortui*	H1H2		C	2HS4HS11	III	57-58-59
2J01_1443	23S rRNA *T. Thermophilus*	H1H2		C	HS12HS13	III	57-58-59
1NKW_2495	23S rRNA *D. Radiodurans*	H1H2		C	HS2HS3HS5	V	H90-91-92
1S72_2551	23S rRNA *D. Radiodurans*	H1H2		C	HS2HS2HS5	V	H90-91-92
2AW4_2516	23S rRNA *E. Coli*	H1H2		C	HS2HS3HS5	V	H90-91-92
2J01_2516	23S rRNA *T. Thermophilus*	H1H2		C	HS2HS3HS5	V	H90-91-92
1NKW_11	5S rRNA *D. Radiodurans*	H2H3		C	HS3HS3HS3		H1-2-4
1S72_8	5S rRNA *H. Marismortui*	H2H3		C	HS4HS3HS3		H1-2-4
2AW4_9	5S rRNA *E. Coli*	H2H3		C	HS3HS3HS3		H1-2-4
2J01_10	5S rRNA *T. Thermophilus*	H2H3		C	HS42HS1		H1-2-4
1UN6_8	5S rRNA *X. Laevis*	H2H3		C	HS42H		H1-2-4
2J00_45	16S rRNA *T. Thermophilus*	H1H3		C	HS5HS9HS2	5'	H4-5-15
2AVY_45	16S rRNA *E. Coli*	H1H3		C	HS5HS9HS2	5'	H4-5-15
2J00_954	16S rRNA *T. Thermophilus*			C	HS6HS10HS3	3'm	H30-31-32
2AVY_954	16S rRNA *E. Coli*			C	HS6HS10HS3	3'm	H30-31-32
2J00_1072	16S rRNA *T. Thermophilus*	H1H2		C	2HS3HS3	3'm	H35-36-37
2AVY_1072	16S rRNA *E. Coli*	H1H2		C	2HS3HS3	3'm	H35-36-37
2J00_1115	16S rRNA *T. Thermophilus*	H1H2		C	HS5HS5HS8	3'm	H38-39-40
2AVY_1115	16S rRNA *E. Coli*	H1H2		C	HS5HS5HS8	3'm	H38-39-40
1NYI_14	Hammerhead *Synthetic*	H2H3		C	HS7HS4HS2		I-II-III
1E8O_102	ALU domain SRP *H. Sapiens*	H1H3		C	2HS4H		H1.1-1.2-2
1L9A_127	ALU domain SRP *Synthetic*	H1H3		C	2HS6HS2		H6-7-8
1LNG_144	SRP 19-7S.S *M. Jannaschii*	H1H3		C	HS2HS4HS1		H5-6-8
2CZJ_6	tmRNA *T. Thermophilus*	H1H3		C	HS12HS6HS3		P1-2a-10
1U8D_20	Riboswitch G *B. Subtilis*	H1H3		C	HS5HS10HS2		P1-2-3
2B57_20	Riboswitch G *B. Subtilis*	H1H3		C	HS5HS10HS2		P1-2-3
2EES_20	Riboswitch G *B. Subtilis*	H1H3		C	HS5HS10HS2		P1-2-3
1Y26_20	Riboswitch A *V. Vulnificus*	H1H3		C	HS5HS8HS2		P1-2-3
2HO7_5	Riboswitch glmS *T. Tengcongensis*	H1H2		C	2HS2HS6		P1-2.1-2.2
2NZ4_5	Riboswitch glmS *B. Anthracis*	H1H2		C	2HS2HS6		P1-2.1-2.2
1Y0Q_195	GI Intron *Twort*			C	HS2HS3HS9		P9-9.0-9.1
1X8W_137	GI Intron *Tetrahymena*	H1H2		C	HS2HS2HS4		P5a-5b-5c
2A64_61	RNase P type B *B. Stearothermophilus*	H1H2		C	2HS5H		P5-5.1-7
2A64_139	RNase P type B *B. Stearothermophilus*	H1H3		C	HS5HS5H		P7-10.1-11
1NBS_132	RNase P type B *B. Subtilis*	H1H3		C	HS5HS5H		P7-8-9-10

Table A.2: List of RNA 3D structures containing 62 4-way junctions. The name describes the PDB code and the number of the first residue of helix H_1 in the junction. The nomenclature is based on [90] and the helices are numbered according to the scheme in Leffers *et al.* [80].

Name	RNA Type	Coaxial stacks	Helical alignments	Family	Nomenclature	Domain	Helix Numbers
1U9S_78	RNase P type A *T. thermophilus*	H1H4, H2H3		H	$2HS_1HS_2H$		P7-8-9-10
2A2E_70	RNase P type A *T. Maritima*	H1H4, H2H3		H	HS_13HS_1		P7-8-9-10
1NBS_89	RNase P type B *B. Subtilis*	H1H4, H2H3		H	$2HS_1HS_2H$		P7-8-9-10
2A64_90	RNase P type B *B. Stearothermophilus*	H1H4, H2H3		H	$2HS_1HS_2H$		P7-8-9-10
1M5O_13	Hairpin Ribozyme *S. Tobacco ringspot virus*	H1H4, H2H3		H	4H		A-B-C-D
1S72_1827	23S rRNA *H. Marismortui*	H1H4, H2H3		H	$HS_3HS_3HS_3HS_4$	IV	H64-65-66-67
2AW4_1771	23S rRNA *E. Coli*	H1H4, H2H3		H	$HS_3HS_3HS_2HS_3$	IV	H64-65-66-67
2J01_1771	23S rRNA *T. Thermophilus*	H1H4, H2H3		H	$HS_3HS_3HS_2HS_3$	IV	H64-65-66-67
1KH6_4	IRES *Hepatitis C Virus*	H1H4, H2H3		cH	HS_22HS_1H		III-IIIa-IIIb-IIIc
2AVY_141	16S rRNA *E. Coli*	H1H4, H2H3		cH	$HS_3HS_7HS_4HS_1$	5'	H7-8-9-10
2J00_141	16S rRNA *T. Thermophilus*	H1H4, H2H3		cH	$HS_1HS_4HS_3HS_1$	5'	H7-8-9-10
1NKW_2621	23S rRNA *D. Radiodurans*	H1H4, H2H3		cH	$HS_2HS_1HS_4HS_2$	VI	H94-95-96-97
1S72_2678	23S rRNA *H. Marismortui*	H1H4, H2H3		cH	$HS_22HS_2HS_1$	VI	H94-95-96-97
2AW4_2642	23S rRNA *E. Coli*	H1H4, H2H3		cH	$HS_2HS_1HS_3HS_1$	VI	H94-95-96-97
2J01_2642	23S rRNA *T. Thermophilus*	H1H4, H2H3		cH	$HS_2HS_1HS_3HS_1$	VI	H94-95-96-97
3F2Q_7	Riboswitch (FMN) *F. Nucleatum*	H1H4, H2H3		cH	$HS_6HS_3HS_1HS_7$		P1-P2-X-P6
3F2Q_31	Riboswitch (FMN) *F. Nucleatum*	H1H4, H2H3		cH	$2HS_2HS_3HS_2$		X-P3-P4-P5
1NKW_1457	23S rRNA *D. Radiodurans*	H1H2, H3H4		cH	$HS_1HS_3HS_6HS_4$	III	H56-57-58-59
2AW4_1443	23S rRNA *E. Coli*	H1H2, H3H4		cH	$3HS_3HS_4$	III	H56-57-58-59
2AVY_568	16S rRNA *E. Coli*	H1H4, H2H3		cL	$HS_7HS_4HS_{10}HS_1$	C	H19-20-24-25
2J00_568	16S rRNA *T. Thermophilus*	H1H4, H2H3		cL	$HS_7HS_4HS_{10}HS_1$	C	H19-20-24-25
1NKW_1282	23S rRNA *D. Radiodurans*	H1H4, H2H3		cL	HS_52HS_2H	III	H47A-47-48-61
1S72_1373	23S rRNA *H. Marismortui*	H1H4, H2H3		cL	HS_72HS_3H	III	H47A-47-48-61
2AW4_1269	23S rRNA *E. Coli*	H1H4, H2H3		cL	HS_52HS_2H	III	H47A-47-48-61
2J01_1269	23S rRNA *T. Thermophilus*	H1H4, H2H3		cL	HS_52HS_2H	III	H49A-49-50-51
1EFW_6	Asp-tRNA *T. Thermophilus*	H1H4, H2H3		cL	$HS_2HS_1HS_5H$		H1-2-3-4
1EHZ_6	Phe-tRNA *Yeast*	H1H4, H2H3		cL	$HS_2HS_1HS_5H$		H1-2-3-4
1N78_506	Glu-tRNA *T. Thermophilus*	H1H4, H2H3		cL	$HS_2HS_1HS_4H$		H1-2-3-4
1QRS_6	Gln-tRNA *E. Coli*	H1H4, H2H3		cL	$HS_2HS_1HS_5H$		H1-2-3-4
1U0B_6	Cys-tRNA *E. Coli*	H1H4, H2H3		cL	$HS_2HS_1HS_4H$		H1-2-3-4
2GIS_7	Riboswitch (SAM I) *Synthetic*	H1H4, H2H3		cL	$HS_6HS_1HS_8HS_3$		P1-2A-3-4

2AVY_114	16S rRNA *E. Coli*	H1H4		cK	$HS_6{}^2HS_2$	5'	H7-11-12-13A
2J00_114	16S rRNA *T. Thermophilus*	H1H4		cK	$HS_6{}^2HS_2$	5'	H7-11-12-13A
1NKW_2263	23S rRNA *D. Radiodurans*	H3H4		cK	$HS_4HS_4HS_1{}^1HS_1$	V	H82-83-86-87
1S72_2318	23S rRNA *H. Marismortui*	H3H4		cK	$HS_3HS_3HS_1{}^1HS_1$	V	H82-83-86-87
2AW4_2284	23S rRNA *E. Coli*	H3H4		cK	$HS_3HS_3HS_1{}^1HS_1$	V	H82-83-86-87
2J01_2284	23S rRNA *T. Thermophilus*	H3H4		cK	$HS_3HS_3HS_1{}^1HS_1$	V	H83A-83-86-87
1NKW_1360	23S rRNA *D. Radiodurans*	H3H4		cK	$HS_1HS_3HS_4HS_2$	III	H51-52-53-54
1S72_1452	23S rRNA *H. Marismortui*	H3H4		cK	$HS_1HS_3HS_2HS_1$	III	H51-52-53-54
2AW4_1346	23S rRNA *E. Coli*	H3H4		cK	$HS_2HS_3HS_2HS_1$	III	H49A-49-50-51
2J01_1347	23S rRNA *T. Thermophilus*	H3H4		cK	$HS_1HS_3HS_2H$	III	H51-52-53-54
2AVY_18	16S rRNA *E. Coli*	H1H2		cK	$HS_7HS_{10}HS_1HS_3$	C	H2-3-19-27
2J00_18	16S rRNA *T. Thermophilus*	H1H2		cK	$HS_7HS_{10}HS_1HS_3$	C	H2-3-19-28
1U9S_118	RNase P type A *T. thermophilus*	H3H4	H1H2	π	$HS_{12}HS_7HS_1HS_4$		P11-12-13-14
2A2E_110	RNase P type A *T. Maritima*	H3H4	H1H2	π	$HS_{10}HS_7HS_2HS_4$		P11-12-13-14
1NKW_1682	23S rRNA *D. Radiodurans*		H1H4	cW	$HS_{13}{}^2HS_{12}HS_5$	IV	H61-62-63-64
1S72_1743	23S rRNA *H. Marismortui*		H1H4	cW	$HS_{13}{}^2HS_{12}HS_5$	IV	H61-62-63-64
2AW4_1665	23S rRNA *E. Coli*		H1H4	cW	$HS_{13}{}^2HS_{12}HS_5$	IV	H61-62-63-64
2J01_1665	23S rRNA *T. Thermophilus*		H1H4	cW	$HS_{13}{}^2HS_{12}HS_5$	IV	H61-62-63-64
1S72_42	23S rRNA *H. Marismortui*		H2H4	Ψ	$HS_6HS_4HS_3HS_1$	I	H4-5-8-10
1NKW_1824	23S rRNA *D. Radiodurans*		H2H4	Ψ	$HS_1HS_2HS_{22}HS_{10}$	IV	H64-65-66-67
1S72_1888	23S rRNA *H. Marismortui*		H2H4	Ψ	$HS_{12}HS_{20}HS_{10}$	IV	H67-68-69-71
2AW4_1832	23S rRNA *E. Coli*		H2H4	Ψ	$HS_{12}HS_{20}HS_{10}$	IV	H64-65-66-67
2J01_1832	23S rRNA *T. Thermophilus*		H2H4	Ψ	$HS_{12}HS_{20}HS_{10}$	IV	H64-65-66-67
1NKW_244	23S rRNA *D. Radiodurans*		H2H4	Ψ	$HS_5HS_6HS_6HS_2$	I	H14-16-21-22
2AW4_267	23S rRNA *E. Coli*		H2H4	Ψ	$HS_5HS_6HS_6HS_2$	I	H14-16-21-22
1NKW_608	23S rRNA *D. Radiodurans*			X	$HS_{10}HS_9HS_3HS_{11}$	I	H27-28-29-31
2AW4_600	23S rRNA *E. Coli*			X	$HS_2HS_3HS_2HS_5$	I	H27-28-29-31
2J01_600	23S rRNA *T. Thermophilus*			X	$HS_2HS_3HS_2HS_2$	I	H27-28-29-31
2IHX_166	*Sarcoma Virus*			cX	$HS_3HS_4{}^2HS_3$		A-B-C-O3
2AVY_942	16S rRNA *E. Coli*			cX	$HS_2HS_5HS_{11}HS_{10}$	3'M	H29-30-41-42
2J00_940	16S rRNA *T. Thermophilus*			cX	$HS_4HS_5HS_{11}HS_{12}$	3'M	H29-30-41-42

Table A.3: List of helix-helix interactions containing AGPM, ribo-base type I and II or both. The first column denotes interaction type (Int.) such as ribo-base interactions type I (RI) or type II (RII). The location describes the PDB file code and secondary structure location such as next to an internal loop (I. loop) or within a n-way junction (nWJ). Watson-Crick base pairs GC, CG, AU, UA and GU wobble are color-coded for easy identification.

Int.	AGPM & Ribo-base type I		Location	Int.	AGPM & Ribo-base type I		Location
AGPM	G549-U563	G17-C533	23S (1NKW)	AGPM	G2688-U2677	G1644-C1655	23S (1NKW)
RI	C550-G562	U18-A532	H2-H25 (10WJ)	RI	C2689-G2676	U1645-A1654	H60-H96 (3WJ-I. loop)
AGPM	G545-U611	C14-G529	23S (1S72)	AGPM	G2744-U2735	G1704-C1714	23S (1S72)
RI	C546-G610	C15-G528	H2-H25	RI	C2745-G2734	C1705-G1713	H60-H96 (3WJ-I. loop)
AGPM	G539-U554	G17-C523	23S (2AW4)	AGPM	G2709-U2698	G1628-C1638	23S (2AW4)
RI	C540-G553	U18-A522	H2-H25 (10WJ)	RI	C2710-G2697	U1629-A1637	H60-H96 (3WJ-I. loop)
AGPM	G539-U554	G17-C523	23S (2J01)	AGPM	G2709-U2698	G1628-C1638	23S (2J01)
RI	C540-G553	C18-G522	H2-H25 (9WJ)	RI	C2710-G2697	U1629-A1637	H60-H96 (3WJ-I. loop)
AGPM	G659-U650	G645-C640	23S (1NKW)	AGPM	G2702-U2666	G1678-C1982	23S (1NKW)
RI	G660-G649	C646-G639	H29-H31 (4WJ)	RI	C2703-G2665	U1679-A1981	H61-H96 (I. loop-4WJ)
AGPM	G740-U731	G695-G690	23S (1S72)	AGPM	G2758-U2724	G1739-C2040	23S (1S72)
RI	C741-G730	C696-G689	H29-H31 (5WJ)	RI	C2759-G2723	U1740-A2039	H61-H96 (I. loop-4WJ)
AGPM	G649-U639	C634-G629	23S (2AW4)	AGPM	G2722-U2687	G1661-C1999	23S (2AW4)
RI	C650-G638	C635-G628	H29-H31 (4WJ)	RI	C2723-G2686	U1662-A1998	H61-H96 (I. loop-4WJ)
AGPM	G649-U639	C634-G629	23S (2J01)	AGPM	G2722-U2687	G1661-C1999	23S (2J01)
RI	C650-G638	C635-G628	H29-H31 (4WJ)	RI	C2723-G2686	C1662-G1998	H61-H96 (I. loop-4WJ)
AGPM	G610-U670	G634-C615	23S (1NKW)	AGPM	G2320-U2270	C2358-G2353	23S (1NKW)
RI	C611-G669	C635-G614	H27-H28 (4WJ)	RI	C2321-G2269	U2359-A2352	H83-H87 (4WJ)
AGPM	G684-U662	G657-C748	23S (1S72)	AGPM	G2375-C2325	G2416-C2411	23S (1S72)
RI	C685-G661	C658-G747	H27-H28 (5WJ)	RI	C2376-G2324	C2417-G2410	H83-H87 (4WJ)
AGPM	G600-U657	C623-G605	23S (2AW4)	AGPM	G2341-U2291	G2379-C2374	23S (2AW4)
RI	C601-G656	C624-G604	H27-H28 (4WJ)	RI	C2342-G2290	C2380-G2373	H83-H87 (4WJ)
AGPM	G600-U657	G623-C605	23S (2J01)	AGPM	G2341-U2291	G2379-C2374	23S (2J01)
RI	C601-G656	C624-G604	H27-H28 (4WJ)	RI	C2342-G2290	C2380-G2373	H83-H87 (4WJ)
AGPM	G544-C501	G35-C549	16S (2AVY)	AGPM	G105-U62	G384-C379	16S (2AVY)
RI	C545-G500	C36-G548	H3-H18 (5WJ)	RI	C106-G61	C385-G378	H6-H15 (5WJ-I. loop)
AGPM	G544-C501	G35-C549	16S (2J00)	AGPM	G105-U62	G384-C379	16S (2J00)
RI	C545-G500	C36-G548	H3-H18 (5WJ)	RI	C106-G61	C385-G378	H6-H15 (5WJ-I. loop)
Int.	**AGPM & Ribo-base type II**		**Location**	**Int.**	**AGPM & Ribo-base type II**		**Location**
AGPM	G2844-U2822	G2697-C2671	23S (1NKW)	AGPM	G950-U852	C829-G1205	23S (1NKW)
RII	C2845-G2821	G2698-C2670	H96-H101 (I. loop-I. loop)	RII	G951-C851	C830-G1204	H36-H38 (7WJ)
AGPM	G2892-U2864	G2753-C2729	23S (1S72)	AGPM	G1038-U932	U909-A1296	23S (1S72)
RII	C2893-G2863	G2754-C2728	H96-H101 (I. loop-I. loop)	RII	G1039-C931	C910-G1295	H36-H38 (7WJ)
AGPM	G2869-U2847	C2717-G2692	23S (2AW4)	AGPM	G939-U839	C816-G1191	23S (2AW4)
RII	C2870-G2846	G2718-C2691	H96-H101 (I. loop-I. loop)	RII	G940-C838	C817-G1190	H36-H38 (7WJ)
AGPM	G2869-U2847	G2717-C2692	23S (2J01)	AGPM	G939-U839	C816-G1191	23S (2J01)
RII	C2870-G2846	G2718-C2691	H96-H101 (I. loop-I. loop)	RII	G940-C838	C817-G1190	H36-H38 (7WJ)
AGPM	G1861-U1856	G2388-C427	23S (1NKW)	AGPM	G584-U757	G821-C879	16S (2AVY)
RII	C1862-G1855	G2389-C426	H68-X (I. loop-pseudoknot)	RII	G585-C756	U822-A878	H20-H25 (3WJ-4WJ)
AGPM	G1878-U1864	G2409-C414	23S (2AW4)	AGPM	G584-U757	G821-C879	16S (2J00)
RI	C1879-G1863	G2410-C413	H68-X (I. loop-pseudoknot)	RII	G585-C756	C822-G878	H20-H25 (3WJ-4WJ)
AGPM	G1878-U1864	G2409-C414	23S (2J01)	**Int.**	**AGPM**		**Location**
RI	C1879-G1863	G2410-C413	H68-X (I. loop-pseudoknot)	AGPM	A2257-U2241	U2366-A2307	23S (1NKW) H81-X

Appendix B

Supplementary Information for Chapter 3

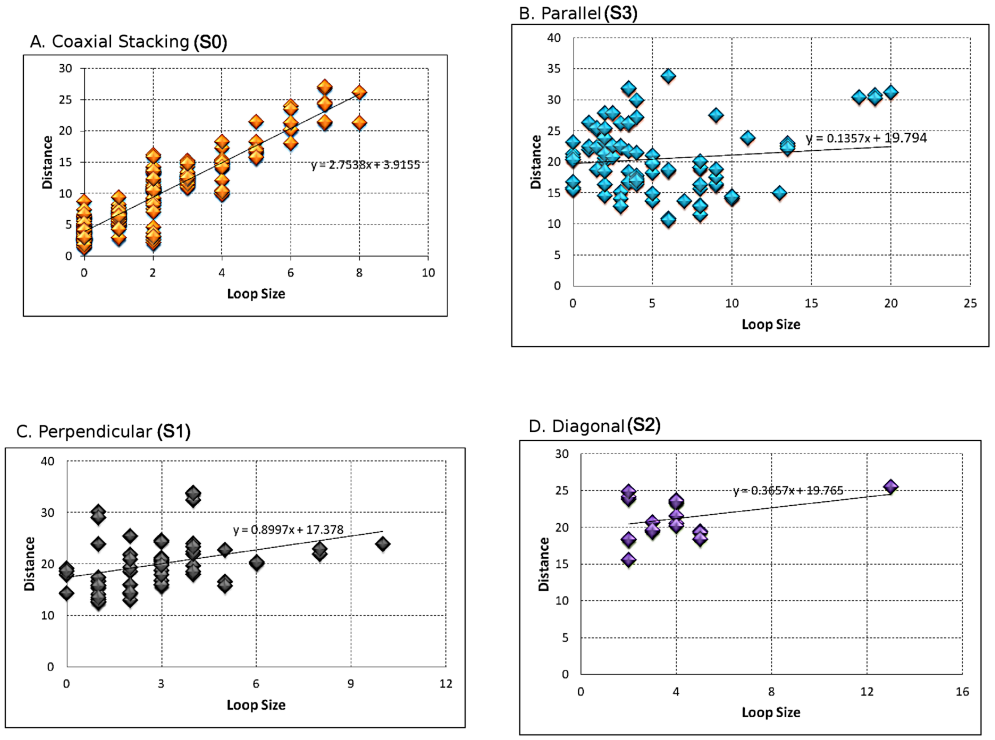


Figure B.1: Distribution of distances with respect to various loop sizes for coaxial stacking of helices **(A)**, parallel **(B)**, perpendicular **(C)**, and diagonal helical arrangements within junctions **(D)**.

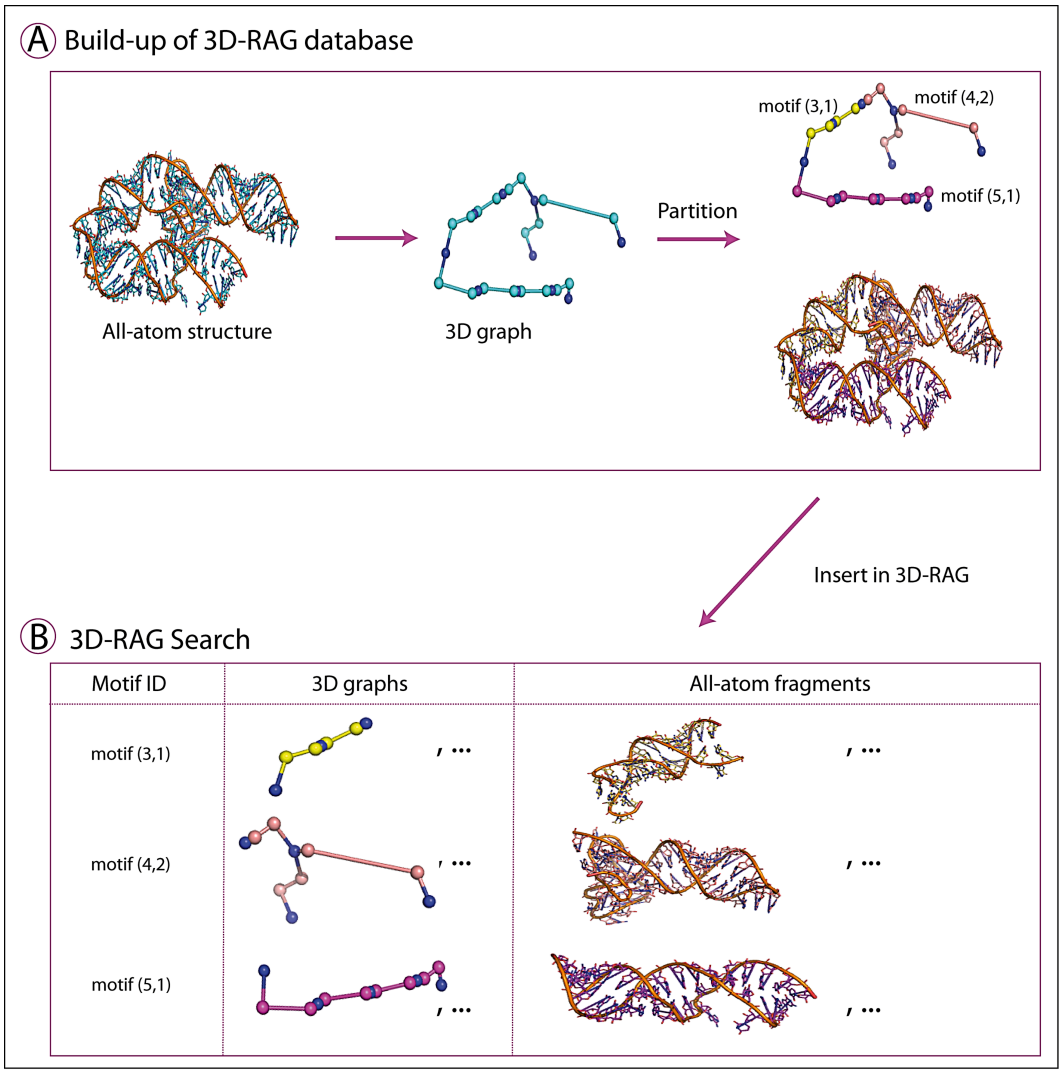


Figure B.2: Illustration of the 3D-RAG build-up and search. **(A)** All-atom structures extracted from known structures are translated into 3D graphs and partitioned into subgraphs based on RAG motif IDs. The subgraphs and all-atom fragments are catalogued in 3D-RAG. **(B)** 3D-RAG can be used for the search of graph similarity. After identifying the motif ID of the target graph, one can search for graph match in the motif ID selected, and extract the corresponding all-atom fragment.

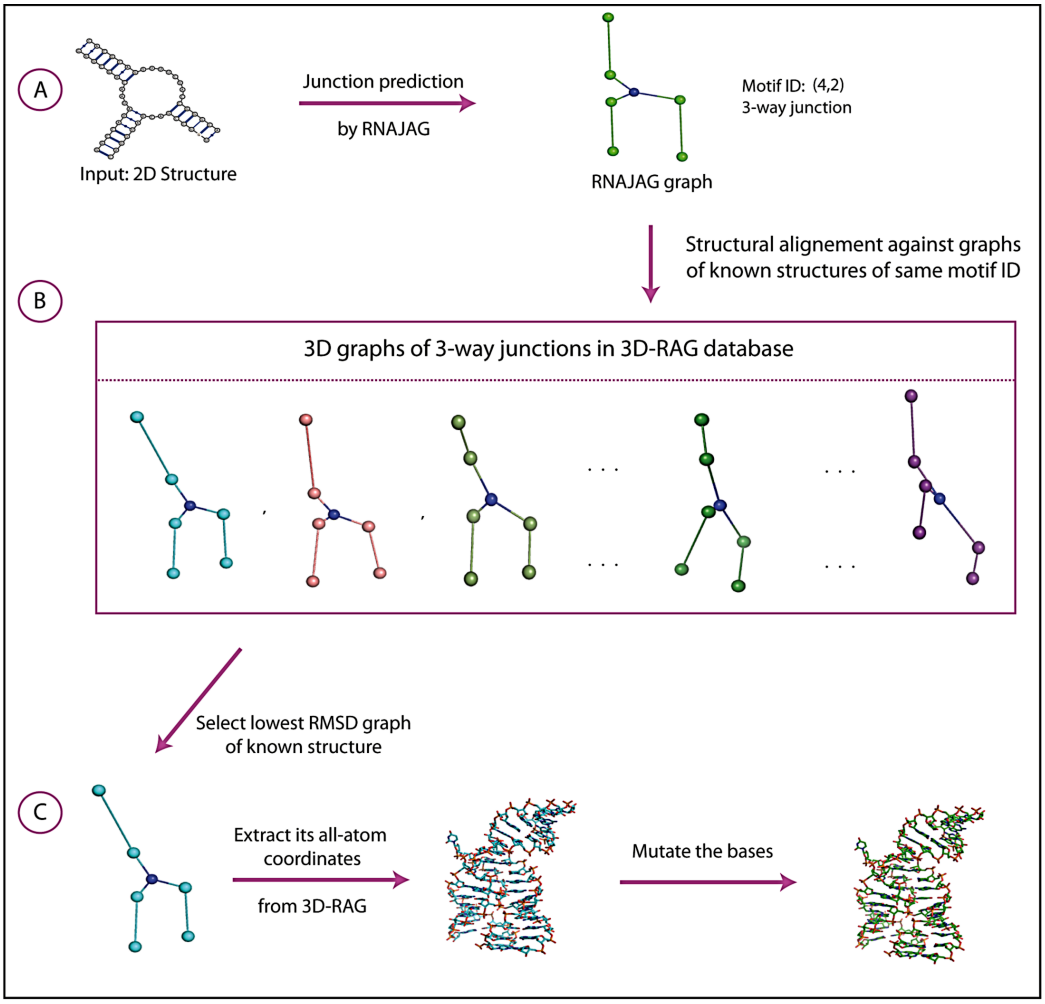


Figure B.3: Illustration of the threading approach for the prediction of the all-atom RNA structure for a 3-way junction. **(A)** Predicted graph by RNAJAG. **(B)** Search for graph similarities in 3D-RAG by superimposing the predicted RNAJAG graph with junction graphs of the same motif ID extracted from known structures. **(C)** Selection of the best graph candidate of known structures with the lowest RMSD, extraction of its all-atom coordinates from the database, and mutation of the bases to match those of the target sequence.

Table B.1: List of RNA 3D structures containing 224 junction data used for distance parameter estimation

PDB	Degree[a]	Stacked Helices[b]	Loop Size[c]	B1[d]	C1[e]	B2[f]	C2[g]	B3[h]	C3[i]	B4[j]	C4[k]	Distance[l] (Å)
1E8O	3WJ	H_1H_3	0	143	E	144	E	103	E	128	E	3.361
1EFW	4WJ	H_1H_4	0	65	B	66	B	7	B	49	B	3.361
1EHZ	4WJ	H_1H_4	0	65	E	66	E	7	E	49	E	3.361
1KH6	4WJ	H_2H_3	0	23	A	24	A	37	A	8	A	7.402
1KH6	4WJ	H_1H_4	0	48	A	49	A	5	A	39	A	2.586
1M5O	4WJ	H_1H_4	0	7	0	8	0	14	0	85	0	2.586
1M5O	4WJ	H_2H_3	0	68	0	69	0	84	0	15	0	2.586
1N78	4WJ	H_1H_4	0	565	C	566	C	507	C	549	C	3.814
1NBS	3WJ	H_1H_3	0	232	A	233	A	133	A	180	A	8.039
1NBS	4WJ	H_1H_4	0	234	A	235	A	90	A	132	A	3.137
1NKW	3WJ	H_1H_2	0	55	0	56	0	69	0	112	0	3.246
1NKW	5WJ	H_3H_4	0	128	0	129	0	142	0	119	0	5.832
1NKW	7WJ	H_2H_3	0	848	0	849	0	954	0	835	0	5.094
1NKW	4WJ	H_2H_3	0	1307	C	1308	C	1662	C	1289	C	5.094
1NKW	4WJ	H_1H_4	0	1992	C	1993	C	1283	C	1665	C	5.094
1NKW	3WJ	H_1H_2	0	2789	0	2790	0	2806	0	2861	0	6.341
1QRS	4WJ	H_1H_4	0	65	D	66	D	7	D	49	D	6.341
1S72	3WJ	H_2H_3	0	52	0	53	0	66	0	109	0	3.727
1S72	7WJ	H_6H_7	0	1267	0	1268	0	1290	0	1089	0	3.467
1S72	4WJ	H_2H_3	0	1400	A	1401	A	1721	A	1382	A	3.467
1S72	3WJ	H_1H_2	0	1551	0	1552	0	1569	0	1634	0	2.451
1S72	4WJ	H_1H_4	0	2050	A	2051	A	1374	A	1725	A	2.451
1S72	4WJ	H_2H_3	0	2712	X	2713	X	2767	X	2682	X	2.451
1S72	3WJ	H_1H_2	0	2831	0	2832	0	2848	0	2909	0	5.714
1U0B	4WJ	H_1H_4	0	65	A	66	A	7	A	49	A	2.927
1U6B	3WJ	H_2H_3	0	93	B	94	B	124	B	51	B	4.534
1U6B	5WJ	H_3H_4	0	143	B	144	B	166	B	41	B	4.259
1U9S	4WJ	H_1H_4	0	230	A	231	A	79	A	110	A	3.227
1UN6	3WJ	H_2H_3	0	65	E	66	E	109	E	14	E	5.312
2A2E	4WJ	H_2H_3	0	87	A	88	A	101	A	73	A	2.779
2A64	3WJ	H_1H_2	0	62	A	63	A	81	A	250	A	3.492
2A64	3WJ	H_1H_3	0	242	A	243	A	140	A	190	A	5.815
2A64	4WJ	H_1H_4	0	244	C	245	C	91	C	139	C	5.815
2A64	6WJ	H_5H_6	0	307	A	308	A	329	A	278	A	4.011
2A64	3WJ	H_2H_3	0	344	A	345	A	384	A	15	A	4.416
2AVY	4WJ	H_1H_4	0	311	A	312	A	115	A	289	A	3.185
2AVY	3WJ	H_1H_2	0	1073	A	1074	A	1083	A	1102	A	3.127
2AW4	3WJ	H_1H_2	0	56	0	57	0	70	0	114	0	3.127
2AW4	5WJ	H_3H_4	0	130	B	131	B	148	B	121	B	5.113
2AW4	7WJ	H_6H_7	0	1163	B	1164	B	1185	B	991	B	4.328
2AW4	4WJ	H_2H_3	0	1294	B	1295	B	1645	B	1276	B	4.303
2AW4	4WJ	H_1H_2	0	1444	B	1445	B	1466	B	1547	B	3.484
2AW4	4WJ	H_1H_4	0	2009	B	2010	B	1270	B	1648	B	3.769
2BTE	5WJ	H_1H_5	0	65	B	66	B	7	B	49	B	3.132
2GDI	3WJ	H_1H_2	0	14	X	15	X	51	X	85	X	4.055
2HOJ	3WJ	H_1H_2	0	14	A	15	A	51	A	85	A	4.630
2J00	4WJ	H_1H_4	0	311	A	312	A	115	A	289	A	3.248
2J00	3WJ	H_1H_2	0	1073	A	1074	A	1083	A	1102	A	2.300
2J01	3WJ	H_1H_2	0	56	A	57	A	70	A	114	A	2.870
2J01	3WJ	H_2H_3	0	68	B	69	B	108	B	16	B	3.913
2J01	7WJ	H_2H_3	0	835	A	836	A	943	A	822	A	3.173
2J01	7WJ	H_6H_7	0	1267	A	1268	A	1290	A	991	A	46.970
2J01	4WJ	H_2H_3	0	1294	A	1295	A	1645	A	1276	A	3.565
2J01	4WJ	H_1H_4	0	2009	A	2010	A	1270	A	1648	A	3.242
2J01	3WJ	H_1H_2	0	2814	A	2815	A	2831	A	2886	A	6.816
2NR0	5WJ	H_1H_5	0	65	F	66	F	7	F	49	F	2.680
2NZ4	3WJ	H_1H_2	0	7	P	8	P	26	P	52	P	3.736
2OIU	3WJ	H_1H_2	0	7	P	8	P	17	P	45	P	5.147

3IIN	3WJ	H_2H_3	0	93	B	94	B	124	B	51	B	4.323
3IVK	4WJ	H_1H_4	0	54	C	55	C	69	C	106	C	4.066
3IVK	4WJ	H_2H_3	0	85	C	86	C	105	C	72	C	2.556
1EFW	4WJ	H_2H_3	1	25	B	27	B	43	B	10	B	2.556
1EHZ	4WJ	H_2H_3	1	25	E	27	E	43	E	10	E	2.556
1L9A	3WJ	H_1H_3	1	220	B	222	B	128	B	177	B	4.903
1LNG	3WJ	H_1H_3	1	231	B	233	B	145	B	187	B	6.645
1N78	4WJ	H_2H_3	1	525	C	527	C	543	C	510	C	8.253
1NBS	4WJ	H_2H_3	1	112	A	114	A	129	A	91	A	4.620
1NKW	3WJ	H_2H_3	1	70	9	72	9	109	9	18	9	8.715
1NKW	4WJ	H_3H_4	1	1458	D	1460	D	1481	E	1563	E	8.715
1NKW	4WJ	H_3H_4	1	2349	0	2351	0	2360	0	2326	0	6.801
1NKW	4WJ	H_2H_3	1	2653	A	2655	A	2711	A	2625	A	6.801
1QRS	4WJ	H_2H_3	1	25	D	27	D	43	D	10	D	6.801
1S72	4WJ	H_3H_4	1	2407	A	2409	A	2418	A	2381	A	6.801
1S72	4WJ	H_1H_4	1	2804	X	2806	X	2679	X	2770	X	6.801
1U0B	4WJ	H_2H_3	1	25	A	27	A	43	A	10	A	8.536
1U9S	4WJ	H_2H_3	1	94	A	96	A	107	A	80	A	6.633
1U9S	4WJ	H_3H_4	1	194	A	196	A	217	A	176	A	5.966
2A2E	4WJ	H_1H_4	1	202	A	204	A	71	A	102	A	7.882
2A64	6WJ	H_1H_2	1	21	A	23	A	44	A	338	A	8.510
2A64	4WJ	H_2H_3	1	113	C	115	C	136	C	92	C	8.510
2AVY	4WJ	H_1H_4	1	219	A	221	A	142	A	198	A	6.703
2AVY	5WJ	H_4H_5	1	337	A	339	A	350	A	316	A	6.242
2AVY	4WJ	H_1H_4	1	879	0	881	0	569	0	821	0	6.242
2AVY	3WJ	H_1H_2	1	1116	A	1118	A	1155	A	1184	A	9.218
2AW4	4WJ	H_3H_4	1	2370	B	2372	B	2381	B	2347	B	6.833
2AW4	4WJ	H_2H_3	1	2674	A	2676	A	2731	A	2646	A	6.833
2AW4	4WJ	H_1H_4	1	2769	A	2771	A	2643	A	2735	A	6.833
2CKY	3WJ	H_1H_2	1	5	A	7	A	39	A	75	A	5.778
2J00	4WJ	H_1H_4	1	219	D	221	D	142	D	198	D	7.584
2J00	5WJ	H_4H_5	1	337	A	339	A	350	A	316	A	6.609
2J00	4WJ	H_1H_4	1	879	B	881	B	569	B	821	B	6.609
2J00	3WJ	H_1H_2	1	1116	A	1118	A	1155	A	1184	A	9.068
2J01	3WJ	H_1H_2	1	1444	A	1445	A	1466	A	1547	A	3.774
2J01	4WJ	H_3H_4	1	2370	A	2372	A	2381	A	2347	A	7.014
2J01	4WJ	H_2H_3	1	2674	A	2676	A	2731	A	2646	A	6.480
2J01	4WJ	H_1H_4	1	2769	A	2771	A	2643	A	2735	A	10.131
2QBZ	3WJ	H_1H_2	1	54	X	56	X	85	X	120	X	5.786
3DIL	5WJ	H_4H_5	1	140	A	142	A	161	A	114	A	6.614
1NKW	3WJ	H_1H_2	2	32	0	35	0	457	0	484	0	11.347
1NKW	3WJ	H_1H_2	2	1319	0	1322	0	1621	0	1638	0	9.623
1NKW	3WJ	H_1H_3	2	1656	0	1659	0	1311	0	1643	0	8.539
1NKW	6WJ	H_1H_2	2	2057	0	2060	0	2222	0	2414	0	3.612
1NKW	3WJ	H_1H_2	2	2496	0	2499	0	2524	0	2546	0	3.424
1NKW	4WJ	H_1H_4	2	2748	A	2751	A	2622	A	2716	A	3.424
1S72	3WJ	H_1H_2	2	29	0	32	0	451	0	479	0	12.250
1S72	7WJ	H_2H_3	2	928	0	931	0	1039	0	915	0	10.593
1S72	3WJ	H_1H_2	2	1205	0	1208	0	1159	0	1194	0	10.494
1S72	3WJ	H_1H_2	2	1412	0	1415	0	1680	0	1697	0	9.769
1S72	4WJ	H_3H_4	2	1511	A	1514	A	1672	A	1494	A	9.769
1S72	3WJ	H_1H_3	2	1715	0	1718	0	1404	0	1703	0	9.078
1S72	6WJ	H_1H_2	2	2115	0	2118	0	2276	0	2470	0	3.196
1S72	3WJ	H_1H_2	2	2552	0	2555	0	2580	0	2602	0	6.175
1X8W	3WJ	H_1H_2	2	138	A	141	A	162	A	180	A	14.948
1Y26	3WJ	H_1H_3	2	72	X	75	X	21	X	54	X	11.897
2A2E	4WJ	H_3H_4	2	169	A	172	A	189	A	151	A	11.371
2AVY	3WJ	H_1H_3	2	392	A	395	A	46	A	369	A	9.675
2AVY	3WJ	H_1H_3	2	1034	A	1037	A	1003	A	1027	A	8.979
2AVY	3WJ	H_2H_3	2	1044	A	1047	A	1210	A	997	A	13.719
2AW4	3WJ	H_1H_2	2	32	B	35	B	445	B	473	B	12.438
2AW4	7WJ	H_2H_3	2	835	B	838	B	940	B	822	B	13.396

2AW4	3WJ	H_1H_3	2	1101	B	1104	B	1055	B	1090	B	10.142
2AW4	3WJ	H_1H_2	2	1306	B	1309	B	1605	B	1622	B	9.864
2AW4	4WJ	H_3H_4	2	1402	X	1405	X	1597	X	1385	X	9.864
2AW4	3WJ	H_1H_3	2	1639	B	1642	B	1298	B	1627	B	7.829
2AW4	6WJ	H_1H_2	2	2074	B	2077	B	2243	B	2435	B	100.760
2AW4	3WJ	H_1H_2	2	2517	B	2520	B	2545	B	2567	B	4.715
2AW4	3WJ	H_1H_2	2	2812	B	2815	B	2831	B	2888	B	8.960
2B57	3WJ	H_1H_3	2	72	A	75	A	21	A	54	A	12.682
2CZJ	3WJ	H_1H_3	2	65	B	68	B	5	B	49	B	7.922
2EES	3WJ	H_1H_3	2	72	A	75	A	21	A	54	A	12.497
2J00	3WJ	H_1H_3	2	392	A	395	A	46	A	369	A	12.033
2J00	3WJ	H_2H_3	2	1044	A	1047	A	1210	A	997	A	15.777
2J01	3WJ	H_1H_2	2	32	A	35	A	445	A	473	A	9.660
2J01	5WJ	H_4H_5	2	1195	A	1198	A	1247	A	812	A	10.374
2J01	3WJ	H_1H_2	2	1306	A	1309	A	1605	A	1622	A	9.136
2J01	4WJ	H_3H_4	2	1402	A	1405	A	1597	A	1385	A	12.796
2J01	3WJ	H_1H_3	2	1639	A	1642	A	1298	A	1627	A	9.462
2J01	6WJ	H_1H_2	2	2074	A	2077	A	2243	A	2435	A	4.429
2J01	3WJ	H_1H_2	2	2517	A	2520	A	2545	A	2567	A	4.925
3BWP	5WJ	H_2H_3	2	40	A	43	A	65	A	28	A	12.188
3EOH	5WJ	H_2H_3	2	40	A	43	A	65	A	28	A	12.160
3F2Q	4WJ	H_2H_3	2	46	X	49	X	60	X	33	X	10.721
3F2Q	4WJ	H_1H_4	2	80	X	83	X	32	X	64	X	11.081
3F4E	4WJ	H_2H_3	2	80	Y	83	Y	32	X	64	Y	10.881
3KTW	3WJ	H_1H_3	2	231	C	234	C	144	C	189	C	8.296
1NKW	3WJ	H_1H_3	3	348	0	352	0	308	0	336	0	13.140
1NKW	3WJ	H_2H_3	3	745	0	749	0	773	0	713	0	10.239
1NKW	3WJ	H_1H_3	3	1111	0	1115	0	1066	0	1102	0	14.143
1NKW	7WJ	H_6H_7	3	1174	0	1178	0	1196	0	1002	0	11.980
1NKW	3WJ	H_2H_3	3	2179	0	2183	0	2202	0	2076	0	13.915
1S72	3WJ	H_2H_3	3	67	9	71	9	110	9	14	9	11.281
1S72	4WJ	H_2H_3	3	1844	A	1848	A	1883	A	1832	A	11.281
1S72	3WJ	H_2H_3	3	2241	0	2245	0	2256	0	2134	0	9.900
1U6B	5WJ	H_1H_2	3	9	B	13	B	37	B	1	C	10.788
1Y0Q	5WJ	H_4H_5	3	138	A	142	A	155	A	127	A	9.545
2AVY	3WJ	H_2H_3	3	651	A	655	A	751	A	588	A	8.167
2AVY	3WJ	H_2H_3	3	857	A	861	A	868	A	829	A	13.616
2AW4	3WJ	H_1H_3	3	337	0	341	0	297	0	325	0	13.616
2AW4	3WJ	H_2H_3	3	732	B	736	B	760	B	700	B	9.793
2AW4	4WJ	H_3H_4	3	1525	B	1529	B	1542	B	1467	B	15.441
2AW4	4WJ	H_2H_3	3	1788	0	1792	0	1827	0	1776	0	15.441
2AW4	4WJ	H_1H_4	3	1975	0	1979	0	1772	0	1830	0	15.441
2AW4	3WJ	H_2H_3	3	2196	B	2200	B	2223	B	2093	B	12.115
2J00	3WJ	H_2H_3	3	651	A	655	A	751	A	588	A	10.730
2J00	3WJ	H_2H_3	3	857	A	861	A	868	A	829	A	11.950
2J01	3WJ	H_1H_3	3	337	A	341	A	297	A	325	A	14.835
2J01	3WJ	H_2H_3	3	732	A	736	A	760	A	700	A	12.963
2J01	4WJ	H_2H_3	3	1788	A	1792	A	1827	A	1776	A	12.204
2J01	4WJ	H_1H_4	3	1975	A	1979	A	1772	A	1830	A	9.544
2J01	3WJ	H_2H_3	3	2196	A	2200	A	2223	A	2093	A	12.133
3F2Q	4WJ	H_2H_3	3	27	X	31	X	84	X	15	X	12.703
3F4E	4WJ	H_1H_4	3	27	X	31	X	84	Y	15	X	12.901
3F4E	4WJ	H_2H_3	3	46	X	50	X	59	Y	33	X	13.379
1NKW	6WJ	H_2H_3	4	156	0	161	0	189	0	45	0	8.530
1NKW	4WJ	H_3H_4	4	1415	0	1420	0	1611	0	1398	0	15.674
1NKW	10WJ	H_9H_{10}	4	2781	B	2786	B	2864	B	2771	B	15.674
1NYI	3WJ	H_2H_3	4	104	B	109	B	118	B	23	A	16.868
1S72	6WJ	H_2H_3	4	149	0	154	0	182	0	42	0	11.583
1S72	5WJ	H_1H_2	4	239	0	244	0	267	0	431	0	16.054
1S72	3WJ	H_1H_3	4	344	0	349	0	302	0	332	0	15.742
1S72	3WJ	H_2H_3	4	823	0	828	0	853	0	791	0	12.121
1S72	5WJ	H_4H_5	4	1300	0	1305	0	1349	0	905	0	14.963

1S72	4WJ	H_1H_4	4	2015	A	2020	A	1828	A	1887	A	14.963
2AVY	3WJ	H_1H_2	4	672	A	677	A	713	A	734	A	14.500
2AVY	4WJ	H_2H_3	4	764	0	769	0	810	0	577	0	14.500
2AW4	6WJ	H_2H_3	4	179	B	184	B	212	B	46	B	11.259
2J00	4WJ	H_2H_3	4	178	D	183	D	194	D	144	D	11.259
2J00	3WJ	H_1H_2	4	672	A	677	A	713	A	734	A	13.244
2J00	4WJ	H_2H_3	4	764	B	769	B	810	B	577	B	13.244
2J01	6WJ	H_2H_3	4	179	A	184	A	212	A	45	A	9.727
3DIL	5WJ	H_1H_2	4	9	A	14	A	78	A	166	A	16.113
3E5C	3WJ	H_2H_3	4	35	A	40	A	47	A	9	A	14.228
1NKW	3WJ	H_1H_2	5	696	0	702	0	786	0	807	0	16.640
1NKW	4WJ	H_1H_4	5	1971	0	1977	0	1683	0	1755	0	16.959
1S72	3WJ	H_1H_2	5	774	0	780	0	866	0	887	0	12.621
1S72	4WJ	H_1H_4	5	2029	A	2035	A	1744	A	1820	A	12.621
1U8D	3WJ	H_1H_2	5	21	A	27	A	43	A	75	A	21.282
2AW4	3WJ	H_1H_2	5	25	B	31	B	474	B	515	B	18.269
2AW4	3WJ	H_1H_2	5	683	B	689	B	773	B	794	B	212.300
2AW4	4WJ	H_1H_4	5	1988	B	1994	B	1666	B	1764	B	16.700
2AW4	10WJ	H_9H_{10}	5	2805	B	2811	B	2889	B	2791	B	197.120
2J01	3WJ	H_1H_2	5	683	A	689	A	773	A	794	A	14.648
2J01	4WJ	H_1H_4	5	1988	A	1994	A	1666	A	1764	A	15.557
1NKW	3WJ	H_1H_2	6	24	0	31	0	485	0	526	0	21.379
1NKW	4WJ	H_3H_4	6	1538	E	1545	E	1558	E	1485	E	21.379
1S72	3WJ	H_1H_2	6	21	0	28	0	480	0	522	0	21.920
2AVY	5WJ	H_4H_5	6	106	A	113	A	314	A	61	A	22.433
2J00	5WJ	H_2H_3	6	106	A	113	A	314	A	61	A	24.173
2J00	3WJ	H_1H_3	6	1033	A	1040	A	1001	A	1029	A	18.903
2J01	3WJ	H_1H_2	6	24	A	31	A	474	A	516	A	23.546
1NKW	10WJ	H_4H_5	7	1274	B	1282	B	1994	B	588	B	23.546
2AVY	4WJ	H_1H_2	7	19	A	27	A	556	A	916	A	27.524
2AVY	4WJ	H_2H_3	7	176	A	184	A	193	A	146	A	24.226
2AW4	10WJ	H_4H_5	7	1261	B	1269	B	2011	B	579	B	208.340
2J00	4WJ	H_1H_2	7	19	A	27	A	556	A	916	A	28.859
2J01	9WJ	H_4H_5	7	1261	A	1269	A	2011	A	589	A	40.051
2J01	5WJ	H_4H_5	7	2580	A	2588	A	2606	A	2508	A	22.892
3F2Q	4WJ	H_1H_4	7	97	X	105	X	8	X	86	X	23.366
3F4E	4WJ	H_1H_4	7	97	Y	105	Y	8	X	86	Y	23.324
1NKW	5WJ	H_2H_3	8	2425	0	2434	0	2475	0	2407	0	28.317
2AW4	5WJ	H_2H_3	8	2446	B	2455	B	2500	B	2064	B	99.709
2J01	5WJ	H_2H_3	8	2446	A	2455	A	2496	A	2064	A	20.078
1U9S	4WJ	H_1H_2	12	119	A	132	A	168	A	222	A	20.284

[a] Degree of each RNA junction; 3WJ, for example, represents three-way junction.

[b] Two helices involved in coaxial stacking where H_1H_2, for instance, denotes helices 1 and 2.

[c] Loop size between coaxially stacked helices.

[d,f,h,j] Each Base position at the helix end of coaxially stacked helices.

[e,g,i,k] Chain ID of each base at the helix end of coaxially stacked helices.

[l] Distance between coaxially stacked helices.

Table B.2: List of 200 RNA junctions from the PDB database. Each junction is listed with its junction family and coaxial stacking arrangement from the native structure and RNAJAG prediction.

3-way junction						
PDB_ResID	Nts	Native		Predictions		RMSD
		Stacking	Family	Stacking	Family	
2A64_128	20	H1H3	C	H1H3	C	2.91
3U5F_1032	21	H2H3	A	H2H3	A	4.21
1NBS_48	22	H1H3	C	H1H3	C	2.88
3IZ9_3356	22	H1H2	A	H1H2	A	6.63
3U5F_630	24	H2H3	A	H2H3	A	5.31
3IZ9_1470	24	H1H2	A	H1H2	A	6.15
1S1I_1371	24	H1H2	A	H1H2	A	6.26
3BBN_947	27	H1H3	C	H1H3	C	3.52
3BBN_529	27	H2H3	A	H2H3	A	4.86
3D2V_5	30	H1H2	A	H1H2	A	2.07
4A1B_2641	30	H1H2	C	H1H2	C	3.03
3IZ9_2844	30	H1H2	C	H1H2	C	3.13
3JYV_1246	30	H1H3	B	**H2H3**	B	7.46
1S1I_1363	30	H1H3	A	H1H3	A	9.49
4A1C_47	31	H1H2	A	H1H2	**C**	3.05
1FJG_42	32	H1H3	C	H1H3	C	2.38
2ZJR_55	32	H1H2	A	H1H2	A	6.45
1PNU_32	32	H1H2	A	H1H2	A	8.75
1C2W_1055	33	H1H3	C	H1H3	C	3.98
3IZ9_333	33	H1H3	A	H1H3	A	4.15
1C2W_2517	33	H2H3	A	H2H3	A	5.10
3IZ9_1224	33	H1H3	C	H1H3	C	6.42
1S72_1119	33	H1H3	C	H1H3	C	6.46
3MOJ_10	33	H1H2	C	H1H2	**A**	8.40
1S1I_20	33	H1H2	A	H1H2	A	8.88
3V2F_27	33	H1H2	A	H1H2	A	8.88
3IZ9_1474	33	H1H2	C	H1H2	C	10.46
1S1I_2435	34	H1H2	C	H1H2	C	2.71
2QBG_2455	34	H1H2	C	H1H2	C	2.96
1PNU_2447	34	H1H2	C	H1H2	C	3.44
1UN6_6	34	H2H3	C	H2H3	C	4.71
2P89_11	34	H1H2	C	H1H2	C	5.10
2XZM_1300	34	H1H2	C	H1H2	C	7.37
3U5F_1278	34	H1H3	B	H1H3	B	8.55
1PNU_9	35	H2H3	C	H2H3	C	3.79
3V2F_11	35	H2H3	C	H2H3	C	4.01
3SD3_20	35	H1H3	A	H1H3	A	4.43
2AW7_999	35	H1H3	B	H1H3	**A**	4.63
3P49_9	35	H2H3	B	H2H3	B	8.56
1E8O_5	36	H1H3	C	H1H3	C	4.80
3IZ9_9	36	H2H3	C	H2H3	C	4.86
1S1I_9	36	H2H3	C	H2H3	**A**	6.30
3V2F_308	36	H1H3	C	H1H3	C	7.79
1FJG_963	36	H2H3	B	H2H3	B	8.20
3IZD_14	36	H1H2	C	**H2H3**	**A**	8.22
3U5F_1208	36	H2H3	B	H2H3	B	8.60

4A1B_1516	37	H1H2	A	H1H2	A	6.28
1C2X_10	37	H2H3	C	H2H3	C	7.44
1C2W_32	37	H1H2	A	H1H2	A	7.55
3U5H_9	38	H2H3	C	H2H3	C	5.23
3IZF_1504	38	H1H2	A	**H1H3**	A	8.05
3JYV_1298	38	H1H2	C	H1H2	**B**	9.82
4A1B_853	39	H2H3	A	H2H3	A	7.02
3IZF_818	39	H2H3	A	H2H3	A	7.03
3IZ9_500	39	H1H2	A	H1H2	A	9.36
1S72_2078	40	H2H3	A	H2H3	A	3.90
2XZM_1179	40	H2H3	B	H2H3	B	10.19
1S72_776	40	H2H3	B	**H1H2**	B	10.59
3SUX_23	41	H1H3	A	H1H3	A	5.14
1RMN_5	41	H2H3	C	H2H3	C	5.32
2XZM_1125	41	H1H2	B	H1H2	B	8.21
1FJG_909	41	H1H2	B	H1H2	B	10.02
1FOQ_24	42	H1H3	B	H1H3	**A**	4.85
3IZF_494	43	H1H2	A	H1H2	A	6.54
4A1B_547	43	H1H2	A	H1H2	A	7.05
3U5F_1156	43	H1H2	B	H1H2	B	10.20
4A1B_2291	44	H1H2	A	H1H2	A	6.06
1PNU_1263	44	H1H2	A	H1H2	A	6.24
3JYX_338	44	H1H3	C	H1H3	C	6.69
1C2W_697	45	H2H3	A	H2H3	A	8.60
3IZF_429	45	H2H3	B	H2H3	B	10.53
3IZ9_2399	46	H2H3	A	H2H3	A	3.81
2QBG_2072	46	H2H3	A	H2H3	A	4.74
1C2W_2092	46	H2H3	A	H2H3	A	5.49
1J2B_7	46	H1H3	A	H1H3	A	5.61
1FJG_805	46	H2H3	A	H2H3	A	5.65
3BBO_5	46	H1H2	A	H1H2	A	6.90
3IZF_178	46	H1H3	C	H1H3	C	8.04
3IZ9_178	46	H1H3	C	H1H3	C	8.05
3BBN_769	46	H2H3	B	H2H3	B	10.34
1FJG_1094	46	H1H2	C	H1H2	C	10.40
2AW7_1112	46	H1H2	C	H1H2	C	10.47
3JYV_1007	47	H2H3	B	H2H3	B	8.50
2QBZ_40	48	H1H2	A	H1H2	A	6.01
3IZ9_812	48	H2H3	B	**H1H3**	**C**	10.46
3FO4_5	49	H1H3	C	H1H3	C	3.45
1S1I_1510	49	H1H2	C	H1H2	C	5.17
3JYX_2472	49	H1H2	A	H1H2	A	7.69
3IZ9_429	49	H2H3	B	H2H3	B	7.84
2AW7_822	50	H2H3	A	H2H3	A	5.45
4A1B_184	50	H1H3	C	H1H3	C	6.39
1PNU_2026	52	H2H3	A	H2H3	A	6.38
2J37_17	53	H1H3	C	H1H3	C	7.67
3U5H_453	54	H1H2	A	H1H2	A	7.50
1Y26_9	55	H1H3	C	H1H3	C	4.08
3SLQ_9	57	H1H3	C	H1H3	C	2.86
2ZJR_1957	62	H2H3	B	H2H3	B	9.71

1S1I_12	64	H1H2	C	H1H2	C	5.82

4-way junction						
PDB_ResID	Nts	Native		Predictions		RMSD
		Stacking	Family	Stacking	Family	
3IZF_1639	35	H1H2H3H4	cH	H1H2H3H4	cH	6.08
3U5H_1480	35	H3H4	cK	H3H4	cK	13.35
4A1B_1507	35	H3H4	cK	H3H4	cK	13.42
3BBN_94	36	H1H4H2H3	cL	H1H4H2H3	cL	7.85
3JYV_98	36	H1H4H2H3	cL	H1H4H2H3	cL	8.29
2NR0_5	37	H1H4H2H3	cL	H1H4H2H3	cL	2.18
1KH6_5	37	H1H4H2H3	cH	H1H4H2H3	cH	2.76
3F4G_32	37	H1H4H2H3	cH	H1H4H2H3	cH	10.36
3IVK_61	38	H1H2H3H4	cH	H1H2H3H4	**H**	6.81
2A64_91	39	H1H4H2H3	H	H1H4H2H3	H	6.73
4A1B_1429	39	H1H4H2H3	cL	H1H4H2H3	cL	8.45
1C2W_1270	39	H1H4H2H3	cL	H1H4H2H3	cL	24.08
3QSY_7	40	H1H4H2H3	cL	H1H4H2H3	cL	4.29
3U5H_1942	41	H1H4H2H3	H	H1H4H2H3	H	4.07
4A1B_1950	41	H1H4H2H3	H	H1H4H2H3	H	4.12
1S72_1787	41	H1H4H2H3	H	H1H4H2H3	H	4.17
3V2F_1700	41	H1H4H2H3	H	H1H4H2H3	H	4.41
3IZ9_2966	41	H1H4H2H3	cH	H1H4H2H3	cH	6.08
2AKE_7	42	H1H4H2H3	cL	H1H4H2H3	cL	2.36
3AMU_7	42	H1H4H2H3	cL	H1H4H2H3	cL	2.57
1S1I_1333	42	H1H4H2H3	cL	H1H4H2H3	cL	8.45
2AW7_111	42	H1H4H2H3	cL	H1H4H2H3	cL	11.06
1S72_34	42	none	ψ	**H1H4H2H3**	**cL**	14.08
2DER_5	43	H1H4H2H3	cL	H1H4H2H3	cL	2.40
4A1B_2766	43	H1H4H2H3	cH	H1H4H2H3	cH	4.72
3U5H_2764	43	H1H4H2H3	cH	H1H4H2H3	cH	4.93
3V2F_1389	43	H1H2H3H4	cH	H1H2H3H4	cH	7.14
1S72_1413	43	H3H4	cK	H3H4	cK	14.64
3IZ9_1511	43	H3H4	cK	H3H4	cK	14.73
1B23_7	45	H1H4H2H3	cL	H1H4H2H3	cL	1.91
2ZJR_2431	45	H1H4H2H3	cH	H1H4H2H3	cH	5.56
1S72_2519	45	H1H4H2H3	cH	H1H4H2H3	cH	5.70
3V2F_2552	45	H1H4H2H3	cH	H1H4H2H3	cH	5.81
2DU3_7	46	H1H4H2H3	cL	H1H4H2H3	cL	2.01
1F7V_7	46	H1H4H2H3	cL	H1H4H2H3	cL	2.26
2DU6_7	46	H1H4H2H3	cL	H1H4H2H3	cL	2.38
2ZJR_1264	46	H3H4	cK	H3H4	cK	15.08
2OM7_7	47	H1H4H2H3	cL	H1H4H2H3	cL	2.29
1U0B_7	47	H1H4H2H3	cL	H1H4H2H3	cL	2.45
2D6F_7	47	H1H4H2H3	cL	H1H4H2H3	cL	3.02
3EPH_6	48	H1H4H2H3	cL	H1H4H2H3	cL	2.69
1GAX_6	48	H1H4H2H3	cL	H1H4H2H3	cL	2.87
1FIR_7	48	H1H4H2H3	cL	H1H4H2H3	cL	3.00
1PNU_553	48	none	X	**H1H4H2H3**	**H**	7.67
2IHX_8	48	none	X	none	**H**	11.62

1IL2_7	49	H1H4H2H3	cL	H1H4H2H3	cL	3.02
3J16_7	49	H1H4H2H3	cL	H1H4H2H3	cL	3.14
3AL0_7	49	H1H4H2H3	cL	H1H4H2H3	cL	3.36
2CV1_7	49	H1H4H2H3	cL	H1H4H2H3	cL	3.40
1FJG_132	49	H1H4H2H3	cH	H1H4H2H3	cH	6.68
3F4G_8	49	H1H4H2H3	cH	H1H4H2H3	cH	7.81
1S1H_121	49	none	π	none	**cH**	16.75
2K4C_7	50	H1H4H2H3	cL	H1H4H2H3	cL	2.11
1QU2_7	50	H1H4H2H3	cL	H1H4H2H3	cL	2.58
2AZX_7	50	H1H4H2H3	cL	H1H4H2H3	cL	2.61
486D_7	50	H1H4H2H3	cL	H1H4H2H3	cL	2.87
3TUP_7	50	H1H4H2H3	cL	H1H4H2H3	cL	2.90
2ZUF_7	50	H1H4H2H3	cL	H1H4H2H3	cL	2.98
3L0U_7	50	H1H4H2H3	cL	H1H4H2H3	cL	3.02
3A2K_7	50	H1H4H2H3	cL	H1H4H2H3	cL	3.08
3KFU_7	50	H1H4H2H3	cL	H1H4H2H3	cL	3.43
1QF6_7	50	H1H4H2H3	cL	H1H4H2H3	cL	3.62
1EIY_7	50	H1H4H2H3	cL	H1H4H2H3	cL	3.66
2ZJR_1361	50	H1H2H3H4	cH	H1H2H3H4	cH	6.42
3V2F_1293	50	H3H4	cK	H3H4	cK	15.80
2AW7_138	51	H1H4H2H3	cH	H1H4H2H3	cH	6.86
2ZJR_533	51	none	X	none	X	15.12
3Q1Q_7	52	H1H4H2H3	cL	H1H4H2H3	cL	2.98
4AEB_8	52	H1H4H2H3	cL	H1H4H2H3	cL	10.33
4A1B_1650	53	H1H2H3H4	cH	H1H2H3H4	cH	3.61
3U5H_1625	53	H1H2H3H4	cH	H1H2H3H4	cH	3.86
3JYV_598	54	H1H4H2H3	cL	H1H4H2H3	cL	7.99
1C2W_1307	55	H1H2	π	**H1H4H2H3**	**cL**	18.37
1PNU_244	55	none	cW	none	cW	23.42
1U9S_44	56	H3H4	π	H3H4	**cH**	7.09
2QBG_601	56	none	X	none	**H**	12.88
1C2W_1444	56	H1H2	cK	H3H4	cK	15.69
1S72_1848	57	none	π	none	**cH**	11.52
3BBN_512	58	H1H4H2H3	cL	H1H4H2H3	cL	8.79
2QBG_268	58	none	cX	none	cX	19.39
3JYX_1567	59	H2H3	cK	none	cK	25.32
2XZM_1131	60	none	cX	none	cX	17.40
4A1B_1868	64	none	cW	none	cW	18.27
3U5H_1849	64	none	cW	none	cW	18.45
4A1B_2011	65	none	π	none	**cH**	14.88
3IZF_668	66	H3H4	cK	**none**	**cX**	18.31
3V2F_1761	67	none	π	none	**cH**	14.41
3BBO_1765	67	none	π	none	π	15.06
3JYX_1998	68	none	cW	none	cW	19.98
3U5F_1163	70	none	cX	none	cX	17.98
3BBN_887	72	none	cX	none	cX	21.12
2ZJR_1586	76	none	cW	none	cW	19.93
1S72_1703	78	none	cW	none	cW	20.29
3IZ9_1869	78	none	cW	none	cW	20.32
3V2F_1610	84	none	cW	none	cW	21.79
3BBN_1004	85	none	π	none	**cX**	24.47
3BBO_23	93	none	cX	none	cX	23.48
3Q1Q_18	103	H1H2	cK	**none**	cK	24.83

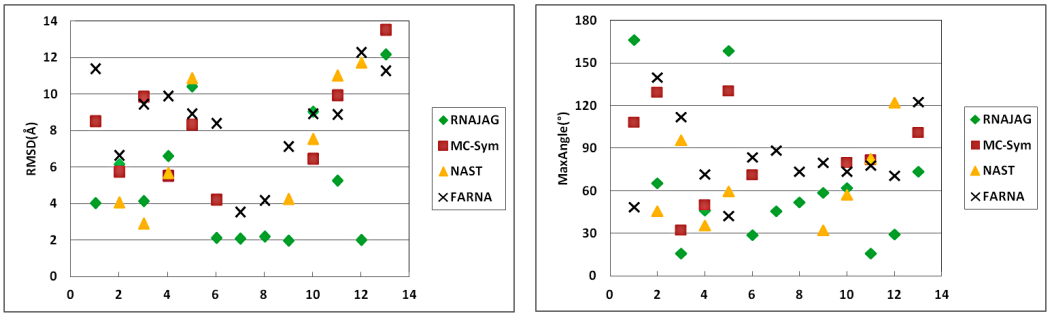


Figure B.4: Distribution of RMSD and MaxAngle for the representative 13 RNA junctions using RNAJAG and other 3D structure prediction programs.

Appendix C

Supplementary Information for Chapter 4

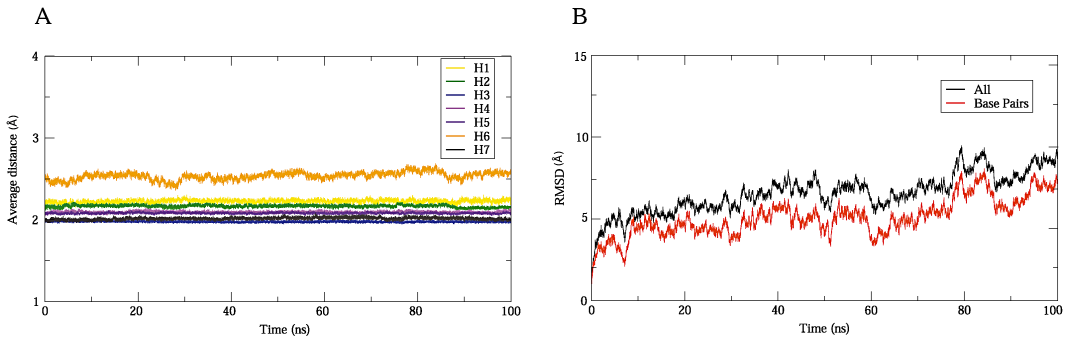


Figure C.1: Average distance of base pairs in helices **(A)** and RMSDs of the entire system (116 residues) and only base pairs (82 residues) with respect to the starting structure, respectively **(B)**.

Sequence	GUAA	GUGA	GCAA	GCGA	GAGA
Count	233	54	22	8	1

Table C.1: Sequences of GNRA loop in 318 FMDV IRES domain 3

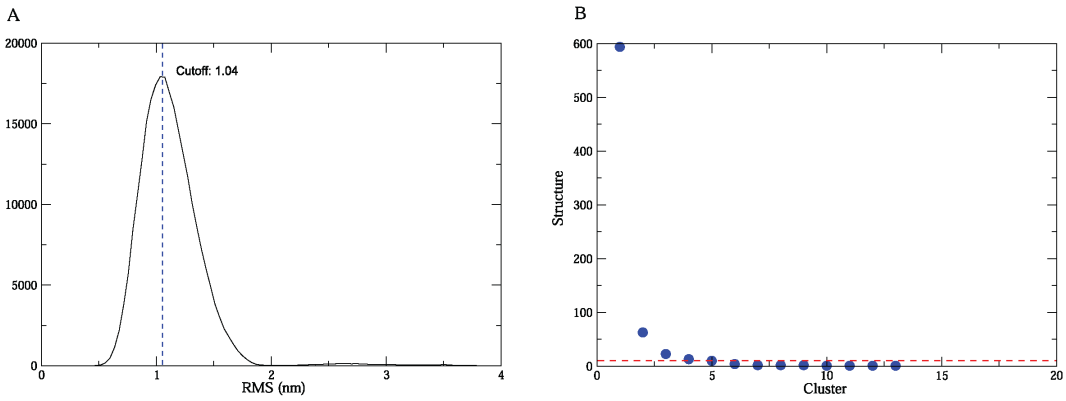


Figure C.2: RMSD distribution and clustering analysis of the 717 structures based on the basal region, G_{86} to U_{133} and C_{249} to C_{299}, in domain 3. The 3D models are obtained using 2D information of FMDV C-S8 IRES domain 3. With equally distributed 101 bins formed between zero and a maximum RMSD value of 3.97 nm, each RMSD value from either upper or lower triangular RMSD matrix (717 $\times$ 717) is put into a right bin. The cutoff value of 1.04 separates the peak **(A)**. Clustering analysis of the 717 structures is based on the overall helical shape of the basal region. The RMSDs of these structures range from 0.44 to 3.97 nm. Thirteen clusters are found, of which the first four clusters contain at least 10 structures **(B)**.

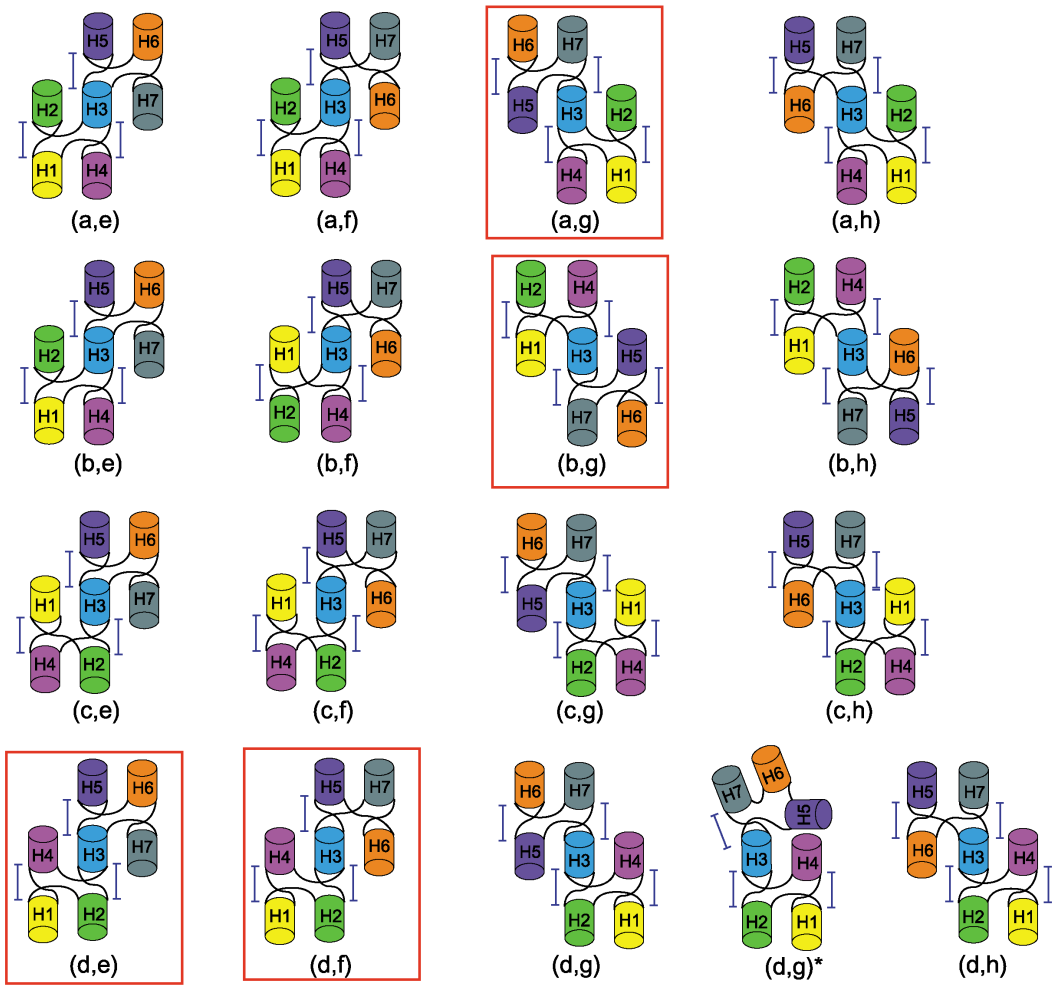


Figure C.3: 17 combinations of two four-way junction topologies including (d,g)*, a modified combination of (d,g),with shown in Figure 4.5B. The four combinations (a,g), (b,g), (d,e) and (d,f) are highlighted with red box as candidate topologies considering potential long-range interactions between H_4 and H_5

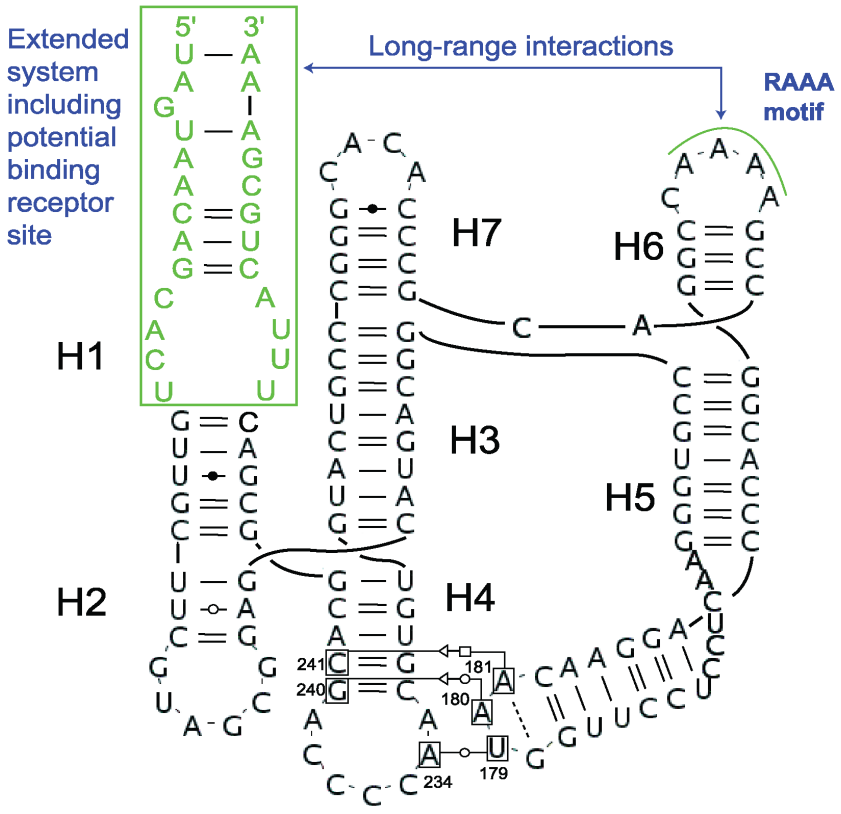


Figure C.4: An extended 2D structure including a potential binding receptor site of RAAA motif. Experimental data suggests long-range interactions between RAAA motif and the extended system ($U_{121}...A_{261}$). In 3D space, the plane of junction I and II are perpendicular that the spatial distance between these structural elements involving RAAA long-range interactions are relatively close.

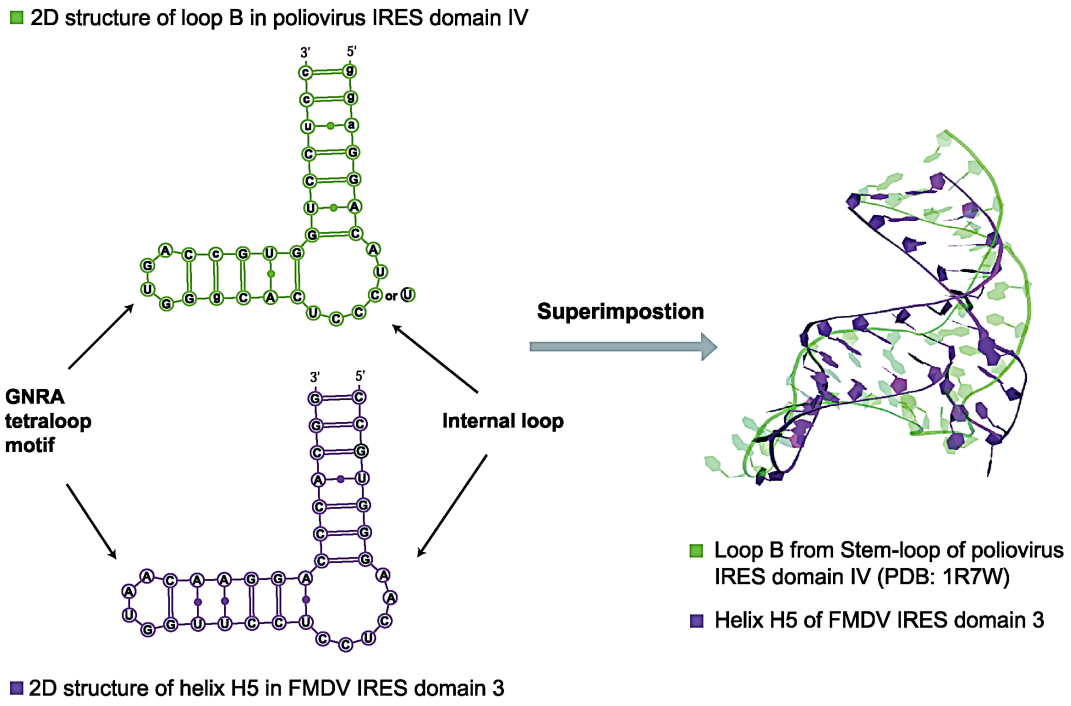


Figure C.5: Structure shape of helix H_5 agrees well with an L-shaped native structure of a poliovirus IRES domain IV. In addition, the helix H_5 containing GNRA motif is compared to the NMR solution data of loop B, equivalent in poliovirus IRES containing GNRA motif. Although the sequences are different, overall shapes of both structures agree well with RMSD value of 7.5Å.

Appendix D

Supplementary Information for Chapter 5

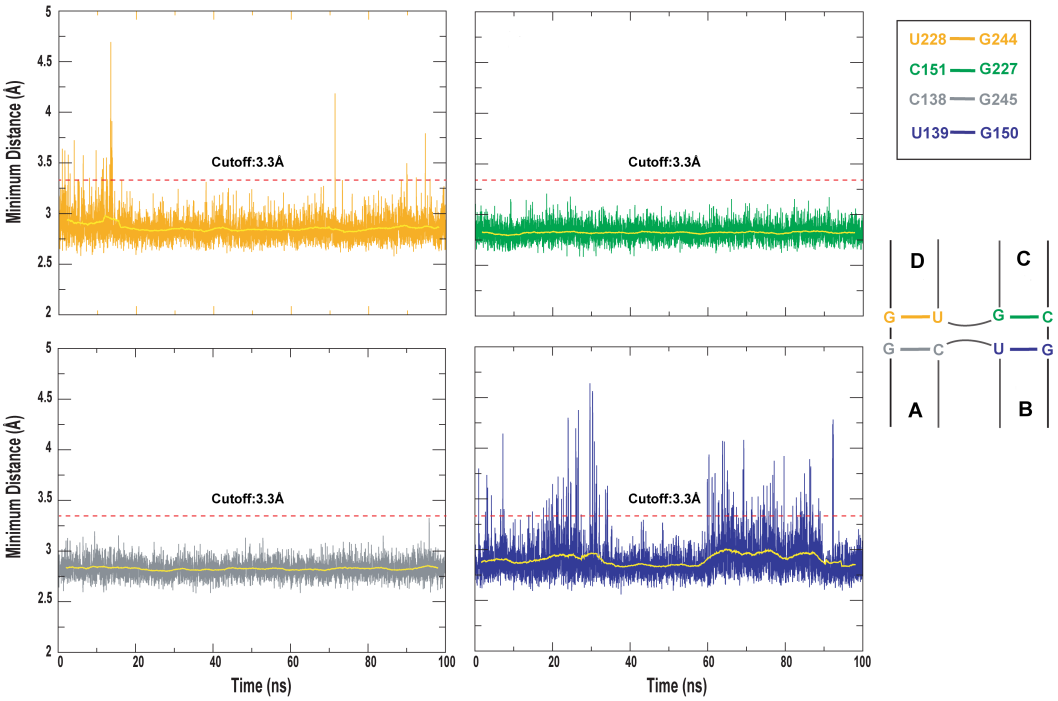


Figure D.1: Distances of heavy atoms in terminal base pairs at the center of the 4H junction. While the distances of G-C base pairs in A (lower left) and C (upper right) are highly stable, the distances of G-U base pairs in B (lower right) and D (upper left) exhibit some fluctuations.

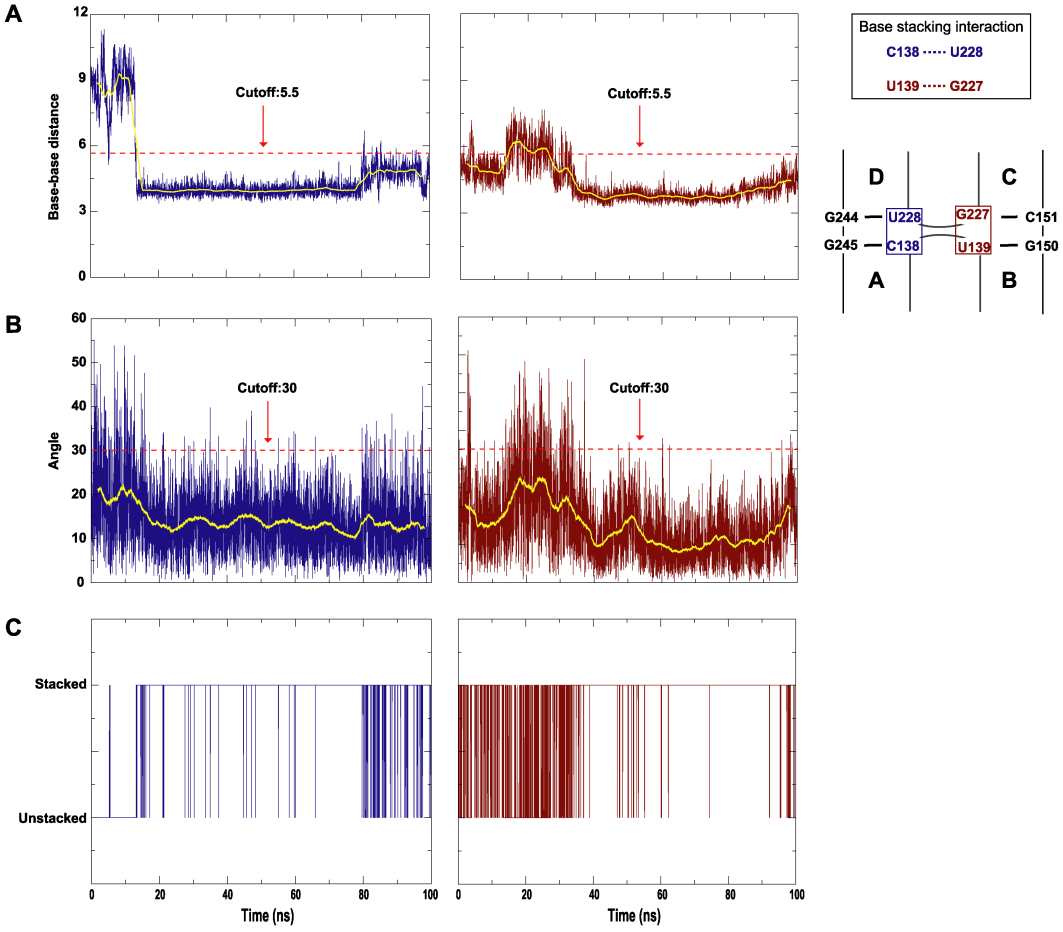


Figure D.2: Base stacking interactions determined by a distance between bases and an angle. **(A)** shows base-base distances with a cutoff value of 5.5Å. **(B)** shows an angle between adjacent bases with a cutoff value of 30°. **(C)** shows base stacking interaction satisfying the distance and angle criteria.

Bibliography

[1] P. L. Adams, M. R. Stahley, A. B. Kosek, J. Wang, and S. A. Strobel. Crystal structure of a self-splicing group i intron with both exons. *Nature*, 430(6995):45–50, 2004.

[2] D. E. Andreev, O. Fernandez-Miragall, J. Ramajo, S. E. Dmitriev, I. M. Terenin, E. Martinez-Salas, and I. N. Shatsky. Differential factor requirement to assemble translation initiation complexes at the alternative start codons of foot-and-mouth disease virus RNA. *RNA*, 13(8):1366–74, 2007.

[3] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon. RNA strand: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9:340, 2008.

[4] M. H. Bailor, X. Sun, and H. M. Al-Hashimi. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science*, 327(5962):202–6, 2010.

[5] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 a resolution. *Science*, 289(5481):905–20, 2000. Ban, N Nissen, P Hansen, J Moore, P

B Steitz, T A GM22778/GM/NIGMS NIH HHS/ GM54216/GM/NIGMS NIH HHS/ New York, N.Y. Science. 2000 Aug 11;289(5481):905-20.

[6] R. T. Batey, S. D. Gilbert, and R. K. Montange. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, 432(7015):411–5, 2004.

[7] G. J. Belsham and N. Sonenberg. RNA-protein interactions in regulation of picornavirus RNA translation. *Microbiol Rev*, 60(3):499–511, 1996.

[8] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 39(Database issue):D32–7, 2011.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–42, 2000.

[10] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. Rnaalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[11] E. Bindewald, R. Hayes, Y. G. Yingling, W. Kasprzak, and B. A. Shapiro. Rnajunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res*, 36(Database issue):D392–7, 2008.

[12] L. Breiman. Random forests. *Machine Learning*, 45(1):52, 2001.

[13] J. M. Burks, C. Zwieb, F. Muller, I. K. Wower, and J. Wower. In silico analysis of ires rnas of foot-and-mouth disease virus and related picornaviruses. *Arch Virol*, 156(10):1737–47, 2011.

[14] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, and J. A. Doudna. Crystal structure of a group i ribozyme domain: principles of RNA packing. *Science*, 273(5282):1678–85, 1996.

[15] J. H. Cate, M. M. Yusupov, G. Z. Yusupova, T. N. Earnest, and H. F. Noller. X-ray crystal structures of 70s ribosome functional complexes. *Science*, 285(5436):2095–104, 1999. Cate, J H Yusupov, M M Yusupova, G Z Earnest, T N Noller, H F GM-17129/GM/NIGMS NIH HHS/ GM-59140/GM/NIGMS NIH HHS/ New York, N.Y. Science. 1999 Sep 24;285(5436):2095-104.

[16] T. H. Chang, J. T. Horng, and H. D. Huang. Rnalogo: a new approach to display structural RNA alignment. *Nucleic Acids Res*, 36(Web Server issue):W91–6, 2008.

[17] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, (117):5179–5197, 1995.

[18] C. C. Correll and K. Swinger. Common and distinctive features of gnra tetraloops based on a guaa tetraloop structure at 1.4 a resolution. *RNA*, 9(3):355–63, 2003.

[19] M. Costa and F. Michel. Frequent use of the same tertiary motif by self-folding rnas. *EMBO J*, 14(6):1276–85, 1995.

[20] M. Costa and F. Michel. Rules for RNA recognition of gnra tetraloops deduced by in vitro selection: comparison with in vivo evolution. *EMBO J*, 16(11):3289–302, 1997.

[21] J. A. Cruz, M. F. Blanchet, M. Boniecki, J. M. Bujnicki, S. J. Chen, S. Cao, R. Das, F. Ding, N. V. Dokholyan, S. C. Flores, L. Huang, C. A. Lavender, V. Lisi, F. Major, K. Mikolajczak, D. J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K. M. Weeks, C. Waldsich, M. Wildauer, N. B. Leontis, and E. Westhof. RNA-puzzles: a casp-like evaluation of RNA three-dimensional structure prediction. *RNA*, 18(4):610–25, 2012.

[22] J. A. Cruz and E. Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136(4):604–9, 2009.

[23] R. Das and D. Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A*, 104(37):14664–9, 2007.

[24] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate shape-directed RNA structure determination. *Genome Res*, 106(1):97–102, 2009.

[25] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D.G. Knowles et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Proc Natl Acad Sci U S A*, 22(9):1775–89, 2012.

[26] D. E. Draper. Strategies for RNA folding. *Trends Biochem Sci*, 21(4):145–9, 1996.

[27] O. Dror, R. Nussinov, and H. J. Wolfson. The arts web server for aligning RNA tertiary structures. *Nucleic Acids Res*, 34(Web Server issue):W412–5, 2006.

[28] Z. Du, N. B. Ulyanov, J. Yu, R. Andino, and T. L. James. Nmr structures of loop b rnas from the stem-loop iv domain of the enterovirus internal ribosome entry site: a single c to u substitution drastically changes the shape and flexibility of RNA. *Biochemistry*, 43(19):5757–71, 2004.

[29] T. Elgavish, J. J. Cannone, J. C. Lee, S. C. Harvey, and R. R. Gutell. Aa.ag@helix.ends: A:a and a:g base-pairs at the ends of 16 s and 23 s rrna helices. *J Mol Biol*, 310(4):735–53, 2001.

[30] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, C. C. Mello Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669): 806–11, 1998.

[31] N. Fernandez, O. Fernandez-Miragall, J. Ramajo, A. Garcia-Sacristan, N. Bellora, E. Eyras, C. Briones, and E. Martinez-Salas. Structural basis for the biological relevance of the invariant apical stem in ires-mediated translation. *Nucleic Acids Res*, 39(19):8572–8585, 2011.

[32] Fernandez, N. and Garcia-Sacristan, A. and Ramajo, J. and Briones, C. and Martinez-Salas, E. Structural analysis provides insights into the modular organization of picornavirus IRES *Virology*, 409(2):251–61, 2011.

[33] O. Fernandez-Miragall and E. Martinez-Salas. Structural organization of a viral ires depends on the integrity of the gnra motif. *RNA*, 9(11):1333–44, 2003.

[34] O. Fernandez-Miragall, R. Ramos, J. Ramajo, and E. Martinez-Salas. Evidence of reciprocal tertiary interactions between conserved motifs involved in organizing RNA structure essential for internal initiation of translation. *RNA*, 12(2):223–34, 2006.

[35] M. G. Gagnon, A. Mukhopadhyay, and S. V. Steinberg. Close packing of helices 3 and 12 of 16 s rrna is required for the normal ribosome function. *J Biol Chem*, 281(51):39349–57, 2006.

[36] M. G. Gagnon and S. V. Steinberg. Gu receptors of double helices mediate trna movement in the ribosome. *RNA*, 8(7):873–7, 2002.

[37] H. H. Gan, D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick. Rag: RNA-as-graphs database–concepts, analysis, and features. *Bioinformatics*, 20(8):1285–91, 2004.

[38] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res*, 31(11):2926–43, 2003.

[39] J. Gillespie, M. Mayne, and M. Jiang. RNA folding on the 3d triangular lattice. *BMC Bioinformatics*, 10:369, 2009.

[40] B. L. Golden, H. Kim, and E. Chase. Crystal structure of a phage twort group i ribozyme-product complex. *Nat Struct Mol Biol*, 12(1):82–9, 2005.

[41] J. Gorodkin and I. L. Hofacker. From structure prediction to genomic screens for novel non-coding rnas. *PLoS Comput Biol*, 7(8):e1002100, 2011.

[42] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31(1):439–41, 2003.

[43] N. Guex and M. C. Peitsch. Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–23, 1997.

[44] F. Guo, A. R. Gooding, and T. R. Cech. Structure of the tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Mol Cell*, 16(3):351–62, 2004.

[45] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys*, 38(3):221–43, 2005.

[46] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–66, 2002.

[47] I. L. Hofacker and P. F. Stadler. Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, 22(10):1172–6, 2006.

[48] S. Hohng, T. J. Wilson, E. Tan, R. M. Clegg, D. M. Lilley, and T. Ha. Conformational flexibility of four-way junctions in RNA. *J Mol Biol*, 336(1):69–79, 2004.

[49] S. R. Holbrook. RNA structure: the long and the short of it. *Curr Opin Struct Biol*, 15(3):302–8, 2005.

[50] S. R. Holbrook. Structural principles from large rnas. *Annu Rev Biophys*, 37:445–64, 2008.

[51] J. Hsin, A. Arkhipov, Yin Y., J. E. Stone, and K. Schulten. Using vmd: an introductory tutorial. *Curr Protoc Bioinformatics*, Chapter 5:Unit 5 7, 2008.

[52] W. Humphrey, A. Dalke, and K. Schulten. Vmd: visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, 1996.

[53] J. A. Izzo, N. Kim, S. Elmetwaly, and T. Schlick. Rag: an update to the RNA-as-graphs resource. *BMC Bioinformatics*, 12:219, 2011.

[54] L. Jaeger, E. J. Verzemnieks, and C. Geary. The UA_handle: a versatile submotif in stable RNA architectures. *Nucleic Acids Res*, 37(1):215–30, 2009.

[55] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, 15(2):189–99, 2009.

[56] F. Jossinet and E. Westhof. Sequence to structure (s2s): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics*, 21(15):3320–1, 2005.

[57] A. V. Kazantsev, A. A. Krivenko, D. J. Harrington, S. R. Holbrook, P. D. Adams, and N. R. Pace. Crystal structure of a bacterial ribonuclease p RNA. *Proc Natl Acad Sci U S A*, 102(38):13392–7, 2005.

[58] J. S. Kieft, K. Zhou, A. Grech, R. Jubin, and J. A. Doudna. Crystal structure of an RNA tertiary domain essential to hcv ires-mediated translation initiation. *Nat Struct Biol*, 9(5):370–4, 2002.

[59] N. Kim, K. N. Furh, and T. Schlick. Graph applications to RNA structure and function. biophysics of RNA folding. *Springer series in Biophysics forthe Life Sciences*, 3, 2012.

[60] N. Kim, H. H. Gan, and T. Schlick. A computational proposal for designing structured RNA pools for in vitro selection of rnas. *RNA*, 13(4):478–92, 2007.

[61] N. Kim, J. Shin, S. Elmetwaly, H. H. Gan, and T. Schlick. RAGPOOLS: RNA-As-Graph-Pools–a web server for assisting the design of structured RNA pools for in vitro selection. *Bioinformatics*, 23(21):2959–60, 2007.

[62] N. Kim, J. A. Izzo, S. Elmetwaly, H. H. Gan, and T. Schlick. Computational generation and screening of RNA motifs in large nucleotide sequence pools. *Nucleic Acids Res*, 38(13):e139, 2010.

[63] N. Kim, N. Shiffeldrim, H. H. Gan, and T. Schlick. Candidates for novel RNA topologies. *J Mol Biol*, 341(5):1129–44, 2004.

[64] S. H. Kim and T. R. Cech. Three-dimensional model of the active site of the self-splicing rrna precursor of tetrahymena. *Proc Natl Acad Sci U S A*, 84(24):8788–92, 1987.

[65] S. H. Kim, J. L. Sussman, F. L. Suddath, G. J. Quigley, A. McPherson, A. H. Wang, N. C. Seeman, and A. Rich. The general structure of transfer RNA molecules. *Proc Natl Acad Sci U S A*, 71(12):4970–4, 1974.

[66] D. J. Klein, P. B. Moore, and T. A. Steitz. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J Mol Biol*, 340(1):141–77, 2004.

[67] P. S. Klosterman, D. K. Hendrix, M. Tamura, S. R. Holbrook, and S. E. Brenner. Three-dimensional motifs from the scor, structural classification of RNA database: extruded strands, base triples, tetraloops and u-turns. *Nucleic Acids Res*, 32(8):2342–52, 2004.

[68] D. R. Koessler, D. J. Knisley, J. Knisley, and T. Haynes. A predictive model for secondary RNA structure using graph theory and a neural network. *BMC Bioinformatics*, 11 Suppl 6:S21, 2010.

[69] V. G. Kolupaeva, C. U. Hellen, and I. N. Shatsky. Structural analysis of the interaction of the pyrimidine tract-binding protein with the internal ribosomal entry site of encephalomyocarditis virus and foot-and-mouth disease virus rnas. *RNA*, 2(12):1199–212, 1996.

[70] A. S. Krasilnikov, Y. Xiao, T. Pan, and A. Mondragon. Basis for structural diversity in homologous rnas. *Science*, 306(5693):104–7, 2004.

[71] A. Laederach, J. M. Chan, A. Schwartzman, E. Willgohs, and R. B. Altman. Coplanar and coaxial orientations of RNA bases and helices. *RNA*, 13(5):643–50, 2007.

[72] C. Laing, S. Jung, A. Iqbal, and T. Schlick. Tertiary motifs revealed in analyses of higher-order RNA junctions. *J Mol Biol*, 393(1):67–82, 2009.

[73] C. Laing and T. Schlick. Analysis of four-way junctions in RNA structures. *J Mol Biol*, 390(3):547–59, 2009.

[74] C. Laing and T. Schlick. Computational approaches to 3d modeling of RNA. *J Phys Condens Matter*, 22(28):283101, 2010.

[75] C. Laing and T. Schlick. Computational approaches to RNA structure prediction, analysis, and design. *Curr Opin Struct Biol*, 21(3):306–18, 2011.

[76] C. Laing, D. Wen, J. T. Wang, and T. Schlick. Predicting coaxial helical stacking in RNA junctions. *Nucleic Acids Res*, 40(2):487–498, 2011.

[77] C. Laing, D. Wen, J. T. Wang, and T. Schlick. Predicting coaxial helical stacking in RNA junctions. *Nucleic Acids Res*, 40(2):487–98, 2012.

[78] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–8, 2007.

[79] H. Leffers, J. Kjems, L. Ostergaard, N. Larsen, and R. A. Garrett. Evolutionary relationships amongst archaebacteria. a comparative study of 23 s ribosomal rnas of a sulphur-dependent extreme thermophile, an extreme halophile and a thermophilic methanogen. *J Mol Biol*, 195(1):43–61, 1987.

[80] V. Lehnert, L. Jaeger, F. Michel, and E. Westhof. New loop-loop tertiary interactions in self-splicing introns of subgroup ic and id: a complete 3d model of the tetrahymena thermophila ribozyme. *Chem Biol*, 3(12):993–1009, 1996.

[81] S. Lemieux and F. Major. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res*, 34(8):2340–6, 2006.

[82] N. B. Leontis, A. Lescoute, and E. Westhof. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol*, 16(3):279–87, 2006.

[83] N. B. Leontis, J. Stombaugh, and E. Westhof. Motif prediction in ribosomal rnas lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie*, 84(9):961–73, 2002.

[84] N. B. Leontis, J. Stombaugh, and E. Westhof. The non-watson-crick base pairs and their associated isostericity matrices. *Nucleic Acids Res*, 30(16):3497–531, 2002.

[85] N. B. Leontis and E. Westhof. A common motif organizes the structure of multi-helix loops in 16 s and 23 s ribosomal rnas. *J Mol Biol*, 283(3):571–83, 1998.

[86] N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

[87] A. Lescoute and E. Westhof. The interaction networks of structured rnas. *Nucleic Acids Res*, 34(22):6587–604, 2006.

[88] D. M. Lilley. Folding of branched RNA species. *Biopolymers*, 48(2-3):101–112, 1998.

[89] D. M. Lilley. Structures of helical junctions in nucleic acids. *Q Rev Biophys*, 33(2):109–59, 2000.

[90] D. M. Lilley, R. M. Clegg, S. Diekmann, N. C. Seeman, E. von Kitzing, and P. Hagerman. Nomenclature committee of the international union of biochemistry and molecular biology (nc-iubmb). a nomenclature of junctions and branchpoints in nucleic acids. recommendations 1994. *Eur J Biochem*, 230(1):1–2, 1995.

[91] J. Lipfert, J. Ouellet, D. G. Norman, S. Doniach, and D. M. Lilley. The complete vs ribozyme in solution studied by small-angle x-ray scattering. *Structure*, 16(9):1357–67, 2008.

[92] S. Lopez de Quinto, E. Lafuente, and E. Martinez-Salas. Ires interaction with translation initiation factors: functional characterization of novel RNA contacts with eif3, eif4b, and eif4gii. *RNA*, 7(9):1213–26, 2001.

[93] S. Lopez de Quinto and E. Martinez-Salas. Conserved structural motifs located in distal loops of aphthovirus internal ribosome entry site domain 3 are required for internal initiation of translation. *J Virol*, 71(5):4171–5, 1997.

[94] S. Lopez de Quinto and E. Martinez-Salas. Interaction of the eif4g initiation factor with the aphthovirus ires is essential for internal translation initiation in vivo. *RNA*, 6(10):1380–92, 2000.

[95] N. Luz and E. Beck. Interaction of a cellular 57-kilodalton protein with the internal translation initiation site of foot-and-mouth disease virus. *J Virol*, 65(12):6486–94, 1991.

[96] E. Martinez-Salas. The impact of RNA structure on picornavirus ires activity. *Trends Microbiol*, 16(5):230–7, 2008.

[97] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999.

[98] S. E. McDowell, N. Spackova, J. Sponer, and N. G. Walter. Molecular dynamics simulations of RNA: an in silico single molecule approach. *Biopolymers*, 85(2):169–84, 2007.

[99] E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *J Am Chem Soc*, 127(12):4223–31, 2005.

[100] K. Meyer, A. Petersen, M. Niepmann, and E. Beck. Interaction of eukaryotic initiation factor eif-4b with a picornavirus internal translation initiation site. *J Virol*, 69(5):2819–24, 1995.

[101] M. M. Meyer, T. D. Ames, D. P. Smith, Z. Weinberg, M. S. Schwalbach, S. J. Giovannoni, and R. R. Breaker. Identification of candidate structured rnas in the marine organism 'candidatus pelagibacter ubique'. *BMC Genomics*, 10:268, 2009.

[102] V. Mlynsky, P. Banas, D. Hollas, K. Reblova, N. G. Walter, J. Sponer, and M. Otyepka. Extensive molecular dynamics simulations showing that canonical g8 and protonated a38h+ forms are most consistent with crystal structures of hairpin ribozyme. *J Phys Chem B*, 114(19):6642–52, 2010.

[103] A. Mokdad, M. V. Krasovska, J. Sponer, and N. B. Leontis. Structural and evolutionary classification of g/u wobble basepairs in the ribosome. *Nucleic Acids Res*, 34(5):1326–41, 2006.

[104] M. Nanias, M. Chinchio, J. Pillardy, D. R. Ripoll, and H. A. Scheraga. Packing helices in proteins by global optimization of a potential energy function. *Proc Natl Acad Sci U S A*, 100(4):1706–10, 2003.

[105] P. Nissen, J. A. Ippolito, N. Ban, P. B. Moore, and T. A. Steitz. RNA tertiary interactions in the large ribosomal subunit: the a-minor motif. *Proc Natl Acad Sci U S A*, 98(9):4899–903, 2001.

[106] H. F. Noller. RNA structure: reading the ribosome. *Science*, 309(5740):1508–14, 2005.

[107] A. Pacheco, S. Lopez de Quinto, J. Ramajo, N. Fernandez, and E. Martinez-Salas. A novel role for gemin5 in mrna translation. *Nucleic Acids Res*, 37(2):582–90, 2009.

[108] M. Parisien, J. A. Cruz, E. Westhof, and F. Major. New metrics for comparing and assessing discrepancies between RNA 3d structures and models. *RNA*, 15(10):1875–85, 2009.

[109] M. Parisien and F. Major. The mc-fold and mc-sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–5, 2008.

[110] J. C. Penedo, T. J. Wilson, S. D. Jayasena, A. Khvorova, and D. M. Lilley. Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *RNA*, 10(5):880–8, 2004.

[111] A. Perez, I. Marchan, D. Svozil, J. Sponer, 3rd Cheatham, T. E., C. A. Laughton, and M. Orozco. Refinement of the amber force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J*, 92(11):3817–29, 2007.

[112] Phillips, J. C. and Braun, R. and Wang, W. and Gumbart, J. and Tajkhorshid, E. and Villa, E. and Chipot, C. and Skeel, R. D. and Kale, L. and Schulten, K. Scalable molecular dynamics with NAMD *J Comput Chem*, 26(16):1781–802, 2005.

[113] E. V. Pilipenko, V. M. Blinov, B. K. Chernov, T. M. Dmitrieva, and V. I. Agol. Conservation of the secondary structure elements of the 5'-untranslated region of cardio- and aphthovirus rnas. *Nucleic Acids Res*, 17(14):5701–11, 1989.

[114] E. V. Pilipenko, T. V. Pestova, V. G. Kolupaeva, E. V. Khitrina, A. N. Poperechnaya, V. I. Agol, and C. U. Hellen. A cell cycle-dependent protein serves as a template-specific translation initiation factor. *Genes Dev*, 14(16):2028–45, 2000.

[115] H. W. Pley, K. M. Flaherty, and D. B. McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372(6501):68–74, 1994.

[116] R. Ramos and E. Martinez-Salas. Long-range RNA interactions between structural domains of the aphthovirus internal ribosome entry site (ires). *RNA*, 5(10):1374–83, 1999.

[117] F. Razga, J. Koca, J. Sponer, and N. B. Leontis. Hinge-like motions in RNA kink-turns: the role of the second a-minor motif and nominally unpaired bases. *Biophys J*, 88(5):3466–85, 2005.

[118] K. Reblova, F. Lankas, F. Razga, M. V. Krasovska, J. Koca, and J. Sponer. Structure, dynamics, and elasticity of free 16s rrna helix 44 studied by molecular dynamics simulations. *Biopolymers*, 82(5):504–20, 2006.

[119] K. Reblova, J. Sponer, and F. Lankas. Structure and mechanical properties of the ribosomal l1 stalk three-way junction. *Nucleic Acids Res*, 40(13):6290–303, 2012.

[120] M. E. Robertson, R. A. Seamons, and G. J. Belsham. A selection system for functional internal ribosome entry site (ires) elements: analysis of the requirement for a conserved gnra tetraloop in the encephalomyocarditis virus ires. *RNA*, 5(9):1167–79, 1999.

[121] M. Rother, K. Rother, T. Puton, and J. M. Bujnicki. Moderna: a tool for comparative modeling of RNA 3d structure. *Nucleic Acids Res*, 39(10):4007–22, 2011.

[122] M. Sarver, C. L. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis. Fr3d: finding local and composite recurrent structural motifs in RNA 3d structures. *J Math Biol*, 56(1-2):215–52, 2008.

[123] T. Schlick. A modular strategy for generating starting conformations and data structures of polynucleotide helices for potential energy calculations. *J of Comp Chem*, 9:861–89, 1988.

[124] T. Schlick, R. Collepardo-Guevara, L. A. Halvorsen, S. Jung, and X. Xiao. Biomolecularmodeling and simulation: a field coming of age. *Q Rev Biophys*, 44(2):191–228, 2011.

[125] A. Serganov, Y. R. Yuan, O. Pikovskaya, A. Polonskaia, L. Malinina, A. T. Phan, C. Hobartner, R. Micura, R. R. Breaker, and D. J. Patel. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mrnas. *Chem Biol*, 11(12):1729–41, 2004.

[126] P. Serrano, J. Gomez, and E. Martinez-Salas. Characterization of a cyanobacterial rnase p ribozyme recognition motif in the ires of foot-and-mouth disease virus reveals a unique structural element. *RNA*, 13(6):849–59, 2007.

[127] Y. Song, E. Tzima, K. Ochs, G. Bassili, H. Trusheim, M. Linder, K. T. Preissner, and M. Niepmann. Evidence for an RNA chaperone function of polypyrimidine tract-binding protein in picornavirus translation. *RNA*, 11(12):1809–24, 2005.

[128] J. Sponer, JV. Burda, Sabat M., J. Leszczynski, and P. Hobza. Interaction between the guanine-cytosine watson-crick dna base pair and hydrated group iia (mg2+, ca2+, sr2+, ba2+) and group iib (zn2+, cd2+, hg2+) metal cations. *J Phys Chem A*, 102(29):5951–5957, 1998.

[129] I. A. Stassinopoulos and G. J. Belsham. A novel protein-RNA binding assay: functional interactions of the foot-and-mouth disease virus internal ribosome entry site with cellular proteins. *RNA*, 7(1):114–22, 2001.

[130] S. V. Steinberg and Y. I. Boutorine. G-ribo: a new structural motif in ribosomal RNA. *RNA*, 13(4):549–54, 2007.

[131] M. Tamura and S. R. Holbrook. Sequence and structural conservation in RNA ribose zippers. *J Mol Biol*, 320(3):455–74, 2002.

[132] D. Tirumalai and C. Hyeon. Theory of RNA folding: From hairpins to ribozymes. springer series in biophysics. *Springer Series in Biophysics*, 13(6995):27–47, 2009.

[133] N. Toor, K. S. Keating, S. D. Taylor, and A. M. Pyle. Crystal structure of a self-spliced group ii intron. *Science*, 320(5872):77–82, 2008.

[134] T. Tuschl, C. Gohlke, T. M. Jovin, E. Westhof, and F. Eckstein. A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, 266(5186):785–9, 1994.

[135] R. Tyagi and D. H. Mathews. Predicting helical coaxial stacking in RNA multibranch loops. *RNA*, 13(7):939–51, 2007.

[136] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. Gromacs: fast, flexible, and free. *J Comput Chem*, 26(16):1701–18, 2005.

[137] F. Walter, A. I. Murchie, D. R. Duckett, and D. M. Lilley. Global structure of four-way RNA junctions studied using fluorescence resonance energy transfer. *RNA*, 4(6):719–28, 1998.

[138] Z. Weinberg, J. X. Wang, J. Bogue, J. Yang, K. Corbino, R. H. Moy, and R. R. Breaker. Comparative genomics reveals 104 candidate struc-

tured rnas from bacteria, archaea, and their metagenomes. *Genome Biol*, 11(3):R31, 2010.

[139] T. J. Wilson, M. Nahas, T. Ha, and D. M. Lilley. Folding and catalysis of the hairpin ribozyme. *Biochem Soc Trans*, 33(Pt 3):461–5, 2005.

[140] B. T. Wimberly, D. E. Brodersen, Jr. Clemons, W. M., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30s ribosomal subunit. *Nature*, 407(6802):327–39, 2000.

[141] Y. Xin, C. Laing, N. B. Leontis, and T. Schlick. Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA*, 14(12):2465–77, 2008.

[142] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*, 31(13):3450–60, 2003.

[143] I. Yildirim, S. D. Kennedy, H. A. Stern, J. M. Hart, R. Kierzek, and D. H. Turner. Revision of amber torsional parameters for RNA improves free energy predictions for tetramer duplexes with gc and igic base pairs. *J Chem Theory Comput*, 8(1):172–181, 2012.

[144] M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller. Crystal structure of the ribosome at 5.5 a resolution. *Science*, 292(5518):883–96, 2001.

[145] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, 3rd Cheatham, T. E., and P. Jurecka. 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(9):2886–2902, 2011.

[146] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, 1981.

[147] W. G. Scott, J. B. Murray, J. R. Arnold, B. L. Stoddard, and A. Klug. Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science*, 274(5295):2065–9, 1996.

[148] S. Jung and T. Schlick. Candidate RNA structures for domain 3 of the foot-and-mouth-disease virus internal ribosome entry site. *Nucleic Acids Res*, 41(3):1483–95, 2013.

[149] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, 102(5):615–23, 2000.

[150] A. Rich and D. R. Davies. A new two-stranded helical structure: polyadenylic acid and polyuridylic acid. *J Am Chem Soc*, 78(14):3548–9, 1956.

[151] J. D. Robertus, J. E. Ladner, J. T. Finch, D. Rhodes, R. S. Brown, B. F. Clark, and A. Klug. Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature*, 250(467):546–51, 1974.

[152] Encode Project Consortium, I. Dunham, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[153] Z. Weinberg, J. Perreault, M. M. Meyer, and R. R. Breaker. Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature*, 462(7273):656–9, 2009.

[154] I. Tinoco, Jr. and C. Bustamante. How RNA folds. *J Mol Biol*, 293(2):271–81, 1999.

[155] S. Hanlon. The importance of London dispersion forces in the maintenance of the deoxyribonucleic acid helix. *Biochem Biophys Res Commun*, 23(6):861–7, 1966.

[156] I. Tazawa, T. Koike, and Y. Inoue. Stacking properties of a highly hydrophobic dinucleotide sequence, N6, N6-dimethyladenylyl(3' leads to 5')N6, N6-dimethyladenosine, occurring in 16–18-S ribosomal RNA. *Eur J Biochem*, 109(1):33–8, 1980.

[157] A. Sarai, J. Mazur, R. Nussinov, and R. L. Jernigan. Origin of DNA helical structure and its sequence dependence. *Biochemistry*, 27(22):8498–502, 1988.

[158] L. Nasalean, J. Stombaugh, C. L. Zirbel, and N. B. Leontis. RNA 3D structural motifs: definition, identification, annotation, and database searching. *Springer Series in Biophysics In Non-protein Coding RNAs*, 1–26, 2009.

[159] H. M. Martinez, J. V. Maizel, Jr., and B. A. Shapiro. RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J Biomol Struct Dyn*, 25(6):669–83, 2008.

[160] F. Jossinet, T. E. Ludwig, E. Westhof. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics*, 26(16):2057–9, 2010.

[161] S. Sharma, F. Ding, and N. V. Dokholyan. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17):1951–2, 2008.

[162] J. Frellsen, I. Moltke, M. Thiim, K. V. Mardia, J. Ferkinghoff-Borg, and T. Hamelryck. A probabilistic model of RNA conformational space. *PLoS Comput Biol*, 5(6)e1000406, 2009.

[163] A. Y. Sim, M. Levitt, and P. Minary. Modeling and design by hierarchical natural moves. *Proc Natl Acad Sci U S A*, 109(8):2890–5, 2012.

[164] N. Kim, K. N. Fuhr, and T. Schlick. Graph applications to RNA structure and function. *Springer Series in Biophysics of RNA Folding*, 3:23–51, 2013.

[165] M. S. Waterman. Secondary structure of single-stranded nucleic acids. Advances in Mathematics Supplementary Studies. *Springer Series in Biophysics of RNA Folding*, 1:167–212, 1978.

[166] B. A. Shapiro and K.Z. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *Comput Appl Biosci*, 6:309–18, 1990.

[167] S. Y. Le, R. Nussinov, and J.V. Maizel. Tree graphs of RNA secondary structures and their comparisons. *Comput Biomed Res*, 22:461–73, 1989.

[168] G. Benedetti and S. Morosetti. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophys Chem*, 59:179–84, 1996.

[169] A. Gopal, Z. H. Zhou, C. M. Knobler, and W. M. Gelbart. Visualizing large RNA molecules in solution. *RNA*, 18:284–99, 2012.

[170] T. E. Cheatham, 3rd. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr Opin Struct Biol*, 14:360–7, 2004.

[171] M. A. Ditzler, M. Otyepka, J. Sponer, and N. G. Walter. Molecular dynamics and quantum mechanics of RNA: conformational and chemical change we can believe in. *Acc Chem Res*, 43:40–7, 2010.

[172] A. Villa, J. Wohnert, and G. Stock. Molecular dynamics simulation study of the binding of purine bases to the aptamer domain of the guanine sensing riboswitch. *Nucleic Acids Res*, 37:4774–86, 2009.

[173] P. Auffinger, L. Bielecki, and E. Westhof. Symmetric K+ and Mg2+ ion-binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J Mol Biol*, 335:555–71, 2004.

[174] W. Li, B. Ma, and B. A. Shapiro. Binding interactions between the core central domain of 16S rRNA and the ribosomal protein S15 determined by molecular dynamics simulations. *Nucleic Acids Res*, 31:629–38, 2003.

[175] I. Besseova, K. Reblova, N. B. Leontis, and J. Sponer. Molecular dynamics simulations suggest that RNA three-way junctions can act as flexible RNA structural elements in the ribosome. *Nucleic Acids Res*, 38:6247–64, 2010.

[176] P. Koehl. Electrostatics calculations: latest methodological advances. *Curr Opin Struct Biol*, 16:142–51, 2006.

[177] J. A. Nelson and O. C. Uhlenbeck. When to believe what you see. *Mol Cell*, 23:447–50, 2006.

[178] A. Mokdad and A. D. Frankel. ISFOLD: structure prediction of base pairs in non-helical RNA motifs from isostericity signatures in their sequence alignments. *J Biomol Struct Dyn*, 25:467–72, 2008.

[179] P. S. Pang, E. Jankowsky, L. M. Wadley, and A. M. Pyle. Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data. *J Exp Zool B Mol Dev Evol*, 304:50–63, 2005.

[180] A. Serganov, A. Polonskaia, A. T. Phan, R. R. Breaker, D. J. Patel. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, 441(7097):1167–71, 2006.

[181] N. Foloppe and A. D. MacKerell, Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comput Chem*, 21:86–104, 2000.

[182] M. Chastain and I. Tinoco, Jr. Structural elements in RNA. *Prog Nucleic Acid Res Mol Biol*, 41:131–77, 1991.

[183] A. I. Murchie, J. B. Thomson, F. Walter, and D. M. Lilley. Folding of the hairpin ribozyme in its natural conformation achieves close physical proximity of the loops. *Mol Cell*, 1:873–81, 1998.

[184] D. R. Duckett, A. I. Murchie, and D. M. Lilley. The global folding of four-way helical junctions in RNA, including that in U1 snRNA. *Cell*, 83:1027–36, 1995.

[185] D. Boehringer, R. Thermann, A. Ostareck-Lederer, J. D. Lewis, and H. Stark. Structure of the hepatitis C virus IRES bound to the human 80S ribosome: remodeling of the HCV IRES. *Structure*, 13:1695–706, 2005.

[186] C. Branlant, A. Krol,J. P. Ebel, H. Gallinaro, E. Lazar, and M. Jacob. The conformation of chicken, rat and human U1A RNAs in solution. *Nucleic Acids Res*, 9:841–58, 1981.

[187] A. Hampel, and R. Tritz. RNA catalytic properties of the minimum (-)sTRSV sequence. *Biochemistry*, 28:4929–33, 1989.

[188] R. A. Poot, N. V. Tsareva, , I. V. Boni, and J. V. Duin. RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proc Natl Acad Sci U S A*, 94:10110–5, 1997.

[189] R. K. Montange and R. T. Batey. Riboswitches: emerging themes in RNA structure and function. *Annu Rev Biophys*, 37:117–33, 2008.

[190] N. Deng and P. Cieplak. Free Energy Profile of RNA Hairpins: A Molecular Dynamics Simulation Study. *Biophys J*, 98(4):627–36, 2010.

[191] J. BEZDEK. Fuzzy models what are they, and why?. *IEEE Transactions on Fuzzy Systems*, 1:1–5, 1993.

[192] S. E. Melcher, T. J. Wilson, and D. M. Lilley. The dynamic nature of the four-way junction of the hepatitis C virus IRES. *RNA*, 9(7):809–20, 2003.

[193] B. Masquida, B. Beckert, and F. Jossinet. Exploring RNA structure by integrative molecular modelling. *New Biotechnology*, 27(3):170–83, 2010.

[194] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, 77:6309–13, 1980.

[195] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for Loop Matchings. *SIAM J Appl Math*, 35:68–82, 1978.

[196] P. Gendron, S. Lemieux, and F. Major. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol*, 308:919–36, 2001.

[197] S. R. Holbrook, J. L. Sussman, R. W. Warrant, and S. H. Kim. Crystal structure of yeast phenylalanine transfer RNA. II. Structural features and functional implications. *J Mol Biol*, 123:631–60, 1978.

[198] C. E. Hajdin, F. Ding, N. V. Dokholyan, and K. M. Weeks. On the significance of an RNA tertiary structure prediction. *RNA*, 16:1340–9, 2010.

[199] C. Laing*, S. Jung*, N. Kim, S. Elmetwaly, M. Zahran, and T. Schlick. Predicting Helical Topologies in RNA Junctions as Tree Graphs. *PLOS ONE*, In Press.

[200] E. Tan, T. J. Wilson, M. K. Nahas, R. M. Clegg, D. M. Lilley, and T. Ha. A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc Natl Acad Sci U S A*, 100:9308–13, 2003.

[201] J. Nowakowski, P. J. Shim, C. D. Stout, and G. F. Joyce. Alternative conformations of a nucleic acid four-way junction. *J Mol Biol*, 300:93–102, 2000.

[202] U. D. Priyakumar, and A. D. MacKerell, Jr. Role of the adenine ligand on the stabilization of the secondary and tertiary interactions in the adenine riboswitch. *J Mol Biol*, 396:1422–38, 2010.

[203] W. Huang, J. Kim, S. Jha, and F. Aboul-ela. The impact of a ligand binding on strand migration in the SAM-I riboswitch. *PLoS Comput Biol*, 9:e1003069, 2013.

[204] K. Chen, J. Eargle, J. Lai, H. Kim, S. Abeysirigunawardena, M. Mayerle, S. Woodson, T. Ha, and Z. Luthey-Schulten. Assembly of the five-way junction in the ribosomal small subunit using hybrid MD-Go simulations. *J Phys Chem B*, 116:6819–31, 2012.

[205] J. Perard, C. Leyrat, F. Baudin, E. Drouet, and M. Jamin. Structure of the full-length HCV IRES in solution. *Nat Commun*, 4:1612, 2013.

[206] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79:926–35, 1983.

[207] D. A. Case, T. A. Darden, T. E. Cheatham, III., C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, et al. AMBER 12. *University of California, San Francisco*, 2012.

[208] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7:155–62, 1964.

[209] A. Amadei, A. B. Linssen, and H. J. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–25, 1993.

[210] P. B. Rupert, and A. R. Ferre-D'Amare. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature*, 410:780–6, 2001.

[211] Z. Y. Zhao, T. J. Wilson, K. Maxwell, and D. M. Lilley. The folding of the hairpin ribozyme: dependence on the loops and the junction. *RNA*, 6:1833–46, 2000.

[212] G. L. Conn, D. E. Draper, E. E. Lattman, and A. G. Gittis. Crystal structure of a conserved ribosomal protein-RNA complex. *Science*, 284:1171–4, 1999.

[213] C. Geary, A. Chworos, and L. Jaeger. Promoting RNA helical stacking via A-minor junctions. *Nucleic Acids Res*, 39:1066–80, 2011.